Geometric Design of Roads Handbook

Keith M. Wolhuter

CRC Press
Taylor & Francis Group
A SPON PRESS BOOK

Geometric Design of Roads Handbook

Geometric Design of Roads Handbook

Keith M. Wolhuter

Consultant Engineer, South Africa

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A SPON PRESS BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150303

International Standard Book Number-13: 978-1-4822-8872-8 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Writing a book is an extremely demanding and selfish task. This book is dedicated to my wife, Helen, who spent many weeks, months and years alone while I was in my study writing it.

And also to

Dr. Graham LD Ross, who showed me that geometric design is a three-dimensional jigsaw puzzle and is as much art as it is science.

Contents

List of figures

List of tables

Foreword

Keith M. Wolhuter has had a career spanning more than 50 years in both the public and private sectors including planning, design, construction, research and policy/administration. Although most of his work was accomplished in South Africa and adjacent countries, Mr. Wolhuter was exposed to highway design criteria, solutions and research in other parts of the world as well. He applied this knowledge, adapted to conditions in South Africa, in the development of design criteria and design solutions. His education, experience and qualifications are well recognized in the highway design community.

This comprehensive geometric design text developed by Mr. Wolhuter is 'state of the art' and has application worldwide. It is much more than the regurgitation of design criteria. The text incorporates a philosophy reflective of multimodal design, context-sensitive solutions, application of human factors and safety. Design criteria will evolve as future research and experience continue; however, the concepts, procedures and multimodal approach presented in this book have lasting relevance.

While all of the basics are covered related to geometric design, Keith M. Wolhuter has also covered all facility types incorporating capacity/traffic operations, multimodal considerations, constructability and the social and economic environment in the design process. Other aspects that influence 'good' design include drainage and environmental issues. These are also addressed. As a geometric design text it is unique in that it is developed in two parts: Part I, Geometric Design and Part II, Supporting Disciplines. This is a 'how to do it' book.

The text can serve as the basis for several graduate-level courses related to geometric design, traffic operations/capacity analysis and public transport. Because much of the material is based on experience in successful project execution, it demonstrates to the student the development of real-world solutions. This is more than a typical textbook.

Practicing engineers will find it extremely valuable in the development of solutions that can be implemented. Geometric design criteria are guidelines and not rules that have to be rigidly applied. It is this book that guides the planner and designer in the application of the design criteria/guidelines in creating designs that will operate efficiently and safely for all roadway users within the facilities' cultural, economic and environmental context.

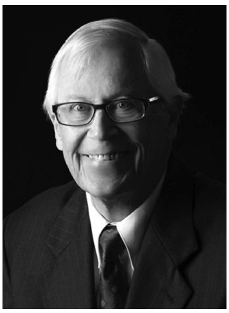
It is not possible to cover in depth all the design concepts and procedures discussed in this book. Consequently, references to additional literature on each topic are provided. These cited references are those consulted in the development of the material and are provided for readers' further consideration on the subject.

I commend Keith M. Wolhuter for undertaking a selfless monumental task to provide a book that has far-reaching value for the highway and traffic engineering profession.

Joel P. Leisch, PE, BSCE, MSCE, LMITE, FASCE

Author

KEITH M. WOLHUTER AND HIS ROLE IN GEOMETRIC DESIGN IN SOUTHERN AFRICA



Keith M. Wolhuter, PrEng, BSc, BEng, MEng, MSAICE, graduated from the University of Stellenbosch in South Africa and joined the Cape Provincial Roads Department in January 1960.

At that time, the department had a training programme that attached young engineers for a period of about three years to construction units headed by experienced construction engineers. This was followed by a further three years as resident engineers in their own right. After this, career paths split into design, construction and maintenance.

After his activity in the field, Mr. Wolhuter joined the Geometric Design Programme, where he ultimately became the head of the Interchange Design Section. This section was responsible for the departmental design of freeways and interchanges as well as overview of design by consultants.

In 1968, Mr. Wolhuter accepted a partnership in the consultancy BA Kantey and Partners to run their Bloemfontein office. His duties, in addition to office management, included the design, contract documentation and supervision of contract construction of rural roads as well as planning and design of freeways and interchanges on behalf of the Pretoria office.

His last project as head of the Bloemfontein office was the planning and design of section of the Maluti Route. The Maluti Route more or less follows the boundary between Lesotho and South Africa. The section of interest was that between Harrismith and Qwaqwa. The main project was the Maluti Route itself but included a spur over the Drakensberg Mountain to Bergville in Natal and also a spur to Phutaditjaba, the capital of Qwaqwa.

Four road authorities were involved. The main client was the Free State Roads Department with its set of geometric standards. The spur over the Drakensberg Mountain brought the standards of the Natal Department of Transport into play, and the spur to Phutaditjaba involved the standards of the then Bantu Affairs Department. A major subsidy by the National Department of Transport required that its geometric design standards also demanded attention.

The Council for Scientific and Industrial Research (CSIR) believed that this state of affairs was unacceptable. On behalf of the Committee of State Road Authorities, which is the South African equivalent of the American Association of State Highway and Transportation Officials, it approached Mr. Wolhuter to seek uniformity of practice between the national and four provincial departments. He resigned his partnership and joined CSIR to undertake this task. It culminated in the issue in 1984 of *Transportation Recommendations for Highways (TRH) 17: The Geometric Design of Rural Roads*. TRH 17 was the basic guideline for rural geometric design countrywide and also became the standard teaching text for all South African universities and technicons.

Mr. Wolhuter acquired his MEng (Transportation) from the University of Pretoria in 1992.

During the late 1980s and 1990s, international freight transport by road in southern Africa had increased exponentially and the governments of the Southern African Development Community embarked on a process of securing uniformity of geometric design practice between themselves with respect to their agreed International Trunk Road Network. Mr. Wolhuter was briefed to develop this Code of Practice. It was based on the updated TRH 17 and the documentation of the various member states and was issued in 1998. The Ministers of Transport of Angola, Botswana, Democratic Republic of the Congo, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia and Zimbabwe ratified the document in 2002.

In 1998, the South African National Road Agency Limited (SANRAL) was created by the Department of Transport as a parastatal responsible for the management, maintenance and development of South Africa's national road network. This body realised various paradigm shifts of note, notably

- The emergence in 1998 of context-sensitive design (also known as context-sensitive solutions because achievement of context sensitivity does not necessarily always require design inputs)
- The quantification of consistency of design
- The concept of the design domain develop by the Canadian road authorities
- Design variances
- Road safety audits
- The transformation of geometric design to a multidisciplinary team effort

with the consequent need for updating the National Roads Department geometric design standards and indeed TRH 17 itself.

CSIR was requested to redraft the geometric design standards for SANRAL and it briefed Messrs Burrell, Mitchell and Wolhuter to undertake this task. A CD titled 'Geometric Design Guidelines' was issued by SANRAL in November 2002. A feature of this document was that it was based almost exclusively on research that had taken place after the 'Thinking beyond the Pavement' Workshop held in 1998.

Since then Mr. Wolhuter has been a member of the team that developed the geometric design section of the Botswana *Road Design Manual*. In 2014 he was asked to develop a geometric design manual for Rwanda but, having achieved the age of 78, graciously declined this task.

Glossary

acceleration lane an auxiliary lane used by an entering vehicle to accelerate before entering the travelled way.

access control the condition whereby the road agency either partially or fully controls the right of abutting landowners to have access to and from a public highway or road.

access interchange an interchange providing access to a freeway from the adjacent non-freeway road network. Referred to in American practice as a service interchange.

arterial highway designed to move relatively large volumes of traffic at high speeds over long distances. Typically, arterials offer little or no access to abutting properties. In the latter case, referred to as a freeway.

auxiliary lane short lane located immediately adjacent to the basic or through lane to accommodate some or other special circumstance such as a turning movement to right or to left, acceleration to or deceleration from the speeds prevailing on the travelled way, or heavy vehicles reduced to crawling speeds on steep upgrades or downgrades.

average daily traffic (ADT) the number of vehicles per day passing a point on the highway during a defined period. If this period extends from 1 January to 31 December, reference is to annual average daily traffic (AADT).

average running speed the distance summation for all vehicles divided by the running time summation for all vehicles. Also referred to as space mean speed, whereas time mean speed is simply the average of all recorded speeds.

axis of rotation the line about which the pavement is rotated to superelevate or bank the roadway. This line normally maintains the highway profile.

barrier sight distance the limiting sight distance below which overtaking is legally prohibited.

boulevard the area separating sidewalks from the through lanes or a landscaped cross-section including through lanes and adjacent service lanes.

bridge a structure erected with a deck for carrying traffic over or under an obstruction and with a clear span of 6 metres or more. Where the clear span is less than 6 metres, reference is to a culvert.

broken-back curve two curves in the same direction with a tangent shorter than 500 metres long connecting them.

camber the slope from a high point (typically at the centreline of the highway) across the lanes of a highway. Negative camber refers to a central low point, usually with a view to drainage of a small urban street or alley.

capacity the maximum number of vehicles that can pass a point on a highway or in a designated lane in one hour without the density being so great as to cause unreasonable delay or to restrict the driver's freedom to manoeuvre under prevailing roadway and traffic conditions.

carriageway roadway forming part of a divided highway and intended for movement in one direction only – hence dual carriageway as an alternative name for divided highway.

catchwater drain located above a cut face to ensure that storm water does not flow down the cut face causing erosion of the cut face followed by deposition of silt on the roadway.

channel grading where side channels are designed to gradients that differ from those of the road centreline, typically on either side of the highest points on crest curves and the lowest points on sag curves where the centreline gradient is less than 0.5 per cent.

channelisation the use of pavement markings or islands to direct traffic through an intersection.

clearance profile describes the space that is exclusively reserved for provision of the road or highway. It defines the minimum height of the soffit of any structure passing over the road and the closest approach of any lateral obstacle to the cross-section.

cloverleaf interchange an interchange with loop ramps in all quadrants to accommodate left turns and outer connectors for the right turns.

collector a road characterised by a roughly even distribution of its access and mobility functions.

collector–distributor road a road used at an interchange to remove weaving from the through lanes and to reduce the number of entrances to and exits from the through lanes.

compound curve a combination of two or more curves in the same direction without intervening tangents between them.

criterion a yardstick according to which some or other quality of the road can be measured. Guideline values are specific numerical values of the criterion. For example, delay is a criterion of congestion.

critical length of grade the maximum length of a specific upgrade on which a loaded truck can operate without an unreasonable reduction in speed. Very often, a speed reduction of 15 km/h or more is considered 'unreasonable'.

cross fall in the case of cross fall, the high point is at the roadway edge. See camber.

cross-over crownline the line across which an instantaneous change of camber takes place. In the case of a normally cambered road, the centreline is a special case of the cross-over crownline. The cross-over crownline can be located anywhere on the road surface and need not even be parallel to the road centreline.

crosswalk a demarcated area or lane designated for the use of pedestrians across a road or street.

crown runoff also referred to as tangent runout. The rotation of the outer lane of a two-lane road from zero cross fall to normal camber (NC).

culvert a structure, with a clear span of less than 6 metres, usually for conveying water under a roadway but can also be used as a pedestrian or stock crossing.

cut section of highway or road below natural ground level. Sometimes referred to in other documents as a cutting or excavation.

cycle lane a portion of the roadway that has been designated by road markings, striping and signing as being exclusively for the use of cyclists.

cycle path also known as a bike way. A path physically separated from vehicular traffic by an open space or barrier and either within the road reserve or in an independent reserve.

decision sight distance sometimes referred to as anticipatory sight distance, allows for circumstances where complex decisions are required or unusual manoeuvres have to be carried out. As such, it is significantly longer than stopping sight distance.

density the number of vehicles occupying a given length of road. Usually averaged over time and expressed as vehicles per kilometre.

depressed median a median lower in elevation than the travelled way and so designed to carry portions of the storm water falling on the road.
design domain the range of values of a design criterion that are applicable to a given design, e.g. lane widths of more than 3.3 metres.
design hour the hour in which the condition being designed for, typically the anticipated flow, is expected to occur. This is often the thirtieth highest hour of flow in the design year.
design speed the maximum safe speed that can be maintained over a specified section of road or highway when conditions are so favourable that the design features of the highway govern.
design vehicle a compilation of the 85th percentile values of the various parameters of the vehicle type being designed for, e.g. length, width, wheelbase, overhang, height, ground clearance, etc., and not a commercially available vehicle.
design year the last year of the design life of the road or any other facility, often taken as twenty years, although, for costly structures such as major bridges, a longer period is usually adopted.
directional distribution (split) the percentages of the total flow moving in opposite directions, e.g. 50:50, 70:30, with the direction of interest being quoted first.
divided highway a highway with separate carriageways for traffic moving in opposite directions.
driveway a road providing access from a public road to a street or road usually located on an abutting property.
eighty-fifth percentile speed the speed below which 85 per cent of the vehicles travel on a given road or highway.
escarpment a long steep slope separating a plateau from terrain at a lower general level.
footway the rural equivalent of the urban sidewalk.
freeway highest level of arterial characterised by full control of access and high design speeds.
frontage road a road adjacent and parallel to but separated from the highway for service to abutting properties and for control of access. Sometimes also referred to as a service road.
gap the elapsed time between the back of one vehicle passing a point on the road or highway and the nose of the following vehicle passing the same point. A lag is the unexpired portion of a gap, i.e. the elapsed time between the arrival of a vehicle on the minor leg of an intersection and the nose of the next vehicle on the major road crossing the path of the entering vehicle.
gore area the paved triangular area between the through lanes and the exit or entrance ramps at interchanges plus the graded areas immediately beyond the nose (off-ramp) or merging end (on-ramp).
grade the straight portion of the grade line between two successive vertical curves.
grade line the line describing the vertical alignment of the road or highway.
grade separation a crossing of two highways or roads, or a road and a railway, at different levels.
gradient the rate of slope of the grade between two adjacent vertical points of intersection (VPI), typically expressed in percentage form as the vertical rise or fall in metres/100 metres. In the direction of increasing stake value, upgrades are taken as positive and downgrades as negative.
guideline a design value establishing an approximate threshold, which should be met if considered practical. It is a recommended value whereas a standard is a prescriptive value allowing for no exceptions.

high occupancy vehicle (HOV) lane a lane designated for the exclusive use of buses and other vehicles carrying more than two passengers.
high-speed typically where speeds of 80 km/h or faster are being considered.
horizontal sight distance the sight distance dictated by lateral obstructions alongside the road and measured at the centre of the inside lane.
interchange a system of interconnecting roads (referred to as ramps) in conjunction with one or more grade separations providing for the movement of traffic between two or more roadways, which are at different levels at their crossing point.
intersection sight distance the sight distance required within the quadrants of an intersection to safely allow turning and crossing movements.
kerb concrete, often precast, element adjacent to the travelled way and used for drainage control, delineation of the pavement edge or protection of the edge of surfacing. Usually applied only in urban areas. In American practice, the term is spelt ‘curb’.
kerb ramp the treatment at intersections for gradually lowering the elevation of sidewalks to the elevation of the street surface to allow for the passage of wheelchairs or bicycles.
***K*-value** the distance over which a one per cent change in gradient takes place.
level of service (LOS) a qualitative concept, from LOS A to LOS F, which characterises acceptable degrees of congestion as perceived by drivers. Capacity is defined as being at LOS E.
low speed typically where speeds of 70 km/h or slower are being considered.
median the portion of a divided highway separating the two travelled ways for traffic in opposite directions. The median thus includes the inner shoulders.
median opening an at-grade opening in the median to allow vehicles to cross from a carriage way to the adjacent carriageway on a divided road.
modal transfer station the public facility at which passengers change from one mode of transport to another, e.g. rail to bus, passenger car to rail.
mountainous terrain longitudinal and transverse natural slopes are severe and changes in elevation abrupt. Many trucks operate at crawl speeds over substantial distances.
normal crown (NC) the typical cross-section on a tangent section of a two-lane road or four-lane undivided road with the high point located at the centreline of the road.
overpass a grade separation where the subject highway passes over an intersecting highway. A grade separation could thus be an ‘overpass‘ or an ‘underpass’ depending on which of the crossing roads is being considered.
outer separator similar to the median but located between the travelled way of the major road and the travelled way of parallel lanes serving a local function if these lanes are contained within the reserve of the major road. If they fall outside this reserve, reference is to a frontage road.
partial cloverleaf (Par-Clo) interchange an interchange with loop ramps in one, two or three (but usually only two) quadrants. A Par-Clo A interchange has the loops in advance of the structure and Par-Clo B interchange has the loops beyond the structure. A Par-Clo AB interchange has its loops on the same side of the crossing road.
passenger car equivalents (units) (PCE or PCU) a measure of the impedance offered by a (usually heavy) vehicle to the passenger cars in the traffic stream. Usually quoted as the number of passenger cars required to offer a similar level of impedance to the other cars in the stream.
passing sight distance the total length of visibility, measured from an eye height of 1.05 metres to an object height of 1.3 metres, necessary for a passenger car to overtake a slower moving vehicle. It is measured from the point at which the initial acceleration commences to the point where the overtaking vehicle is once again back in its own lane.
PC (point of curvature) beginning of horizontal curve, often referred to as the BC.

PI (point of intersection) point of intersection of two tangents.

PRC (point of reverse curvature) point where a curve in one direction is immediately followed by a curve in the opposite direction. Typically applied only to kerb lines.

PT (point of tangency) end of horizontal curve, often referred to as EC.

PVC (point of vertical curvature) the point at which a grade ends and the vertical curve begins, often also referred to as BVC.

PVI (point of vertical intersection) the point where the extensions of two grades intersect. The initials are sometimes reversed to VPI.

PVT (point of vertical tangency) the point at which the vertical curve ends and the grade begins. Also referred to as EVC.

quarter link an interchange with at-grade intersections on both highways or roads and two ramps (which could be a two-lane two-way road) located in one quadrant. Because of its appearance, also known as a jug handle Interchange.

ramp a one-way, often single-lane, road providing a link between two roads that cross each other at different levels.

relative gradient the slope of the edge of the travelled way relative to the gradeline.

reverse camber (RC) a superelevated section of roadway sloped across the entire travelled way at a rate equal to the normal camber.

reverse curve a combination of two curves in opposite directions with a short intervening tangent.

right of way the strip of land acquired by the road authority for provision of a road or highway.

road bed the extent of the road between shoulder breakpoints.

road prism the lateral extent of the earthworks.

road reserve the strip of land acquired by the road authority for provision of a road or highway. See right-of-way.

road safety audit a structured and multidisciplinary process leading to a report on the crash potential and safety performance of a length of road or highway, which report may or may not include suggested remedial measures.

roadside a general term denoting the area beyond the shoulder breakpoints.

roadway the lanes and shoulders excluding the allowance (typically 0.5 metres) for rounding of the shoulders.

rolling terrain the natural slopes consistently rise above and fall below the highway grade with, occasionally, steep slopes presenting some restrictions on highway alignment. In general, rolling terrain generates steeper gradients, causing truck speeds to be lower than those of passenger cars.

rural road or highway characterised by low-volume high-speed flows over extended distances. Usually without significant daily peaking but could display heavy seasonal peak flows.

service interchange american nomenclature describing an interchange providing access to or from a local area to an interchange. In this document, the term ‘access interchange’ is used.

shoulder usable area immediately adjacent to the travelled way provided for emergency stopping, recovery of errant vehicles and lateral support of the roadway structure.

shoulder breakpoint the hypothetical point at which the slope of the shoulder intersects the line of the fill slope. Sometimes referred to as the hinge point.

side friction (*f*) the resistance to centrifugal force keeping a vehicle in a circular path. The designated maximum side friction (fmax) represents a threshold of driver discomfort and not the point of an impending skid.

sidewalk the portion of the cross-section reserved for the use of pedestrians.

sight triangle the area in the quadrants of an intersection that must be kept clear to ensure adequate sight distance between the opposing legs of the intersection.
simple curve a curve of constant radius without entering or exiting transitions.
single point urban interchange a diamond interchange where all the legs of the interchange meet at a common point on the crossing road.
speed profile the graphical representation of the 85th percentile speed achieved along the length of the highway segment by the design vehicle.
standard a design value that may not be transgressed, e.g. an irreducible minimum or an absolute maximum. In the sense of geometric design, not to be construed as an indicator of quality, i.e. an ideal to be strived for.
stopping sight distance the sum of the distance travelled during a driver's perception/reaction time and the distance travelled thereafter while braking to a stop.
superelevation the amount of cross-slope provided on a curve to help counterbalance, in combination with side friction, the centrifugal force acting on a vehicle traversing the curve.
superelevation runoff also referred to as superelevation development. The process of rotating the outside lane from zero cross fall to reverse camber (RC), thereafter rotating both lanes to the full superelevation selected for the curve.
systems interchange interchange connecting two freeways, i.e. a node in the freeway system.
tangent the straight portion of a highway between two horizontal curves.
tangent runoff see crown runoff.
traffic composition the percentage of vehicles other than passenger cars in the traffic stream, e.g. 10 per cent trucks, 5 per cent articulated vehicles (semi-trailers), etc.
transition curve a spiral located between a tangent and a circular curve.
travelled way the lanes of the cross-section. The travelled way excludes the shoulders.
trumpet interchange a three-legged interchange containing a loop ramp and a directional ramp, creating between them the appearance of the bell of a trumpet.
turning roadway channelised turn lane at an at-grade intersection.
turning template a graphic representation of a design vehicle's turning path for various angles of turn. If the template includes the paths of the outer front and inner rear points of the vehicle, reference is to the swept path of the vehicle.
underpass a grade separation where the subject highway passes under an intersecting highway.
urban road or highway characterised by high traffic volumes moving at relatively low speeds and pronounced peak or tidal flows. Usually within an urban area but may also be a link traversing an unbuilt-up area between two adjacent urban areas, hence displaying urban operational characteristics.
value engineering a management technique in which intensive study of a project seeks to achieve the best functional balance among cost, reliability and performance.
verge the area between the edge of the road prism and the reserve (right of way) boundary.
warrant a guideline value indicating whether or not a facility should be provided. For example, a warrant for signalisation of an intersection would include the traffic volumes that should be exceeded before signalisation is considered as a traffic control option. Note that, once the warranting threshold has been met, this is an indication that the design treatment should be considered and evaluated and not that the design treatment is automatically required.
white line break point a point where a sharp change of direction of the white edge line demarcating the travelled way edge takes place. Usually employed to highlight the presence of the start of a taper from the through lane at an interchange. In countries where the edge of the carriageway is demarcated by yellow lines, the equivalent is the 'yellow line break point'.

yellow line break point a point where a sharp change of direction of the yellow edge line demarcating the travelled way edge takes place. Usually employed to highlight the presence of the start of a taper from the through lane at an interchange. In countries where the edge of the carriageway is demarcated by white lines (e.g. the United States), the equivalent is the 'white line break point'.

Part I

Geometric design

Chapter 1

What a geometric designer needs to know

CONTENTS

1.1 INTRODUCTION

Geometric design is defined as the design of the visible elements of the road. In effect, the geometric designer is the architect of the road. As such, geometric design includes a fair measure of art amongst the science. It follows that designers should be sensitive to the perceptions of the people who are using the road, whether they are drivers or passengers in a vehicle or outside vehicles. If they are outside the confines of a vehicle, people may be pedestrians heading for a specific destination, relaxing at pavement cafes or perhaps window shopping. They may also be members of the community through which the road passes and this could include children playing in the street. Whatever the reason for their being either in the road reserve or adjacent to it, they have the fullest right to be there. To develop a design that will balance all these conflicting needs, it is necessary for the designer to go beyond

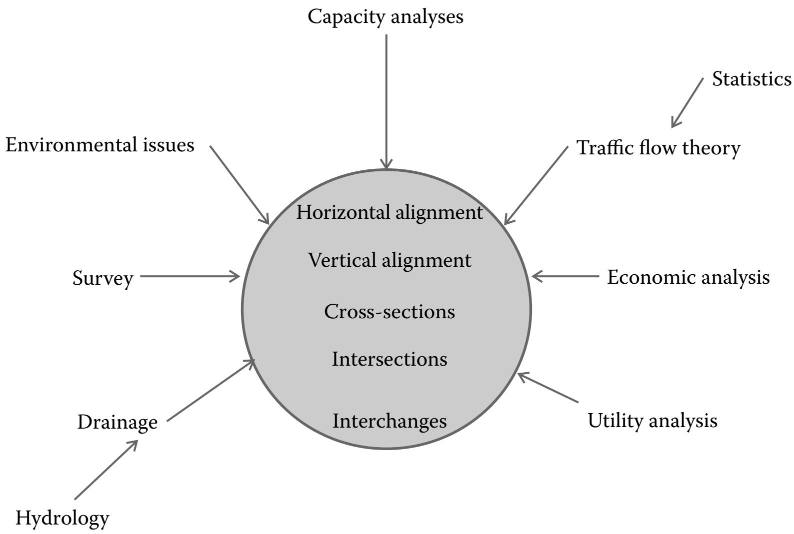


Figure 1.1 The structure of geometric design. (From Wolhuter KM. *Lecture T1: Introduction. Geometric design course*. South African Road Federation, Johannesburg, 2003a.)

merely seeking to ensure that vehicles can travel between origin and destination with the minimum of impedance to their passage.

Engineers entering the field of geometric design sometimes believe that it can be successfully pursued simply by following whatever directives are provided in their client's geometric design manual. This belief is remote from the truth because it does not go far enough. A successful design will result in a road that provides safety, accessibility, mobility and convenience to the movement of people and goods with a minimum of side effects. In addition, it should be visually pleasing regardless of the point from where it is observed, whether from within a vehicle or outside it. In short, the road should blend with its environment to the extent that it should look as though it always was there and that the landscape would have been the poorer without it.

The intention of this book is to guide designers towards an understanding of

- The philosophies behind geometric design
- The techniques whereby acceptable designs can be achieved
- The numerous disciplines that underpin geometric design

The overall structure of the field is illustrated in Figure 1.1.

1.2 HOW TO USE THIS BOOK

Fiction is intended to be read from start to finish, as these books have plots, character development and story lines. *Geometric Design of Roads Handbook* is not intended to be read in this fashion. It is hoped that, when a designer is confronted by a problem outside of his or her experience, reference to the relevant section of the Handbook would either suggest an answer to the problem or, at least, direct the designer towards some other source of information.

If the Handbook is, however, read in the manner of a work of fiction, it will be noted that comments made in one chapter may be repeated in several subsequent chapters and sometimes more than once in a chapter. This is intentional as reference to one chapter or even only a portion of one should not require searching elsewhere for additional information.

1.3 SAFETY AND GEOMETRIC STANDARDS

There is a widespread belief, not only amongst designers but also in courts of law, that a road designed within the limits of prescribed geometric standards is safe. If this were true, there would never be any crashes on the road and nobody would ever be killed or hurt. Obviously, this is not the case. A road can usually be made safer or it may become less safe but it can never be made safe. In short, road safety is a relative rather than an absolute construct.

Before pinning blind trust on any given geometric standards, it would be useful for the designer to have an understanding of their background. Some are based on recent and rigorous research, whereas others are based on historical research that cannot take account of improvements in vehicle and tyre design over the last 50 or more years. Some are based on professional experience and judgment in the belief that they have proved to work in practice. Crash rates may, however, demonstrate that this belief is sometimes unfounded.

Black spots occur on roads on which every element conforms to standards and suggest that many drivers are making similar mistakes at the same places. It is possible that the road design has inadvertently created a misleading impression and that the ‘driver error’, which is so often assumed to play a role in the majority of crashes, is, in fact, a design error.

The ‘low-volume’ road typically has a gravel wearing course and this generates a new list of problems. Not least of which is the fact that, in many countries worldwide, the majority of the road network is either gravelled or unsurfaced. The skid resistance, tangentially and radially, of a gravel surface is lower than that of a bituminised or concrete surface. It is susceptible to scour, which places a limitation on the maximum acceptable gradient and superelevation. This suggests that a gravel road should have higher geometric standards than those applied to a surfaced road. However, normal practice often is to design a low volume road, which will in the foreseeable future be surfaced, to the standards of a surfaced road and it is suggested that this may not necessarily be wise.

The low-volume road requires regular replacement of the gravel wearing course, and the availability of suitable material within an acceptable distance of the route is also an issue.

1.4 INTERNATIONAL NOMENCLATURE

Possibly because the Transportation Research Board, based in Washington, DC, is the most prolific publisher of road research worldwide, international terminology tends to follow American practice. That will also be the practice adopted in this book.

Britain, however, also has an influence on terminology and this may be attributed to its historical colonial role. The most obvious illustration of this role is to be found in the names of the various parts of the passenger car. Some examples are offered in Table 1.1.

Geometric design also has its fair share of differences in nomenclature other than the four-legged intersection of the United States finding its counterpart in the British four-armed junction as referred to in Table 1.1. The American painted island is reflected in the British ghost island. A median created in an American two-lane two-way road as part of the layout of an intersection is, in the United Kingdom, single dualling. What is referred to in the United States as the ‘right of way’ is referred to in the United Kingdom as the ‘road

Table 1.1 Differences in nomenclature

USA	*United Kingdom*
Right of way	Road reserve
Intersection	Junction
Four-legged	Four-armed
Fender	Bumper
Horn	Hooter
Hood	Bonnet
Trunk	Boot

Source: Wolhuter KM. *Lecture T1: Introduction. Geometric design course*. South African Road Federation, Johannesburg, 2003a.

reserve'. South African practice is also to use the term 'road reserve' with 'right of way', being a legal term referring to drivers yielding right of way to vehicles approaching from the right at a roundabout. (South Africa and the other commonwealth countries all drive on the left.) Right of way applies also to at-grade railway crossings in favour of the train.

Components of the cross-section are also referred to by various names, some of which may be regional or only local. For example, the outer separator, which is the island between a major road and the adjacent parallel service road (also referred to as a frontage road), has at least 27 different names in the United States, varying by state. It is trusted that designers will be aware of the nomenclature in use in the area for which they are designing and, furthermore, that they will also be familiar with general American usage.

1.5 THE BROADER APPROACH TO DESIGN

1.5.1 Human factors

With the notable exception of tourism, transportation is not a demand of itself. It is a derived demand as discussed in Chapter 2. As such, the needs of the people served by the road must take precedence above all else. In the case of people in vehicles, human factors have a major role in the selection of the values of the various parameters employed in the design. These include

- Basic lanes
- Lane balance
- Application of auxiliary lanes
- Ramp spacing
- Decision sight distance
- All right exits and entrances (left in British Commonwealth countries and Japan)
- Single exit per interchange
- Exit in advance of the crossroad
- Simplified signing

Human factors were previously limited to driver eye-height and reaction time. It has now become desirable to consider additional characteristics of the driver such as

- Age
- Driver behaviour, whether impaired or not

- Decision making
- Desired rates of deceleration as opposed to panic stop deceleration
- Estimation of speed:
  - In relation to that of approaching vehicles
  - Safe speeds for the negotiation of horizontal curves
- Expectancy
- Gap acceptance
- Information processing
- Response to workload
- Visual perception

Without cognisance being taken of these factors, it is possible that a design that is unacceptable in terms of safety may result.

The concept of derived demand also has a major impact on the philosophy brought to bear on geometric design. This is discussed in Chapter 2 in relation to congestion pricing. This is a useful tool for assessing the efficiency of a road networking in addition to the determination of the cost of levied charges or tolls on vehicles entering areas that are already congested.

1.5.2 Design consistency

Consistency of design seeks to avoid the situation of a driver being surprised by some or other unexpected feature of the road. Surprise may cause the driver to carry out an unpremeditated action. Unpremeditated actions, such as a violent swerve, tend to be more extreme than the circumstances warrant, possibly leading to loss of control of the vehicle.

Previously, consistency of design was sought through application of the principle of design speed. Unfortunately, the design speed simply defines the lowest value of any design criterion found along the length of the road. Two long tangents connected by a 60 km/h radius curve could quite correctly and yet totally misleadingly be described as having a design speed of 60 km/h.

The next development was to use the operating speed, defined as the speed not exceeded by 85 per cent of all drivers, as the basis for achieving consistency of design. Until research had been carried out, it had been assumed that the operating speed was of the order of 85 per cent of the design speed. This new lower speed was adopted for design, albeit briefly. It was then found that, under conditions of long tangents and gentle radii of curvature, operating speed could actually be higher than design speed by anything up to 10 or even 20 km/h. In consequence, operating speed regained respectability as a measure of design consistency.

Further research was then undertaken to predict operating speeds on curves of a range of radii. Subsequently, various methods have been developed to quantify consistency of design in terms of variations in speed along the length of the road, alignment indices or driver workload. It is now possible to assess the consistency of design objectively as opposed to reliance on subjective judgement. Obviously, the professional judgement of the skilled designer will always play a role in such assessments but objective techniques, properly used, are a powerful tool.

Consistency of design includes the concept of the ‘self-Explaining Road’. This concept defines the means whereby the designer seeks to inform the driver of what behaviour is expected and supported by the design of the road. It seeks to create harmony between all the elements of the road in the sense, for example, that a wide cross-section implying safety at high speeds should not be superimposed on a horizontal alignment where curve radii are small.

1.5.3 Context-sensitive design

As discussed later, design of roads in urban areas differs completely from that in rural areas. Urban areas include a wide range of vehicle types moving in high volumes and relatively low speeds, whereas in rural areas the differences in vehicle types becomes less critical, speeds tend to be high and volumes low. Within these two areas, many subareas also present themselves such as

- Urban centres
- Urban corridors
- Suburban corridors and nodes
- Industrial corridors
- Residential areas
- Rural town centres
- Transitional areas
- Rural connecting corridors

There are significant differences between these subareas in the types and numbers of vehicles present. Traffic patterns and speeds differ and the needs of the people in each are also going to differ.

In the case of people outside vehicles, it is necessary to consider their reasons for being in or adjacent to the road reserve and to create an environment acceptable to them. Context-sensitive design is thus an issue not only for the drivers of vehicles but for others as well.

Community involvement or public participation in the development of the design from its earliest stages is important for the achievement of a truly context-sensitive design.

1.5.4 The design domain

It may not always be possible to match or exceed some or other prescribed value of a criterion. This could be because topographic or economic or any other constraints and the concept of the design domain and its corollary, the extended design domain, have recently come into play. The problem with venturing into the extended domain is that of legal liability. The further away from the normal design domain, the greater is the possibility and extent of legal liability. The design exception and the basis on which the decision was taken to depart from normally acceptable standards needs to be recorded in detail to be able to demonstrate that the designer and the client have 'exercised their minds' in arriving at whatever value of the criterion was applied and what factors were employed to mitigate its effect on safety.

1.5.5 The design team

As illustrated in Figure 1.1, geometric design is not a process that functions in isolation. There are many disciplines that impact on it. These are introduced under the following subheading and discussed in more detail in Part B. It would be unreasonable to expect the designer to be expert in all of them, but there should, at least, be some understanding of these disciplines and the role that they play in the determination of a successful design.

It follows that geometric designers should realise that they are members of a team that includes

- Town and/or regional planners
- Environmentalists (bearing in mind that this classification includes numerous disciplines from archaeology to zoology)
- Safety specialists
- Representatives of the community

In special circumstances, it may also have to include experts in the fields of

- Sociology
- Education
- Public health
- History, and so forth

1.6 SUPPORTING DISCIPLINES

1.6.1 Introduction

As illustrated in Figure 1.1, there are many disciplines in addition to geometric design itself that impact on the final design that is to be built.

In countries that are well-endowed with wide-ranging professional skills, the geometric designer can focus on his or her field of interest with the ancillary information being provided by others. It would, if only in order to be able to ask the right questions and then to interpret the answers' impact on the design be necessary to have some knowledge of the supporting disciplines. In other countries that are less fortunate, it may be necessary for the geometric designer to become involved in these disciplines as well.

These supporting disciplines are discussed further below.

1.6.2 Systems analysis

All the geometric designer is really doing is seeking an answer to a problem. The first step, thus, is to define the problem. It is remarkable how often a design is produced without much thought being given to the underlying problem. The client may suggest that the problem is that a road is required between two points. If this is the brief provided to the designer, it is to be trusted that the client had at least undertaken a process that led to this conclusion being drawn. Unfortunately, this may not always prove to have been the case.

For example, industries may be set up in areas where their input materials are at hand. On the other hand, if their processes are labour intensive, it may be worthwhile for them to be located adjacent to their potential labour force. As a further alternative, it may financially be more attractive if they were to be located in close proximity to their markets.

These three alternatives result in totally different transportation problems to be resolved. Should the workers be required to commute to the industrial area? If so, should this be by public transport and, if so, should the mode be rail or road? If the industries concerned are to be relocated to the location of the potential source of labour, this would require the movement of materials rather than of people. Raw materials may have to be shipped in and finished products shipped out. Rail transport may thus be more suitable than transport by road. The third alternative would require the movement of both people and raw materials, whereas the final movement of products to the market would be relatively straightforward.

Systems analysis needs to be brought to bear on the transportation problem to be solved. It comprises a hierarchy of steps, as follows.

- Problem definition
- Identification of possible solutions
- Selection of a solution appropriate to the problem being addressed
- Implementation of the proposed solution
- Monitoring the outcome of this implementation

Implementation may lead to the emergence of a new problem or monitoring to identify one previously hidden by the old problem. This completes the loop leading to a further round of systems analysis. Without knowledge of systems analysis and the ability to apply its principles, the likelihood of a successful design being developed is reduced.

1.6.3 Project analysis

One of the steps in the systems analysis process is the selection of the problem solution to be implemented. It is a fact of engineering life that problems having only one unique solution do not exist outside the classroom. There could be many solutions that are patently wrong or unimplementable but, equally, there could be many solutions that, to some greater or lesser extent, address the problem. However many adequate solutions are found, only one can be implemented. Which then is the most adequate solution? Economic analysis offers one way of comparing competing solutions and utility analysis another. Without these analytical tools, the application of systems analysis would not be possible.

Economic analysis involves a determination of the economic benefits accruing from providing the road in comparison with the cost of providing it. Three methodologies could be brought to bear on this form of analysis.

- Rate of return
- The benefit/cost ratio
- Net present worth

The first mentioned would be employed by a national or state government that has to provide a variety of competing services to the population, such as education, health care and transportation, ensuring an equitable allocation of its limited budget. It would also be used by transport economists in allocating financial resource to competing modes of transport or to competing roads such as between A and B vis-à-vis between C and D.

Geometric designers, however, typically have to decide between competing routes that provide the same service, that is, only one of which is going to be selected for design and construction. Either the benefit/cost ratio or the net present worth method could be used for this purpose. Regardless of which method is used, the ultimately ranking of the competing routes would be identical.

Design and construction costs are incurred during a relatively short period of time whereas maintenance and road user costs accrue over the lifetime of the road. In addition to being familiar with the two possible forms of analysis, it is necessary to be aware of the process of discounting future costs to a common baseline year.

Utility analysis comes into play where roads have a social as opposed to an economic value. Roads that serve as links to schools, hospitals or recreational areas would fall into this category. Utility analysis provides a formalised allocation of nonmonetary values to each of the facilities served by the various possible routes under consideration. Typically, roads have a combination of economic worth and social value. Both would have to be borne in mind in the final decision of the route to be designed and built.

1.6.4 Transportation planning

Convenience simply means that people can move between origin and destination without having to suffer inordinate levels of congestion and correspondingly long travel times, which include having to cope with over-saturated intersections. Questions that have to be answered are thus:

- 'How many people would require or wish to move from A to B, C to D, etc.?'
- 'Where are A, B, C, D, etc. relative to each other?'
- 'What would the preferred mode of travel be – car, bus, train etc?'
- 'What would the average occupancy of each vehicle on the network be?'

Origin–destination surveys and other high-level predictive techniques will ultimately provide the geometric designer with information on traffic volumes that have to be accommodated on the roads in the network including the road of immediate interest. These techniques fall in the realms of transportation planning. The geometric designer should have some understanding of transportation planning and the techniques it employs to verify the validity of its findings and its application to road geometry.

1.6.5 Capacity analysis

Roads are never designed to accommodate only the traffic that is anticipated to currently use it. They have to provide some or other acceptable level of service in a future year, typically at the end of the design life of the road. The cross-section of the road is predicated on the volume of traffic anticipated for that future year. The cross-section can be anything from a multi-lane freeway to a two-lane two-way road and, in very limited circumstances, may be only one lane wide with passing bays. It is necessary for the designer to carry out capacity and level of service analyses to determine the cross-section that has to be provided.

Furthermore, roads carrying large volumes of traffic moving at high speeds generally have higher geometric standards than those applied to lower order roads. Level of service and speed inputs thus also have a bearing on more than just the cross-section.

Because urban intersections are usually closely spaced, the efficiency of urban networks is dependent on the capacity and level of service provided by these intersections. A turning vehicle can cause considerable delay to the vehicles following it so that, even where only one vehicle wishes to turn across opposing traffic, it may be advisable to provide a turning lane. The determination of the required length of turning lane becomes an application of traffic flow theory.

1.7 TRAFFIC FLOW THEORY

Traffic flow theory is based on probability. For example, at an intersection, the probability that a gap of sufficient length will appear within a reasonable period of time to enable a vehicle either to turn into or cross an opposing stream of traffic needs to be determined. It is necessary to know what constitutes a 'a gap of sufficient length', that is, a gap that a driver will accept. It is also necessary to know what a 'reasonable period of time' is. Armed with this *a priori* knowledge derived from application of human factors to the design process, the designer will be able to establish for the project in question whether or not adequate gaps will occur with acceptable frequency and, hence, what form of traffic control is required to ensure that the intersection functions as intended.

To carry out such exercises in traffic flow theory, the designer requires a fair knowledge of the application of statistics and statistical distributions.

1.8 ENVIRONMENTAL ISSUES

The geometric designer should be aware of the aesthetic quality of the road as ultimately built. This applies to the ever-changing vista presented to the occupants of moving vehicles. While the road itself would be a prominent component of what is seen, the landscape also forms part of the view. Aesthetics also applies to people who have to live with the road as part of their unchanging view of the area in which they live. An irretrievably scarred hillside would definitely not find favour in the eyes of the local community.

More important than the appearance of the road is the fact that it can be a danger to the health of surrounding communities. Air pollution caused by the exhaust gases emitted by moving vehicles is a distinct health hazard. Nitrogen oxides and sulphur dioxide cause damage to the lungs, and the latter is also responsible for acid rain.

Gravel roads serving farming areas may be traversing areas where fruit and vegetables are grown for the market. These roads could thus present an intolerable nuisance in terms of dust, both from moving vehicles and from maintenance operations.

Noise affects a hearer by

- Causing annoyance
- Interfering with conversation and human performance
- Sleep

It thus produces unacceptable physiological and sociological side effects.

1.9 HYDROLOGY AND HYDRAULICS

Water can provide one of the major constraints on the longevity of any road. Storms of an unanticipated magnitude may cause bridges and culverts to wash away or the approach fills to collapse. They can also lead to mud slides either blocking the road or causing it to slip down the hill side. It is necessary to consider these possibilities in the design of the road and to provide measures suitable to their elimination. This requires knowledge of hydrology and hydraulics.

Hydrology is concerned with the prediction of rainfall, return periods of storms and their associated intensity and runoff. This leads to prediction of the quantity of water that has to be transported either along or across the road reserve. Hydraulics addresses pipe and channel flow and thus defines the facility that has to be provided to accommodate the anticipated flow.

The geometric designer requires some knowledge of these disciplines in support of appropriate bridge and culvert sizing. Sizing and possibly the angle of skew will have an impact on the vertical alignment of the road and may even force a change in the horizontal alignment. Longitudinal drainage within the road reserve is accommodated in channels. Analysis of the flows in channels will define not only the dimensions of the required channel but also its minimum gradient for self-cleansing flow speed or the maximum gradient at which scour will not occur.

1.10 SURVEY

Having completed the design, it is necessary to convert lines on paper to lines on the earth's surface. While the actual setting out of the works would be the work of a surveyor, it would be necessary for the designer to define the route as located in order to brief the surveyor properly. All that is required is the ability to carry out coordinate calculations, usually only a series of joins and polars and, possibly, triangulation. In the case of interchange ramps, defining their coordinates can be a problem that the geometric designer has to be able to resolve.

1.11 URBAN VERSUS RURAL GEOMETRY

There is a significant difference between urban and rural geometry. The idea that a car takes *x* metres to stop from *y* km/h, regardless of where it is, is perfectly true but misses the point completely.

What is typically found in the rural environment is relatively low volumes of vehicles moving at high speed for considerable distances whereas the urban environment is characterised by high volumes moving at low speeds over relatively short distances. High speeds require high values of horizontal and vertical curvature and relatively flat gradients. In consequence, cuts and fills several metres high can occur in the rural environment. The lower speeds prevailing in urban areas mean that it is possible to get the vertical alignment closer to ground level. Ideally, the urban vertical alignment is just slightly below ground level to enable draining the surrounding properties towards the road reserve. Any significant departure from ground level also creates problems with regard to access between adjacent properties and the road reserve.

Intersections in the rural environment are spaced at anything up to kilometres apart and their effect on the smooth flow of traffic is minimal. Urban intersections are very closely spaced and could be as little as 100 m apart. Their effect on traffic flow is significant. In fact, it is the intersection that dictates the efficiency of the urban network. A single vehicle turning left at a rural intersection could, in all probability, execute the turn the moment the intersection is reached and following traffic – if any – would not even have to slow down to avoid the turning vehicle. On the other hand, the same single vehicle in an urban environment could cause a considerable backup queue to generate while awaiting a gap in the opposing traffic stream.

Rural intersections are typically no more than priority-controlled bell mouths between opposing two-lane two-way roads. Their urban counterparts could be multi-lane layouts with dedicated turning lanes, median and other islands, pedestrian cross walks and with a high level of sophistication brought to bear on their signalisation.

Competing activities, that is, activities other than vehicle movement that also require a share of the road space, are very low in the rural areas. Whether in a vehicle or on foot, anybody on the road is going somewhere. In the urban area on the other hand, there is a very high proportion of competing activities and modes of transport. It is this very diversity of activity that gives the urban environment its rich character. Ignoring it in the name of mobility will almost certainly result in the production of an environment that is dull, uninteresting and, in fact, sterile.

In the rural environment, differences between the various vehicles to be found on the roads are, very generally, not critical. Cars, intercity buses and trucks all move at about the same pace and there is little impediment to overtaking. The only time that this does not apply is during holiday seasons when, on certain rural roads, volumes become so high that traffic is brought to a complete standstill. In the urban environment, simply because vehicles are so much closer together, physical dimensions start to play a role. An articulated vehicle takes up as much space as about three passenger cars. Its ability to accelerate from a stopped condition is significantly slower than that of a passenger car. A left turning bus or truck very often requires the active assistance of opposing vehicles to complete the manoeuvre. The urban designer must be more sensitive to the operational differences between vehicles than his or her rural counterpart.

Congestion simply means that there are more vehicles on the road than it can comfortably accommodate. Historically, this problem was addressed by providing more infrastructure. In terms of the supply/demand equation, the focus was on supply. It is, however, also necessary to consider demand. For an unaltered number of person trips, vehicle trips could be slashed by replacing them with bus travel. One bus could take the place of a queue of passenger cars that is three-quarters of a kilometre long. It has often been said at high political levels that 80 per cent of passenger trips should be by bus and that this goal should be achieved in the very near future.

While both the percentage and the time frame could be queried, the urban geometric designer is required to produce street networks supportive of public transport. This does not only mean the provision of bus lanes and careful consideration of the location of bus stops. It also means that pedestrians should not have to walk too far to get to the nearest stop. Obviously, this is going to impact on the layout both of the residential area and the destination area. Public transport requires high population densities to be economically viable. The alternative is for a bus to wind its way for several kilometres through a low-density residential suburb to acquire its full load of passengers. This is another consideration that has to be brought to bear on the design of the residential area. Using the bus as the design vehicle for a certain road would also suggest that maximum gradients and minimum curve radii are going to become critical features of its location. The location and layout of termini will require careful attention. The rural designer is not beset by this problem.

In the rural environment, storm water drainage usually involves removing excess water from the road and onto the surrounding area as soon as possible and providing culverts or bridges to accommodate cross flows that are already concentrated in watercourses. In the urban environment, the road reserve is, in fact, the conduit for storm water being transported from surrounding areas. However, because of the higher speeds found in rural areas, water on the road surface is potentially more hazardous and liable to cause aquaplaning than in urban areas, where water on the road surface tends more to be just a nuisance, although wet urban roads tend to produce large numbers of tail-end accidents. The approach to storm water drainage design differs totally between the rural and the urban environment. This includes, in the urban area, drawing a distinction between major and minor storms.

A further complication in urban design is that provision must be made for numerous other facilities, specifically facilities not usually found in rural areas. Sewerage, water and power reticulation, street lighting, telephone lines and, in older city areas, gas lines are usually located in the road reserve. Sewerage and storm water drains operate on gravity and they can have a serious impact on the location of the road reserve in which they are to be accommodated.

Land acquisition costs are typically 20 per cent of the total cost of provision of a rural road with construction accounting for the other 80 per cent. In the urban environment, the proportion is reversed in spite of the fact that construction is more expensive in urban than in rural areas. This is because land acquisition typically involves developed sites up to and including multi-storey business premises in addition to the fact that urban land prices even for undeveloped properties are significantly higher than in rural areas. The cost of construction in urban areas is also higher than in rural areas. This is because material that has to be dumped to spoil or imported for the design layers has to be carted for longer distances than in the rural situation. Even if the haulage distances were similar, travel times are longer in the urban area. Construction sites are more constricted in the urban area and provision for bypasses can be difficult. Unit prices thus tend to be higher for urban than for rural construction. However, this is not sufficient to change the percentage differences between costs of land acquisition and construction.

1.12 CONCLUSION

There are various definitions in the literature of the goals of transportation. The following is a fairly standard definition:

> Transportation is directed towards the accessible, convenient and economical movement of people and goods with a minimum of environmental side effects and within a concern for the safety of road users, whether within or outside vehicles, and the safety of the community as a whole.

Very simply, the objective of geometric design is to give practical expression to these higher goals.

Traditionally, road users were tacitly understood to be people in cars and design addressed their needs largely to the exclusion of all else. It has now been realised that road users include everybody within and adjacent to the road reserve, *regardless of why they are there*. They may be in the road reserve as pedestrians or cyclists with a set destination in mind or simply walking, jogging or riding for exercise, out on the street taking a break from the office, meeting friends at sidewalk cafes or window shopping. The reasons for people being in the road reserve are probably endless and these reasons are going to dictate their behaviour patterns while in the road reserve. There are also a considerable number of people who live or work in close proximity to the road and whose needs differ from those of the people actually within the road reserve itself. It is necessary for the geometric designer to achieve a design that effects an acceptable trade-off between all these conflicting demands and supports the creation of a sense of place, specifically of one where people would wish to live and work and play.

The person at the drawing board must thus go far beyond knowledge and application of laid down geometric standards. An attitude of caring, encompassing the road user, the community and the environment in which it lives, is required. This attitude finds expression, amongst other things, in

- The concepts of human factors, consistency of design and public involvement referred in Section 1.1
- The need for design to be by a team rather than as a solo effort
- Consideration of the framework of supporting disciplines and their involvement in underpinning successful design as illustrated in Figure 1.1

This does not mean that a bleeding heart is a fundamental requirement for good design. In reverting to the introductory comment on the objectives of transportation, the point is stressed that the designer must be a realist. In addition to an attitude of caring, designers must also bring a host of technical skills to bear on problem definition and solution in their chosen field.

The intention of this book is to set readers on the path towards acquiring those skills required to give sensible expression of a sound philosophy of design.

Chapter 2

Policy

CONTENTS

2.1 INTRODUCTION

There seems to be a mystique attached to policy that is totally unjustified. In fact, a policy could simply be a desired end state, for example, 'Our policy is that, in 2015, we want the situation in this Department to be positioned to…'. Alternatively, a policy could be a statement of intent. 'We do not have sufficient to gravel to maintain our shoulders. It is our policy, therefore, to provide all roads with surfaced shoulders'. Finally, a policy could simply be a statement of preferences: 'All our roads are going to have a design speed of 120 km/h'.

The formulation of policy leads into another activity, also seemingly shrouded in mystery. This is strategy development. In its simplest form, strategy is the enunciation of a series of actions aimed at moving from a current situation towards a desired end state. 'This where we are now and this is what we must do to achieve the desired end state'.

According to the *Concise Oxford English Dictionary*, a policy is defined as 'a course or principle of action adopted or proposed by an organisation or individual'. The definition offered by *The Merriam-Webster Dictionary* is that it is 'a definite course or method of action selected from among alternatives and in light of given conditions to guide and determine present and future decisions'. What these dictionaries are offering as definitions of policy are, in fact, definitions of strategy. It is presumed that they see strategy as part of policy. The common denominator of these definitions of policy is that they are all aimed at providing a uniform approach to problems that may be encountered in the future.

2.2 LAND USE POLICIES

Town planning was previously focussed on somewhat sterile grid layouts dedicated to single usage. Areas could thus be planned to have a residential usage, or be oriented towards commercial or industrial applications. The fact of the matter was that everybody had to move from home to job and back again in the evening. The proliferation of the passenger car resulted in many people being able to enjoy the convenience of using their own car to get to work. Home was generally in some outlying residential suburb and the workplace was either in an industrial area or in an office in the central business district (CBD).

Congestion, air pollution, rat-running and noise were inevitable accompaniments to the single-usage township layout. A further consequence of this layout was that people relied on their cars to get anywhere and did not exercise much, if at all, resulting in a population that included 25 per cent obese people, many of whom were morbidly obese.

More recently, a cellular policy of township layout has emerged and the single-usage layout largely abandoned. Basically, the 'cell' would be surrounded by arterials and have relatively few accesses to the arterial network, with no portion of the internal road system serving as a short cut to anywhere. In the case of existing grid layouts it would thus be easy to eliminate rat-running through the community altogether. The land usage inside the cell could include high- and low-density residential accommodation, local retail outlets, possibly light industries and other employment opportunities, schools, churches and recreational areas. In a cellular multi-usage township the 'home above the shop' is a distinct possibility whereas the single-usage policy precluded this.

A cellular township layout is illustrated in Figure 2.1.

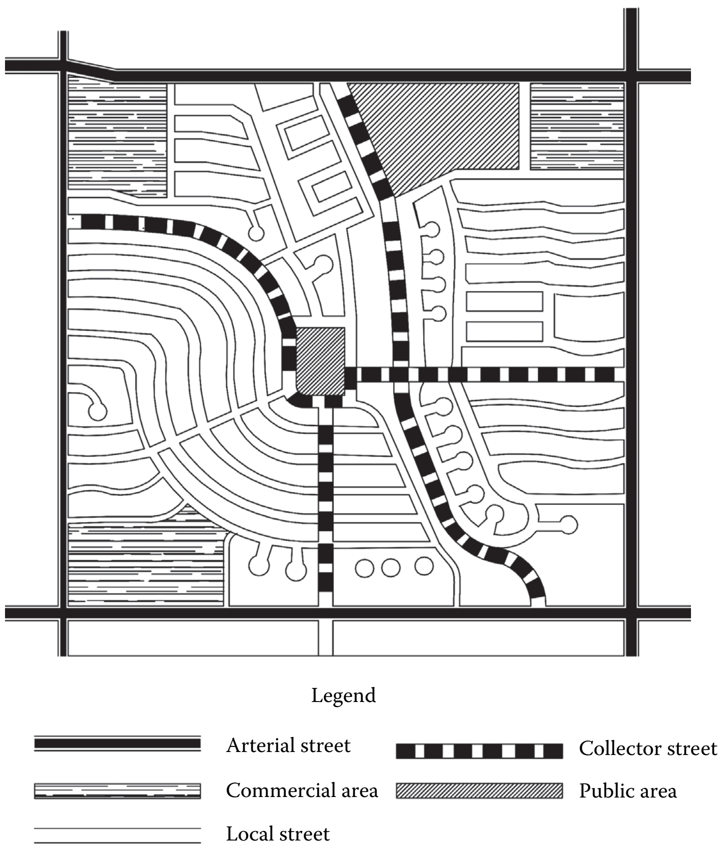


Figure 2.1 Typical cellular township. (From American Association of State Highway and Transportation Officials. *A policy of geometric design of highways and streets.* Washington, DC, 2011a.)

Many people would be able to walk or cycle to work within the cell area, thus getting healthy exercise while simultaneously reducing the pollution caused by hydrocarbons. The employment opportunities in the township would have an obvious although possibly restricted impact on the traffic on the arterials. One of the more obvious benefits of the cellular policy is that it would engender a strong community spirit.

In the case of existing grid layouts, the application of the cell concept is achieved through the application of traffic calming. Traffic calming is a retrofit application to eliminate rat-running, amongst other things. It is aimed at reducing both the speed and the volume of traffic of traffic in a residential area. This is briefly discussed in Chapter 13.

An important aspect of land use is that it should seek to minimise the footprint of the road, which is the extent of land required for the provision of the road. This is required to ensure that land that could have been used for other profitable purposes isn't sterilised by incorporation into the road reserve. An example is the insistence on an overly generously sized median island. Examples exist whereby the Transportation Authority demanded and got a median island with a width of 20 metres on the basis of the distance that an out-of-control vehicle would travel beyond the boundary of the travelled way. The assumption made was that this vehicle would not travel further than about 10 metres away from the travelled way. If two vehicles travelling in opposite direction were to cross the median simultaneously in an out-of control fashion such that they could collide, the likelihood of a fatal crash would be high. Hence the need for the 20-metre width!

Obviously, the likelihood of the precise location of the two vehicles on departing from their carriageway and their subsequent paths being such that a crash will occur is vanishingly small. A 20-metre wide median that is, say, 50 kilometres long occupies an area of 100 hectares forever removed from economic activity.

The need to accommodate the desired cross-section and ancillary features such as drainage of the road and widening of the road reserve to accommodate intersections or interchanges as well as space for utilities must be the basis of the determination of the footprint of the road.

Having established the extent of this footprint, the land required has to be acquired by the Transportation Department. This is often by an agreed on price on the basis of a willing seller/willing buyer process. Unfortunately it also happens that agreement on the purchase price sometimes cannot be achieved, usually because the value that the seller attaches to the land to be acquired is unrealistic. It would then be necessary to go to a process of expropriation whereby the seller is forced to accept the purchase price offered. Normal practice is that the price offered is market related. In some countries, the acquired land is legally transferred into the name of the Transportation Authority through a deed of sale or equivalent mechanism, whereas in others the acquired land is not transferred but remains the property of the owner even though he or she does not derive any benefit from ownership of the expropriated land. Some countries levy taxes to finance the upkeep of the local road system, with these taxes being based on the area of the owned property. In these cases, not taking transfer of the land would be manifestly unfair.

The creation of the road reserve may result in portions of the property being alienated from the remainder of the property, for example, without access from the remainder and being too small to be operated as an economic unit. The alienated portion should then be incorporated in the footprint of the road reserve.

2.3 POLICIES RELATING TO NETWORK PLANNING

Network planning is closely related to land use planning and, in fact, tends to follow it. The proper location and mix of land uses determines to a large extent the efficiency of urban areas in the sense of their ‘liveability’. In short, are they such that people would wish to live, work and play there? This addresses many issues such as

- The noise generated by passing traffic
- The pollution of the air by vehicular exhaust emissions
- The safety of people, particularly children, when within or in close proximity to roads
- Rat-running of commuters bypassing congested arterials by finding routes through adjacent residential areas

The accessibility of areas is an important aspect of liveability, particularly in the sense of journey times, especially commuter travel times, and ease of use of the network. It has long been realised that the provision of additional infrastructure as a means of reducing congestion is simply not effective and that it is necessary to have planning policies supportive of public transport. To illustrate: A bus that transports up to 80 or more passengers occupies significantly less space than the equivalent 60 or more passenger cars required to transport them and hence contributes to a far lesser extent to congestion. Unfortunately, public transport cannot easily match the convenience of a passenger car in terms of availability, comfort and freedom of route selection. Strategies to make public transport more attractive to commuters are discussed in Chapter 15. These include the sale of tickets at stations as opposed to buying them from the bus driver, and through ticketing in addition to the formulation of policies relating to the planning of the network and design of the buses.

Planning is concerned with broad regional issues where town and regional planners and the local communities play a role in ensuring that the links in the road network actually serve the communities through which they pass. Policy decisions are captured in a set of principles that define the manner in which a Transportation Department relates to other government departments, local land owners and communities to properly manage the network that forms their remit.

Policy requires that planning be initiated by an in-depth analysis of the needs of the inhabitants of the area through which the road is intended to pass. Furthermore, the policy usually will dictate that this analysis includes involvement of the communities concerned. This will lead to the definition of a corridor that respects community values and sensitivities. For example, locating a road through a cemetery should only be the absolute last resort before all other alternatives have been investigated and rejected.

The government departments with which Transportation Department interacts in the planning of networks are mainly those at all levels of government with a brief addressing some or other aspect of the environment or service to communities. These include

- Land use including township development, farming activities and mining
- Environmental planning, including provision for migrating animals and avoidance of habitat fragmentation, game farms and timber plantations, and also control of riverine and air pollution
- Water supply including provision of dams and reservoirs, water purification, and potable water reticulation
- Telecommunications, principally the location and use of land lines and cable TV
- Power supply including the location of and horizontal and vertical clearances to overhead power lines and space for transformers and substations
- Provision of gas usually in underground pipes
- Waste management and sewerage, normally the purview of local authorities
- Other transport authorities such as those responsible for railways, airports, sea ports and lakes and canals

In short, government policy requires that roads not be planned in isolation from the requirements of other authorities. Interactions between them are often formulated in legislation as the ultimate form of policy declaration.

Some interactions relate to utilities that are located within the road reserve as listed earlier, pointing out that service providers need to obtain the approval of the Transportation Authority as formulated in way-leaves prior to building their infrastructure.

It is necessary to have policies that define precisely where in the road reserve these utilities may be located. For example, foul water sewers and potable water reticulation are often installed on opposite sides of the road, with the sewers also normally being buried deeper than the fresh water supplies. This is to minimise the risk of the health hazard from pollution. No road authority is normally prepared to countenance a utility located under the road way itself. Furthermore, any underground crossing of the road is required to be sleeved so that the owner of the utility can access it without having to dig a trench across the road.

A further feature of the regulations and policies whereby a Transportation Authority defines the acceptable location of utilities within its road reserve is that these utilities are permitted only at the pleasure of the Transportation Authority. In short, if, for any reason, a portion of an existing road has to be relocated and this relocation impacts on the location of a utility, the service provider would have to undertake the relocation of the utility at its own cost to a new position that also has the approval of the Transportation Authority. Obviously, the Transportation Authority will carefully consider its own future plans and requirements, for example, an anticipated need for future doubling of a road's cross-section in granting approval of the proposed new location.

The location of utilities has been discussed in some depth because it is more likely to impact directly on the detail of geometric design whereas the other topics listed in the preceding text are more in the nature of planning issues relating, for example, to

- Origin–destination matrices
- Trip generation and attraction factors
- Transport economics
- The overall length of the network

A policy decision that is common to many transportation departments is that all road reserves should be fenced. This is a safety issue intended to restrict access to the road reserve to specific and carefully selected points, one of the aspects of access management in fact. In countries where the seasonal migration of wild life is common, it may, however, be considered desirable to leave the migration paths unfenced as an environmental issue. In defence of this decision, it is offered that migrating animals seldom move after dark and find grass more comfortable to sleep on than the road surface other than during periods of intense cold, when the road surface will often be warmer than the surrounding areas. Signage advising of the possible presence of migrating animals would be useful from a road safety aspect.

2.4 THE POLICY OF THE 'COMPLETE STREET'

As stated previously, planning was originally aimed at roads designed with only one goal in mind – the efficient, safe and high-speed transport of people and freight – and all transport policies were directed to this end. Towns that had relied on passing traffic for their existence thus sometimes had high-speed bypasses inflicted on them. These towns and villages generally suffered a serious economic decline and often took several years to regain some sort of right to existence.

Little or no attention was paid to the social aspects of the road network in the sense of the impact that it had on the lives of the people served by it. There is no doubt that streets have a significant impact on the lives of the people living on it. It has been found that people who live alongside a busy arterial generally have fewer friends on the street than do people living on a quiet residential street and tend to live in isolation from their neighbours.

Since then, other disciplines have entered the field of transportation and have successfully made their presence felt. The geometric designer is now not in the commanding role of yester year but is a member of a multidisciplinary team. This team includes

- Town and regional planners
- Environmentalists
- Sociologists
- Architects

Local communities demand that they be members of this team and have been found to have a major influence on the design and location of roads.

The presence of other disciplines on the design team has resulted in design becoming a significantly more holistic process. It is now accepted that every person in the road reserve, whether in a passenger car or bus or driving a multi-axle truck, or outside a vehicle has the fullest right to be there. People outside vehicles can be in the road for a variety of reasons and they all have to be accommodated. In residential areas, children should be expected to be playing in the streets and not be placed at risk by high-speed rat-running passenger cars. It is, in fact, these extravehicular activities that give the urban scene its richness and variety whereas moving vehicles tend to create a sterile if not actually hostile environment. Street design is now based on a policy of context sensitivity and this is also addressed in Chapter 4.

2.5 GEOMETRIC DESIGN AS A 'PRECISE SCIENCE'

Geometric design standards used to be predicated on Newtonian physics and vector diagrams. Much effort was put into acquiring information on coefficients of friction that determined, for example, stopping and other forms of sight distance and minimum speeds on curves and hence acceptable curve radii for various speeds. The engineers responsible for this research had the concept of the safety factor firmly embedded in their minds so that what had started out as a painstakingly accurate study was multiplied by a fairly arbitrary policy decision regarding a fudge factor to derive the geometric standards that have been successfully in force for several years.

Quite fortuitously, the standards that were derived were very much in line with what drivers were prepared to accept or wanted to do. Furthermore, the thought has been expressed more than just once and usually by politicians that 'standards are too high and, no, we don't want to drop our standards. We must seek appropriate standards'. Experience suggests that 'appropriate' is actually just a euphemism for lowered standards.

Sight must not be lost of the fact that the current developers of policy decisions acquired their knowledge and design experience in an era when geometric designers did not take kindly to 'interference' by those not of the fraternity. It was this attitude that resulted in the prescriptive nature still evident in some geometric design manuals today.

The policymakers of the day were principally motivated by thoughts of the capabilities of the design vehicle with a focus on the fast, safe, convenient and affordable movement of people and freight. While thinking about what vehicles can do, that is, the science of limiting performance, they tended to overlook the fact that no vehicle would travel in any way other than what the driver would wish. The thinking now has come to not what the vehicle can do but what the driver wishes to do. The human factors approach to design has now become entrenched in the philosophy of design and the policies that guide it.

2.6 POLICIES RELATED TO DESIGN

The geometric design manual is the principle source of information regarding organisational policy as it impacts on geometric design. Previously, manuals were prescriptive in the extreme and were inclined to use words such as 'shall', 'will' and 'may not'. Little scope was allowed for the designer to apply his or her mind to the creation of innovative solutions to design problems.

The belief was that a road designed to departmental standards was safe, with the corollary that roads that didn't match them were somehow suspect. This belief was widely supported by courts of law. Seeing that the departmental standards were derived from considerations of policy, it followed that departures from the precise scope of the standards implied that policy decisions were also suspect.

However, the winds of change have blown through the fields of geometric design and, as described in the rest of this book, have manifested themselves in fundamental changes in the approach to policy formulation. This arose, in part, from the budgetary restraints that have complicated transport departments approach to their mandate. The need for new roads and the maintenance of existing roads has not changed but the funding available to give effect to them has withered to some extent. The annual increase in budgets has largely been below the inflation rate, implying a net decrease in available funding.

The standard political demand has been for departments to 'do more with less'. This they have achieved largely through changes in their approach to policy formulation that have kept pace with changes in the philosophy of design. For example, it was noted that there were stretches of road that may, in fact, not have been upgraded to match the most recent design standards but, in spite of this, had good safety records. This led directly to the development of the concept of the design domain as discussed in Chapter 4. This became necessary because, in rehabilitation exercises, shrinking funds meant that authorities had a choice between rehabilitating a road from end to end or allocating some of the available funding to the upgrading of the geometry of a horizontal or vertical curve, leaving a portion of the pavement unrehabilitated.

The simplistic approach of the law courts regarding the correlation between standards and safety has been replaced by the mechanism referred to as the design variance whereby the road authorities can prove that they have applied their minds to the standards in use on their road networks. The ruling of a South African judge several years ago that 'bridges must be designed to withstand the worst flood of all times' would receive short shrift today.

2.6.1 Cost-efficient design

Previous polices were directed towards geometric design that minimised the cost of construction. It took some time for departments to realise that, in minimising the cost of construction, the resulting light pavements tended to deteriorate rapidly, imposing a significant maintenance bill on them. The policy was then adapted to minimising the cost of construction AND maintenance. With transport economics gaining ground, the realisation finally dawned that construction and maintenance between them added up to about 25 per cent of the total cost of a road, with road user costs being the remaining 75 per cent. The current approach is that the 'whole life' cost of the road most be minimised to achieve a cost-effective design. The so far unproven belief that 'for every metre of false rise and fall eliminated, the length of the road can be increased by 25 metres' arises from this new approach.

As illustrated in Chapter 16, while economic analysis is an indicator of cost efficiency, it is necessary also to consider the irreducible factors, that is, elements to which it is not possible to attach a cash value. These include the issues of social factors and minimisation of the environmental impact of the road.

Although it doesn't have the formality of a policy with a proper rationale, the design life of a road is often taken as being 20 years. The annual maintenance and road user costs of this period are discounted to the baseline year, which normally is the year in which the road either has or will be opened to traffic. This provides a net present worth of the whole life cost of the road. It is recommended a policy be formulated whereby a range of design lives is analysed in terms of annual maintenance and road user costs with a view to establishing the optimum design life in terms of the minimum present-day discounted costs.

2.6.2 Access control

As discussed in the following chapter, access control deliberately restricts access to any road to ensure that its functionality is not compromised by the number of vehicles on it. This can be achieved, for example, by

- The systematic spacing of successive intersections
- The layout of intersections
- The design and operation of driveways, intersections and/or interchanges
- The design of median openings

In rural areas, access to a road is usually limited to one access per farm, with a farm defined as comprising all those properties that have been consolidated into and operated as a single unit. In urban areas, successive intersections are sometimes located to promote 'green wave' operation. This means that, just as a platoon of vehicles reaches an intersection, its traffic signal changes to showing a green phase. In very special circumstances, it is even possible to have green wave operation simultaneously in both directions. The norm is that signal phasing is set to favour the heavier flow, for example, the peak hour flow toward the CBD in the morning peak and away from the CBD in the evening peak. The opposing flow in either case simply has to take its chances with a less than optimal signal phasing.

Developers of areas immediately adjacent to a freeway often have the belief that direct access to the freeway is absolutely essential for the viability of their parcel of land. This may be true but it is not always possible to grant the requested access without jeopardising the operation of the freeway. Rejection of the application for access is invariably followed by recourse to a court of law with a view to overturning the departmental ruling. Arguments are offered in favour of the massive job creation that will follow the approval of the application, improvement in the tax base of the department concerned, improvement in the general economy of the area and so on.

The problem may be that there simply isn't enough space between existing interchanges to accommodate the interchange needed to provide the requested access. There has to be adequate space between on upstream on-ramp and the following off-ramp to allow for the weaving that has to take place between merging and diverging vehicles. Without this, freeways simply become totally clogged and brought to a standstill. All the rosy visions of the developer are then brought to nought by what used to be an operating freeway being brought to the status of an expensive car park.

In general, courts seem to favour the case of the plaintiff, that is, the developer. The developer is seemingly not required to substantiate any of the claims, fanciful or otherwise, made in respect of the advantages of providing the development. The Transportation Department, on the other hand, is placed in the position of having to prove the merits of its rejection, to prove in fact that 'it has applied its mind' in deciding not to approve the application.

In cases where a development may be granted access to a freeway, the department invariably would not have the funding available to finance the required interchange. If the developer is not prepared to carry this cost, the interchange would simply not be provided. It should be noted that, while weaving within interchanges can be accommodated by collector–distributor roads, weaving between successive interchanges is far more critical to the operation of the freeway as a whole. This is because it involves the operation of the main line carriageways as opposed to interchange ramps. If there is any limitation on the space needed for provision of the interchange it may thus be necessary to provide auxiliary lanes upstream and downstream of the site to facilitate these weaving operations.

2.7 CONCLUSION

The most appropriate policy for transportation departments to adopt is to accept that the rigidity of geometric 'standards' is no longer and, in fact, never was a basis for good design. The fact of the matter is that the 'one size fits all' approach to design is fundamentally flawed. A standard appropriate to the southern border of Texas is quite possibly not going to be appropriate to Alaska.

Designers have to be allowed the flexibility that the politicians were inadvertently asking for with their need for 'appropriate standards'.

As advocated throughout this book, a philosophy and policy of rigid 'design standards' should be replaced by design made possible by 'geometric guidelines'. Obviously, this is not advocating a 'laissez aller' approach to design that is but a short step away from anarchy. Departing from a design value recommended by a guideline would have to be motivated by the designer and formally approved by the Transportation Authority before it could be incorporated in the design of a road.

Chapter 3

Road classification

CONTENTS

3.1 INTRODUCTION

All roads have in common that they are owned by someone, usually a statutory body, are paid for by direct or indirect taxation or tolling and form part of a network. In any form of engineering endeavour, structure is important, and this is addressed through various forms of classification.

The needs for any system of road classification vary, and these are discussed in the following sections. Classification can be based on

- Road status
- Route numbering as a guide to funding
- Route numbering as an aid to navigation and tourism
- Function

with the last mentioned being that usually applied to geometric design.

3.2 ROAD STATUS

Each road authority has its own unique numbering system, usually based on the perceived relative importance of each road in the authority's network. The road authority may refer to roads as being primary, secondary or tertiary, or main, divisional, district or county roads or any other grouping of nomenclatures.

This serves as a guide to various departmental functions such as allocation of maintenance effort or to the splitting of finances between departments and others such as local rate payers. For example, on roads that are very low in the hierarchy of the rural network, the policy may be that the department send a grader out at regular intervals to blade the road, with any other work being undertaken or funded by local farmers. The latter work could include keeping the verges tidy, clearing drainage channels, and so forth.

3.3 ROUTE NUMBERING AS A GUIDE TO FUNDING

In many countries, numbering defines the allocation of funding, whether at the national, state or province or local level. In Britain, numbered classification was originally used for the purpose of keeping track of these funding allocations. The country is divided into nine zones and, within each zone, roads are classified as either A or B routes. The major roads, or A routes, radiating from London are all given single-digit numbers, with the digits increasing in a clockwise direction. The A1 is from London to Edinburgh.

Reference is made in various countries to numbered trunk roads. These are roads connecting major centres and are usually the recommended route for long-distance and freight traffic. In the United States, the interstate highways would be the equivalent of trunk roads. The 14 member states of the Southern African Development Community have jointly created a network of numbered trunk roads serving freight movement and with a common set of minimum geometric standards.

3.4 TOURISM ROUTES

Another form of classification is the numbering of routes from end to end whether or not the various segments of the route have any other number derived from an administrative classification system. This numbering of routes is intended to serve the needs of tourists. Dependent on the extent to which they cross borders between the various authorities, numbered routes such as tourism networks may fall within the remit of any one of them or be a shared responsibility. To this end, road maps show the major routes not only by the convention of line style and colour but also by number. This applies particularly to rural areas and would focus on major rural routes.

Within urban areas, numbering of routes is a convenient way of defining roads that would at least have the status of being collector roads, as discussed further in the following sections.

3.5 ADMINISTRATIVE CLASSIFICATION

Administrative classification addresses the typical hierarchy of government agencies, from central government down to the lowest level of local government. In general, this hierarchy comprises three levels: the national level, the provincial or state level and the local level, which, in turn, may comprise various sub-levels such as county, city, town, village or rural area.

The administrative classification of roads usually has a bearing on the method of financing their construction and maintenance apart from the obvious distinction being drawn between the various tiers of government.

Funding may be generated by fuel levies and these would go directly to the national or central level of government. The lower levels of government would usually receive a grant of funding from this source. In addition, on the basis of its importance to some or other national imperative, the central government may elect to subsidise or bear the total cost of whatever project is being mooted.

3.6 FUNCTIONAL CLASSIFICATION

Road classification implicitly acknowledges that the roads network has a built-in hierarchy of roads based on their functions and capacities. The basic hierarchy comprises arterials, collectors and local roads. Freeways are essentially arterial routes but their operation is different from that of all other roads. For this reason they are usually regarded as a separate hierarchical group.

For geometric design, the most useful form of classification is functional classification, as it defines the spectrum of road usage from pure mobility to pure accessibility. This, in turn, supports the selection of the design speed and the design vehicle. These two parameters, in combination with current and anticipated traffic volumes, define geometric standards of the horizontal and vertical alignment, and intersections or interchanges and definition of the cross-section.

3.6.1 Trip component

A trip may have only a single component, such as passenger car from home to the home of a friend. Most trips, however, comprise several components. A commuter could leave home on a local street and then move to a collector that connects via a transition ramp to a freeway or other arterial. After the relatively long-distance high-speed main movement, the trip has the same elements as those that preceded the main movement but in reverse order. An exit ramp provides a transition to a distributor that may be a moderate speed arterial taking the commuter closer to the intended destination. The arterial would then be left and a short journey on a collector would end at a parking facility that may be on- or off-street parking. This hypothetical trip comprises 10 stages.

It is not necessary for a journey to go through all the elements of this hypothetical trip and some can be left out. As a general rule, though, there should not be gaps that leapfrog more than one of the elements. This is because all the elements of the trip will still be there but will take place on some substitute for the missing element. An off-ramp may connect directly to a parking garage but would join a distribution road within the garage connecting to the various parking areas. Collector roads or lanes would provide a link to the parking bays. The parking aisle within the parking bay would take the vehicle to the individual parking spaces. In short, all the elements of the trip are identifiable. And the roads that are the network equivalents have been highlighted.

It would not be wise for an off-ramp to connect directly to a local street, for example. Two possible inferences can be drawn from this omission.

1. The local street cannot accommodate the flow discharging from the freeway, resulting in a backup onto the mainline lanes of the freeway. Vehicles travelling at a speed of 120 km/h or more may unexpectedly be confronted by a queue of stationary vehicles with a high probability of a multivehicle crash.
2. The flow on the freeway is such that the local street can accommodate it. The turbulence of vehicles diverging from the freeway may have an adverse effect on the mainline lanes, particularly through the lane changing that precedes the departing vehicle arrival in the outside lane prior to the diverging manoeuvre. The level of service on the mainline could thus drop to an unacceptable level, resulting in upstream congestion.

The only conclusion that can be drawn is that the trip components will all be present, *whether or not the required network elements are present*. Provision must therefore be made for them.

3.6.2 Mobility versus accessibility

The example of the elements of a trip as shown previously implies that they happen at different volumes and speeds. At the one end of the scale, the local street is geared to the provision of access to individual properties or other trip ends such as parking garages. The major part of the trip is taken at high speeds over significant distances so that design for mobility becomes the keynote. Unfortunately, these are not the opposite sides of a coin although they are linked in the sense that accessibility is becoming appreciated as a co-equal goal with mobility. It is axiomatic that travel is a derived demand. In short, except when on holiday, people don't normally travel for enjoyment but rather to reach some opportunity available at a destination. This has an important implication in that the traditional assessment of success in transportation being an emphasis on mobility is, in fact, incorrect.

Transportation planners and engineers generally attach significant value to the alleviation of roadway congestion, that is, increasing mobility. If mobility is achieved through adding infrastructural capacity to the network, the outcome could be that desired destinations could be located at greater distances apart. For example, in times gone by, rural villages tended to be spaced at 50 kilometres. Personal transport was by horseback and the farmer located halfway between villages had a travel time of about 1 hour. With personal transport now being by passenger car, a travel time of 1 hour equates to a distance of about 100 kilometres, suggesting that towns located at a spacing of about 200 kilometres would flourish and those in between would tend to decline.

Increased mobility requires more time and money spent on travelling, hence less time and money to spend at the destination. In terms of the concept of the derived demand, people don't necessarily wish to travel as such; they only wish to arrive at their destination. The increased cost in terms of time and money of reaching a destination is thus counter-productive (Levine and Garb, 2002).

Relief from congestion should thus be rejected as an independent goal of transportation policy. Enhanced mobility can be valued only insofar as it enhances accessibility whereas mobility gains resulting in a loss of accessibility represent a failure of transport policy. If destinations are fixed in space, an increase in mobility results in cost per kilometre being reduced – an increase in accessibility. If an increase in mobility makes it possible to live further away from the place of employment, mobility gains result in a loss of accessibility. The emphasis of the focus on mobility when accessibility is perceived as some sort of poor relation should thus be replaced by accessibility as being the prime measure of transportation efficiency.

The proportion of service shown in Figure 3.1, which has been widely used for several years as an illustration of the continuum from pure mobility to pure accessibility, should thus be scrapped as being misleading.

It would appear that congestion is not necessarily bad per se. It becomes bad only when it increases the time and money spent on accessing a destination. In fact, strategies aimed at reducing congestion could be inimical to accessibility by causing travel distances to grow and land uses to spread, as discussed earlier. In contrast, congestion pricing could be designed to promote more appropriate pricing of scarce roadway space and facilitate the alternative to a growth in travel distances.

This does not suggest that congestion pricing should actually be implemented to keep vehicles away from congested areas or propose the levying of a charge for entering an already

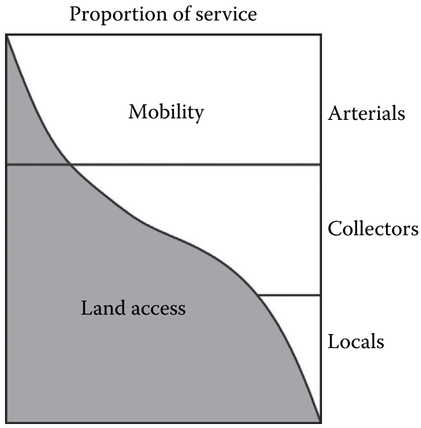


Figure 3.1 Mobility versus accessibility. (From American Association of State Highway and Transportation Officials. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

congested area, as has already been applied by various metropolitan authorities such as that of London. It is rather intended to identify the effect of a proposed project on accessibility overall rather than simply testing its ability to improve capacity and hence mobility (Levine and Garb, 2002).

3.6.3 Definition of urban and rural areas

Everyone seems to know the difference between urban and rural areas on the basis of, 'If you're surrounding by buildings you're in town. If you are surrounded by farm lands, you're in the country'. If it only were so easy. Most countries operate on the basis of the definition of urban areas, with the rural areas being anything that is not classified as urban.

In the United States, urban areas are those with a population of more than 5000 people. There are two categories of urban areas differentiated on the basis of total population. 'Urbanised areas' are those with a population larger than 50,000 and 'small urban areas' have populations of between 5000 and 50,000. According to the Census Bureau, urban areas are those with a population density of 386 people per square kilometre (1000 people per square mile) in the core area and 193 people per square kilometre (500 people per square mile) in the surrounding area. In the U.S. Department of Agriculture's natural resources inventory, urban areas are officially known as developed or built-up areas of more than 4 ha (10 acres).

In Canada, urban areas are designated as 'population centres' if they have a population of at least 1000 people with a density of 400 or more per square kilometre.

In its Road Traffic Act (Act 29 of 1989), South Africa defines an urban area as that falling under the jurisdiction of a municipal or metropolitan authority and that has been subdivided into individual properties including the roads abutting on them as listed on the municipal cadastral map or cadastre, with the latter providing much more detail than the map. The municipal commonage thus falls outside the urban area.

In the United Kingdom, the definition is based on the extent of 'irreversible urban development' indicated on ordnance survey maps. The definition is an extent of at least 20 ha and at least 1500 census residents. Separate areas are linked if less than 200 metres (220 yards) apart. If the urban area includes an Anglican cathedral, it acquires the status of a city.

3.6.4 Functional network components of rural areas

Rural networks comprise

- Arterial roads
  - Major arterial roads
  - Minor arterial roads
- Collector roads
  - Major collector roads
  - Minor collector roads
- Local roads

The principal arterial system accommodates long-distance travel and movement between all urban areas with populations of 50,000 people and most of those with populations of more than 25,000 people. These roads are usually freeways. Minor arterials provide linkages between cities and large towns, including other traffic generators such as major resort areas. The spacing of minor arterials is such that all developed areas are within reasonable distances of major arterials. Trip lengths and travel densities are greater than those of local roads but less than those of major arterials.

Rural collectors serve trips with lengths shorter than those served by arterials at lower speeds than those typically encountered on arterials. Major collectors provide linkages to towns not directly served by arterials, linking them also to larger towns or cities or to arterials, both major and minor. Minor collectors are spaced at intervals consistent with population densities to provide linkage to higher order roads in the network. They also serve smaller communities and provide a service to rural areas by linking them to locally important centres.

The primary function of local roads is to provide access to individual properties adjacent to the network. Travel on local roads is typically over short distances at relatively low speeds. Local roads are all those not classified as arterials or collectors.

3.6.5 Functional network components of urban areas

The components of an urban network are

- Arterial roads
  - Major arterial roads
  - Minor arterial roads
- Urban collector roads
- Urban local streets

The major arterial roads are the effective backbone of any urban system. They carry large volumes of traffic at high speeds from one end of the urban network to the other. They thus perform a valuable service to through travel and rurally oriented traffic. Their importance also derives in large measure from service to major circulation movements within the urban area. Major arterial roads are usually freeways with complete control of access by the use of interchanges. As a general rule, the minimum spacing between successive major interchanges can be as shown in Table 3.1.

The spacing between parallel freeways is addressed in Chapter 11. The concept of the white line breakpoint distance, which is suggested as the form of measurement of interchange spacing, is introduced in Figure 11.6.

Table 3.1 Spacing of interchanges

Configuration	*Urban areas*	*Rural areas*
Access to access	1300 m	2170 m
Access to systems	2100 m	3270 m
Systems to access	1420 m	2170 m

Source: Wolhuter KM and Garner D. *Interchange spacing in Gauteng*, 3rd International Symposium on Highway Geometric Design. Transportation Research Board, Chicago, 2005.

Note: The configurations 'Access' and 'Systems' are defined in Chapter 11. The distances in the table are 'white line distances', which are discussed also in Chapter 11.

Links to freeways are often provided by minor arterials. They are invariably dual carriageways and connect to the local road system via traffic signal-controlled at-grade intersections. To achieve 'green wave' operating conditions on minor arterials their intersections should be spaced at roughly 500 metres, centre-line to centre-line but can actually vary from as little as 200 metres to about 1000 metres in the central business district up to as much as 5000 metres in the outlying urban areas. As a general rule, all arterials not classified as major arterials are minor arterials.

Local streets include through streets, crescents, culs de sac and dead ends. Speeds are low and the principal function of local streets is to provide access to adjacent properties and a link from these properties to the higher order roads in the network. They constitute the lowest level of mobility and do not usually contain bus routes. They can fall prey to rat–running, and through movements are actively discouraged and, where possible, totally eliminated.

Chapter 4

Fundamental design considerations

CONTENTS

4.1 INTRODUCTION

Historically, geometric design focussed on two main functions: mobility and accessibility. Of these, mobility seemed to be the more important of the two, with accessibility being relegated to the Cinderella role of a limited focus on local urban streets, whether residential, commercial or industrial. As illustrated in Figure 4.1, these features were seen as a continuum ranging from pure mobility and a total control of access as typified by the freeway to the pure accessibility of the local cul de sac.

There was little or no attention paid to the social aspects of the road network in the sense of the impact that it had on the lives of the people served by it.

The physical environment was also not treated with much sympathy. It was seen as a virtually endless source of road building material. In many parts of the world, particularly in developed countries, sources of road-building material are now already exhausted. What was then perceived a swamp that had to be drained is now a highly prized wetland. The last 40 years have seen a growing concern for the well-being of the environment. The public has demanded and gotten a greater say in what is designed and ultimately built in all fields of civil engineering. Road networks are no exception. Geometric designers have lost their God-like powers by which complaints by the 'uninformed' were treated with the disdain of those who know what is really required to provide an efficient network.

Apart from the changes to design forced by external circumstances, the designers themselves realised that there was a need for a radical change in the philosophy of design. In spite of the beliefs held by courts of law worldwide, it was realised that adherence to a set of design standards did not necessarily result in a design that was safe nor did a departure from those standards result in a design that was less than safe. It was also realised that there can never be such a thing as a safe road (Hauer, 2000). A safe road would imply that there would never be any crashes on it except those caused by circumstances other than the road

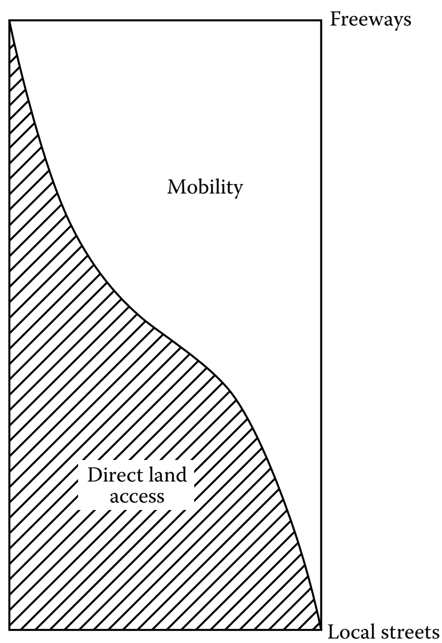


Figure 4.1 Relationship between functionally classified systems.

itself, that is, mechanical malfunction of a vehicle or driver error. Safety is thus a relative concept. A road could be made safer or a design error may cause it to become less safe; it can never be made completely safe.

4.2 OVERVIEW OF DESIGN CONCEPTS

Since the late 1990s, there have been two paradigm shifts in the design process and two changes in design methodology. Geometric design as a 'pure science' based largely on Newtonian physics, the limits of the performance of the vehicles on the road and reliance on a set of standards has had to make way for a significantly more complex approach. What was a vehicle-based approach to design has become a people-based approach. This approach has been broadened to address the needs of people other than the drivers of vehicles. The physical and the social environments in which the road is located have also gained significantly in importance. In consequence, the design approach has become multidisciplinary in nature, with the geometric designer being a member of a team, as opposed to having the previous role of final arbiter on the design of the road.

As an example, it is now realised that design should be based on what the driver wishes to do rather than on what the vehicle can do. For example, calculation of stopping sight distance based on several assumptions regarding road smoothness and dryness, tyre condition, and vehicle braking performance results in the determination of what constitutes a panic stop. This is outside the expectations of the average driver. The adoption of substantial safety factors has, however, resulted in values of stopping sight distance that, more by good luck than good management, more or less match with what is acceptable to drivers. In addition to driver expectations, there are other characteristics of drivers that should be borne in mind in the design process and they are grouped together in the concept of a human factors approach to design as discussed in Section 4.3. The human factors (HF) philosophy was developed in the 1930s and some aspects of it, for example, reaction time, were adopted in highway design at that time. It is only comparatively recently, however, that HF has broadened and achieved the prominence that it now has in design.

Design should also consider road users other than the drivers of vehicles. In the case of freeways, which purely address mobility, there should not be any road users other than the drivers and occupants of moving vehicles. At the other end of the scale, the local street has to accommodate people outside vehicles. The reasons why these people are in the road reserve include

- Being pedestrians moving from one place to another
- Socialising
- Children playing in the street
- Relaxing in sidewalk cafes
- Window shopping and browsing in flea markets
- Cycling or using other nonmotorised forms of transport

The context in which the road is designed thus needs to be considered. Context-sensitive design (CSD), also referred to as context-sensitive solutions (CSS), has become a major concept in the field of geometric design and is addressed in Section 4.4. The latter name is gaining general acceptance because context sensitivity can be achieved without necessarily having to resort to design.

A freeway, with its high volumes of fast-moving vehicles, is a sterile and essentially hostile environment and the design concerns are essentially those of safety and aesthetics. The local street, on the other hand, is a fundamental feature of what is referred to as a 'sense of place'. Whether people are in vehicles or not, they all, and for whatever reason, have a perfect right to be in the road reserve. It should be designed in such a way that they find it to be a pleasant place to be in. This sense of place could, for example, be generated by a combination of

- Low vehicle speeds and volumes
- Wide sidewalks
- Shade trees, shrubs and other plantings
- Street furniture in the form of benches, tables and chairs
- Possibly artwork

A further group of people to be considered are those who live in close proximity to the road. They are subjected to air and noise pollution and placed at risk of injury or worse by vehicles passing through their area. Efforts should be made to reduce the side effects that inevitably result from the provision of a road. These could include a once cohesive community being effectively bisected by a new major arterial passing through the area. Addressing the needs of the local community requires a combination of environmental awareness and context sensitivity.

An important concept is that of consistency of design. This concept is not new and much research has been done over the years to develop a rational methodology whereby consistency can be tested and achieved. It is aimed primarily at removing the element of surprise from the driving task. This is because a driver who is surprised may, under pressure, take a wrong decision possibly resulting in loss of control of the vehicle and, in the worst case, a crash. Design speed was the first measure to be applied to the achievement of the consistency of design but has been shown to be not without its weaknesses. In comparatively recent times, a methodology has been developed whereby consistency can be quantified (Lamm et al., 1999). This methodology has gained international acceptance.

Limited budgets have forced road authorities to rethink their approach to standards. For example, when a road is being rehabilitated, it is possible that the values of some of the standards applied in the original design may have been superseded by higher values and yet crashes are still relatively infrequent on 'substandard' sections of the alignment. Should funding be diverted from essential upgrading and rehabilitation of the pavement to unnecessary improvements in the road geometry? The concept of the design domain discussed in Section 4.6 considers how departures from accepted design standards should be addressed. In an increasingly litigious era, it is essential that designers and the road authority concerned should be able to prove in a court of law that they have 'applied their minds' in the provision of the road, including its design, construction and maintenance.

4.3 HUMAN FACTORS–BASED DESIGN

4.3.1 Introduction

Human factors engineering seeks to relate devices and systems to the capabilities and limitations of the people who use them and is arguably the first discipline to relate technology to human psychology (Campbell et al., 2008). It is thus of cardinal importance in eliminating mismatches between design practices and human frailty and desires.

Although many design practices are based on empirical behavioural data and research, this is not always the case. Some design practices may suffer the shortcomings of being based on

- Engineering judgment rather than rigorous research
- Outdated data that are no longer appropriate to current driver behaviour
- Overly simple models of what drivers see and do
- A large number of assumptions concerning the variables of these models
- Incorrect assumptions regarding road users' capabilities, limitations or desires

These design practices also do not necessarily reflect

- The special needs of some road users such as
  - Older drivers
  - Pedestrians who are visually or hearing impaired
  - Pedestrians who have difficulty in walking
  - People in wheelchairs
- Technological improvements in
  - Vehicle features
  - Communications technology
  - Roadway features
  - Traffic control devices

4.3.2 Driver characteristics

As stated previously, driver characteristics traditionally considered in the design process tended to be limited to driver eye height and reaction time. Driver eye height is measured from the road level at the centreline and is a combination of the height of the driver's seat and the height of his or her upper body. The driver eye height is of cardinal importance in the determination of the various forms of sight distance discussed in Chapter 7. Over the years the driver eye height has become lower in the case of drivers of passenger cars because of changes in vehicle design. The eye height of truck drivers has remained unchanged.

There are, however, many other characteristics that also impinge on the driving task. These include

- Age
- Driver behaviour
- Decision making
- Expectancy
- Gap acceptance
- Information processing
- Preferred limits of forces acting on the human body, such as
  - Deceleration forces
  - Force of gravity
  - Centrifugal forces
- Psychological characteristics
- Response to workload
- Visual perception

Many of these characteristics have yet to be quantified through research.

4.3.2.1 Age

Developed countries such as the United States and the United Kingdom and other European countries have aging populations. The United Kingdom has more people older than 60 years than younger than 16 and there are five times more people older than 85 years than there were 60 years ago. The median age of people in the United Kingdom is 40.5, that of the United States population 37, and that of other European countries ranging from 38.2 to 48.9. The median age implies that in the United Kingdom, for example, 50 per cent of the population is older than 40 years. In African countries, on the other hand, the median age of the population ranges from 15 to 20, with South Africa being an outlier at 24.7 (Wolhuter and Garner, 2004).

Elderly drivers, people who are 70 years of age or older, require more time than others to decide on a preferred course of action. And many of the problems that beset them are found in intersections (Staplin et al., 2001; Hakamies-Blomqvist et al., 2004; Smiley et al., 2008). Heavy mental loads can easily confuse an elderly driver. The presence of other and conflicting vehicles in an intersection generates a substantial mental load, so the layout of the intersection itself should be kept as simple as possible in order not to add unnecessarily to the driver's distress. Furthermore, clear and continuous information on precisely where they are when driving through an intersection is particularly useful to elderly drivers. Because of diminishing night vision, unlit intersections can be a major problem for the elderly.

The road should be so designed that it allows sufficient time to perform tasks one at a time with some recovery time between successive tasks. This includes reading signs. Signs should therefore be located where the driver needs and expects the information to be and of a size that supports easy legibility. Navigational information at intersections should be logical, consequent and hierarchical in nature.

Regarding the detail of intersections layout, various points are worthy of consideration. For example, a 90° angle of skew would accommodate loss of neck mobility and poor peripheral vision in addition to reducing the size of the intersection. Protected left turns at signalised intersections would eliminate the gap acceptance problems of the elderly, as would the conversion of a priority-controlled intersection to a roundabout. Offsetting left turn lanes would be beneficial (and not only to the elderly), as this would afford a longer time in which to observe opposing vehicles and thus form a better estimate of their approach speed.

Elderly pedestrians would benefit from simplicity of intersection layout and having intersections well lit so that they can easily be observed in their entirety. Younger pedestrians have a 15 percentile walking speed of 1.21 m/s compared to the 0.94 m/s of the elderly pedestrian. A walking speed of 0.94 m/s suggests that crossing an undivided four-lane road could require a gap of 16 seconds, which is not likely to occur with any frequency on a street with traffic flows sufficient to warrant this width of cross-section. Even if not provided along the full length of a four-lane road, the inclusion of median islands at the intersections would be helpful to elderly pedestrians.

4.3.2.2 Expectancy

The driving task is based on features of the road that the driver expects to see. This expectancy derives from the sum of the driver's past experience and what was seen over the last few minutes. There are three basic forms of expectancy (Burrell et al., 2002):

- Continuation expectancy
- Event expectancy
- Temporal expectancy

Continuation expectancy is that the events of the immediate past will remain unchanged. For example,

- The radii of upcoming curves will be similar to those observed over the last few minutes
- The speed of the vehicle ahead will remain relatively unaltered

Event expectancy suggests that events that haven't happened will not happen. This results in a disregard for railway crossings where rail traffic is low or for minor intersections. Drivers expect that no hazard will present itself because none has been seen before.

Temporal expectancy suggests that, where events are cyclic, such as at traffic signals, the longer a given state prevails, the greater is the likelihood that a change will occur. Although this is a reasonable expectation, it may result in inconsistent responses. Some drivers will accelerate towards a green signal in anticipation of its changing to red whereas others may decelerate.

There are numerous other expectations that manifest themselves. For example,

- A curve in one direction will be followed by a curve in the other
- A crest curve will be followed by a sag curve
- On- and off-ramps are located on the right hand side of the carriageway (in the case of driving on the right and on the left in the case of driving on the left)
- The rule of left-turning vehicles yielding to straight through traffic creates the expectation that left-turn arrows would be lagging rather than leading (Sampson, 2011)

It is necessary for the designer to place him- or herself in the position of the driver in viewing the design to establish which form of expectancy needs to be addressed. The 'accident black spot' where numerous drivers make the same mistake is not unknown. It can reasonably be inferred that the road misinformed the driver in some or other way. Although driver error has been identified as the cause for a high percentage of crashes, typically 70 per cent or more, it is suggested that some of these could have arisen from the road design seducing the driver into an incorrect response.

4.3.2.3 Gap acceptance

Gap acceptance comes into play when a driver wishes to cross the path of an oncoming vehicle, for example, at an intersection or in the case of a left turn. Previously, the movements of a vehicle on the minor leg of an intersection would be modelled on the basis of several assumptions. These would include

- The length of time it takes to get the vehicle moving from rest
- The rate of acceleration of the vehicle on the minor road
- The speed at which the turns to the left or to the right are executed bearing in mind that the radius of a turn to the left has a larger radius than that to the right and could thus be executed quicker
- The length of the vehicle on the minor road in the event of its crossing rather than turning onto the major road
- The deceleration applied by an opposing vehicle in the major road

With a range of possible values for each of these variables, intersection sight distance could assume a very wide range of values. However, all these problems fall away if the human factors characteristic of gap acceptance is adopted (Fitzpatrick et al., 2005). Gap acceptance tacitly addresses all the variables involved in the manoeuvre as a single value of time in terms of the driver's judgement of the matter.

This and the other human factors listed in the preceding text are discussed further where appropriate in the text.

4.3.3 The self-explaining road

Geometric design standards are often based on distance-speed criteria, such as stopping or passing sight distances, to specify design elements. These are thus based on time and how road users use it (Campbell et al., 2008). Humans are notoriously slow single-stream data processing beings and can take as long as 3 seconds to select the response to initiate in the case of a simple but unexpected situation. In complex situations such as in busy intersections, taking decisions may require a longer period of time.

The preferred human factors approach is that no information should be unexpected. This preference leads to the approach of the 'self-explaining' road. This is described (Theeuwes and Godthelp, 1992) as an environment where drivers know how to behave based on the road design rather than on external agents such as road signs and signals. A road that is not self-explaining has the potential for surprising the driver into an unpremeditated and possibly incorrect response with the consequences that this implies. It could also cause traffic operations to be inefficient and delayed in addition to being unsafe.

Three axioms have been defined (World Road Association, 2008) in pursuit of the concept of the self-explaining road. These are

- The six-second axiom
- The field of view of axiom
- The logic axiom

The first of these suggests that a driver requires 4 to 6 seconds to change to a new driving environment. There are three logical steps in the change process. The first is the identification of the critical point at which the new environment applies. The second is the identification of the nature of the change required, be it to a reduced speed or to a change of direction. The final step is the implementation of the identified response. The second and the third of these have traditionally been applied in the calculation of stopping sight distance. In terms of the human factors approach to design the first step must be added to these. At an approach speed of 100 km/h and a deceleration rate of 3 m/s^2 the length of the transitional area amounts to 300 metres.

The second axiom is that the road must offer a safe field of view with the emphasis on the word 'safe', which is more than merely sufficient. The view can be monotonous to the point where the driver may actually fall asleep at the wheel or be cluttered by objects such as billboards competing for the driver's attention. Optical misguidance and illusions can complicate the driving task unnecessarily. A user-friendly and self-explaining road seeks to avoid these pitfalls.

In terms of driver expectancy, surprises are an unwelcome intrusion in the routine of driving. It follows that situations, where available sight distance is approaching the minimum level for a given design speed, are likely candidates for a surprise. This is simply because drivers do not keep their eyes glued on the immediate view ahead. They look at the instrument panel, rear view and wing mirrors, passing signs, provocative billboards, the view, and so on. By the time their attention is, once again, focused on the road ahead, it is possible that the available sight distance has already been used up and they don't have sufficient time to respond to whatever should already have received their attention.

The third axiom is that roads have to follow the driver's perception logic. This logic is based on the sum of the driver's past experience and the history of the road as seen in the previous few minutes, in fact continuation expectancy as previously described. Much of the driving task occurs as an automated process and unexpected events, in a manner of speaking, switch off the automatic pilot. Some time passes until the event is dealt with and during this time the driver is not wholly in control of a vehicle that may be moving at more than 30 m/s. Consistency of design is a very important feature of the self-explaining road.

Where changes are necessary, these changes should be indicated as early and as clearly as possible. The reason for the change should also be obvious to the driver. It could be occasioned by entering an urban area or a change in topography from hilly to mountainous. Furthermore, changes should not be sudden; for example, a significant drop in design speed should be achieved in a series of steps rather than in a giant leap.

4.4 CONTEXT-SENSITIVE DESIGN

4.4.1 Overview of context-sensitive design

No road is designed or constructed in a vacuum. It could form part of an urban high-speed environment, or perhaps be a scenic drive in an area that attracts tourists. In short, the provision of the road takes place within some or other context. This context will directly influence the parameters of the design and inform the selection of the dimensions of the geometric elements of the road. Context-sensitive design (CSD) thus addresses safety and efficiency within the framework of the road's natural and human environment.

> Context-sensitive design asks questions first about the need and purpose of the transportation project, and then equally addresses safety, mobility, and the preservation of scenic, aesthetic, historic, environmental, and other community values. Context-sensitive design involves a collaborative, interdisciplinary approach in which citizens are part of the design team. (Neumann et al., 1998)

This workshop signalled the development of the movement towards CSD in the United States.

There are two main drivers of CSD. The first is that, starting in the 1970s, many countries enacted legislation protecting the environment. Inevitably, designers with the concepts of mobility and accessibility deeply ingrained in their psyche ran into resistance from the public. Improvements sought by the designers that were were no longer unquestioningly accepted were

- Faster travel times
- Greater safety
- Reduced congestion
- Less delay

Although roads were accepted as being necessary, they were also perceived as having a definite cost in terms of invasion of the environment and disruption of community life. The second driver is that community values had gained a voice at about the time that the world was becoming increasingly litigious. Perhaps the two are not entirely unrelated. Communities had become more aware of what can be described as a 'sense of place'. Design proposals were weighed up regarding their impact on historical, cultural and community values. Where designs were in conflict with these values, proposals would be hotly resisted, if not rejected out of hand.

CSD is also known as context-sensitive solutions (CSS). This is because of the need for an outcome rather than merely a process. The road must be designed, built, maintained and operated, thus involving all the activities of a Transportation Department or Agency. It is also possible that the solution of a problem does not necessarily need a design component and could equally well be solved through the implementation of an operational measure. Many thus believe that CSS is to be preferred as being a more encompassing term.

4.4.2 The contexts of design

The most obvious contexts are the urban and rural environments. The urban environment is generally defined by ‘intensive land use for closely spaced buildings and structures and impermeable surfaces’. Traffic patterns are characterised by high volumes of vehicles moving at low speeds, whereas in the rural environment, traffic volumes are low and speeds high.

The high speeds of the rural environment require high standards of geometry. Gradients are low and radii of horizontal curvature high, as are the *K*-values of the vertical curvature. In consequence, cuts and fills could be high. The lower speeds of the urban environment make it possible to adopt lower design speeds and the accompanying lower geometric standards. In consequence, roads can be held closer to ground level. This eases problems of provision of access to surrounding land uses.

In the rural area, drainage comprises mainly the transporting of water from the catchment area across the road reserve and drainage of the road reserve towards the surrounding area. In urban areas, the road reserve serves also as a drainage conduit of the surrounding land use. Having the road slightly below the natural ground level supports this drainage.

Rural intersections often are no more than simple bell mouths between two two-lane roads. Pastoral farming typically requires large farms, and in such areas intersections could be several kilometres apart. Agrarian farming would result in a closer spacing of intersections but that would still be longer than that found in urban areas. Urban intersections follow each other in quick succession. Their spacing on urban arterials is typically of the order of 500 metres and, in residential areas, spacings of 100 metres and less are not uncommon. These intersections may be between multi-lane roads with medians, channelising islands and a high level of sophistication brought to bear on their signalisation.

In urban areas, the physical size of the various vehicles on the road starts to play a role because they are closer together. Stops and starts are the norm for urban traffic streams, with busses and semitrailers having rates of acceleration significantly lower than those of passenger cars.

It is not possible to endlessly supply more infrastructure to match the increase in vehicles on the road. In terms of the supply/demand equation it is now necessary also to address the demand side of the equation by the promotion of public transport. This requires that roads and streets must be bus-friendly with, where possible, relatively flat grades, generous kerb radii at intersections and careful location of bus stops. Passenger should not have to walk too far to get to the nearest bus stop, with a maximum of 500 metres being a useful target to aim for. This will obviously have an influence on the layout of the residential area.

4.4.3 The subcontexts of design

There are eight subareas within the primary contexts of urban and rural areas (Milton and St Martin, 2005):

- Urban centres
- Urban corridors and nodes

- Suburban corridors and nodes
- Industrial corridors
- Rural town centres
- Transitional areas
- Rural connecting corridors
- Residential areas

They are described in the sections that follow, offering some indication of the features of these areas as they impact on parameters of the transportation facilities they require.

4.4.3.1 Urban centres

The central business districts of large towns and cities are typical urban centres. They are relatively small in relation to the urban areas they serve and are primarily the location of commercial developments and offices. Urban centres include housing, small parks, retail space, recreational opportunities and public transport termini. The employment opportunities offered by urban centres result in high concentrations of traffic converging on them in the early mornings and similar volumes leaving in the late afternoon and evening. Parking is normally off-street in parking garages but short-term parking may be provided on-street.

These centres should be pedestrian friendly, with wide sidewalks, and provide facilities for parking of bicycles. Shelters should be provided for users of public transit. Traffic control would usually be by signalisation.

4.4.3.2 Urban corridors

The major streets within urban corridors often are dual carriageways, with side streets and alley ways serving loading areas, but may also include multi-lane undivided streets. Very often and depending on the volume of traffic to be served, urban corridors take the form of one-way pairs of streets. High-density residential buildings may be located within these corridors, which may also include local concentrations of small shops and municipal libraries and other services as well as community centres.

Urban corridors usually have to accommodate a variety of modes of transport, including

- Passenger cars
- Freight vehicles of all sizes
- Public transit, usually busses but possibly including light rail
- Cyclists
- Pedestrians

Intersection are usually signalised subject to warrants for signalisation being met. Bus stops are frequent and possibly at intervals of 500 metres or less.

4.4.3.3 Suburban corridors and nodes

Suburban corridors differ from urban corridors insofar there is a greater preponderance of passenger cars in the traffic stream. Very often, shops have their own parking areas and access may be via short frontage roads separated from the major road by an island generally referred to as an outer separator.

Roads may have a median, which, where there are many left-turns to off-street parking, may be replaced by two-way left-turn lanes.

Bus routes are normally located on these roads so that bus shelters should be provided at stops. Pedestrian crossings should be provided, preferably in close proximity to these bus stops.

4.4.3.4 Industrial corridors

Industrial corridors have high morning and late afternoon peaks as workers arrive at and then leave their workplaces. Bus volumes as well as cycle and pedestrian volumes are fairly high at these times, with the latter two being low at other times. During the course of the day, the percentage of truck traffic in the stream would tend to be high, with many of these trucks being semitrailers. Industrial corridors are typically truck routes and the geometry of the intersections should be designed to accommodate these vehicles. The major routes are generally arterials with signalised intersections.

4.4.3.5 Rural town centres

The centres of villages and small towns in a rural environment are typically host to a variety of activities such as business, shopping and recreation. Government offices at all three tiers of government, from state to local, could be found at these centres along with other services such as post offices and police stations. Town halls and other public places such as libraries would be located at the town centre. Traffic volumes would be medium but with high volumes of pedestrians and low volumes of buses. Buses would principally provide a link to nearby towns and cities.

Reserve widths tend to be narrow, and some village centres are considered to be historic in nature. There could be a strong argument in favour of vehicle exclusion in these areas, with the road network being to human scale rather than to vehicle scale. This suggests that geometric design would have to be very sensitive to local community values.

4.4.3.6 Transitional areas

At the outskirts of rural towns there may be transitional areas of various types. These areas would be located on roads leading to other towns and villages and could be either residential or commercial in nature. If residential, two possibilities arise: residences could be up-market on large stands or shacks occupied by people who are probably unemployed. The up-market transitional area would tend to have a preponderance of passenger cars whereas shack dwellers would rely on walking or cycling. Because the through route could also serve as a truck route, it would be desirable, in the case of shack dwellers, to provide cycle tracks and pedestrian paths that are well clear of the through route.

In developing countries, shack dwellers often make use of animal-drawn transport, typically donkey carts. These carts often have the same width as a passenger car but travel at no more than a fast walking pace. Because of the speed differential between them and the motorised vehicles, providing paths for them that are clear of the through route would be desirable.

Commercial development would tend towards light industry and small shops. Parking would generally be outside the limits of the road reserve.

4.4.3.7 Rural connecting corridors

These are roads linking rural towns to each other or to urban areas. Design has to take due cognisance of rural land use patterns. Agrarian farming results in high seasonal fluctuations in traffic volumes in the form of farm to market trips whereas pastoral farming does not show these spikes in demand. Intersections are typically widely spaced and have a simple layout. Depending on the topography and hence the alignment of the road, speeds could be high. The safety of the relatively infrequent cyclists and pedestrians on these roads thus suggests that paths close to the reserve boundaries should be considered.

4.4.3.8 Residential areas

Residential areas are served by collectors and local streets. Pedestrian volumes are low but children playing in streets should not be unexpected. Traffic calming measures should be brought to bear on the design of residential streets. Traffic calming comprises reduction in traffic volumes as well as reduction in traffic speeds. Traffic volumes can be kept low by ensuring that residential streets do not offer a convenient alternative to congested arterials. Traffic speeds are generally kept low through the application of speed humps and chicanes. These traffic calming measures are generally a case of retrofitting an existing grid pattern of streets but in a 'green fields' development the same end results can be achieved without the application of essentially artificial measures.

4.5 CONSISTENCY OF DESIGN

Consistency of design is defined as the conformance of a road's geometric and operational features with driver expectancy (Wooldridge et al., 2003). It has long been appreciated that designs demonstrating significant departures from continuation expectancy (summarised as 'What was, will be') tend to have poor crash records. Expectancy suggests that drivers would wish to travel at a constant speed and that side forces experienced on horizontal curves should be low. Over the years, much research has been conducted in an effort to quantify consistency of design. Measures examined include

- Speed measures
- Alignment measures
- Workload measures

4.5.1 Speed measures

4.5.1.1 Design speed

Historically, consistency was evaluated in terms of design speed. Design speed was defined as the maximum speed at which the road could be traversed in safety when only the geometry of the road dictated what this speed could be. A major weakness of design speed as a measure of consistency is that it only defined the minimum geometric standards that could be adopted in the design of a road. Consistency could thus imply that the entire road should have geometric standards that are close to minimum values.

This definition has fallen into disuse and has been replaced by the statement that 'Design speed is the speed selected to determine the various geometric design features of the road for design' (AASHTO, 2004). In short, design speed is the speed selected for design! Design speed is thus no longer a speed per se but rather constitutes a grouping of geometric standards. Consistency of design dictates that it should be logical with regard to the topography, adjacent land use and the functional classification of the road, all of which would have a bearing on expected operating speeds.

4.5.1.2 Operating speed

Operating speed is the speed at which vehicles are observed to be moving during free-flow conditions. The 85th percentile of the observed speed distribution curve is the most frequently used measure of operating speed.

Drivers' selection of speed, especially on lightly trafficked high-speed rural roads, tends to exceed the design speed by a considerable margin. For example, a road with a design speed of 120 km/h may actually have 15 per cent of the vehicles travelling at 140 km/h or more. As it is generally considered to be a fair indication of the level of safety that can be expected, much research has been undertaken to develop predictive models of operating speed. Regression models developed in nine countries for the estimation of operating speed (Lamm et al., 1999, 2001; Operational Effects of Geometrics Committee, 2011)

- Australia
- Canada
- France
- Germany
- Greece
- Italy
- Lebanon
- United Kingdom
- United States

show that the highest operating speed is found in Italy and the lowest in Lebanon, with the others running between these two extremes and generally parallel to them.

These models are all directed towards rural two-lane roads, largely because speeds in urban areas are constrained by speed limits and the presence of other traffic, whereas in rural areas, free-flow speeds can be and are achieved. They are constrained only by the selection of design speed. The majority of crashes occur on two-lane roads and, if the inconsistencies in horizontal alignment could be addressed, the road network would be a safer place than it currently is.

Consistency of design is achieved through a minimum of variation in the speed profile across the length of the road and this is, in effect, an alignment measure.

4.5.2 Alignment measures

In addition to variations in operating speed, speed differentials between various classes of vehicles can also have a negative impact on safety. It is thus not only the variations in the radii of successive horizontal curves that have to be considered but also the impact of gradients on truck speed. Comparison of the actual alignment with recommended standards result in ratings of 'Good', 'Fair' and 'Poor' (Polus et al., 2005).

Some possible alignment indices are suggested in Table 4.1. Of those offered, the Curvature Change curvature change rate has been internationally adopted and is discussed further in Section 4.5.2.1.

4.5.2.1 The curvature change rate

The curvature change rate (CCR; Lamm et al., 2001) is calculated as

$$\mathrm{CCR_s} = \frac{\dfrac{L_{\mathrm{Cl1}}}{2R} + \dfrac{L_{\mathrm{Cr}}}{R} + \dfrac{L_{\mathrm{Cl2}}}{2R}}{L} \cdot \frac{200}{\pi} \cdot 10^3$$

Table 4.1 Alignment indices

Horizontal alignment indices	*Vertical alignment indices*
Curvature change rate (either degrees per kilometre or gons per kilometre, with the European gon = 0.9°)	Vertical curvature change rate
Curve length: roadway length ratio ($\Sigma C/L$)	Average rate of vertical curvature
Average radius of curvature (ΣR/number of curves)	Average gradient
Average tangent length (ΣL/number of tangents)	

where

CCR_s = curvature change rate of the single circular curve with transition curves (gons/km) (Note: The European 400 gons = 360°. The CCR can be expressed in degrees per kilometre by replacing the value 200 in the equation by 180.)

L = overall length of unidirectional curve (m)

L_{Cr} = length of circular curve (m)

L_{Cl1}, L_{Cl2} = length of clothoids preceding and succeeding the circular curve (m)

The CCR, being a change in direction over a unit distance (in this book a kilometre) is the 'bendiness' referred to in some economic analyses. Similar to bendiness is the measure of vertical alignment as 'hilliness', which is the algebraic sum of the rises and falls over a unit distance.

Tangents, being circular curves of infinite radius, acquire CCR_ss of their own if they are independent (long). If they are short, they are simply ignored. To differentiate between long and short tangents, it is necessary to consider the operating speed that can be achieved on them as determined by the preceding curve or will be restricted on it by the following curve. Three possibilities exist.

Case 1: The tangent length is such that it is, at most, just possible on going from a shorter to a longer radius to accelerate to the operating speed of the following curve within the length of the tangent; $T \le T_{min}$.

Case 2: The tangent is sufficiently long to allow acceleration to the maximum operating speed, $V85_{max}$; $T \ge T_{max}$.

Case 3: The tangent length is sufficient to achieve an operating speed higher than that of the succeeding curve but not as high as $V85_{max}$; $T_{min} < T < T_{max}$.

The rate of acceleration or deceleration applied is 0.85 m/s^2, which was determined by the application of car-following techniques in Europe.

Where deceleration from a longer to a shorter radius is involved, it is presumed that the reverse of the above cases applies. For example, in the non-independent Case 1, deceleration to match the shorter radius would commence on the preceding curve.

In the case of a gradient equal to or less than 6 per cent, the operating speed, $V85$, is described by the regression

$$V85 = 105.31 + 2 \cdot 10^{-5} \cdot CCR_s^2 - 0.071 CCR_s$$

with an R^2 value = 0.98.

For gradients steeper than 6 per cent, the regression becomes

$$V85 = 86 - 3.24 \cdot 10^{-9} \cdot CCR_s^3 + 1.61 \cdot 10^{-5} \cdot CCR_s^2 - 4.26 \cdot 10^{-2} \cdot CCR_s$$

with R^2 = 0.88.

These relationships apply to CCR_s values between 0, which corresponds to a tangent and 1600 gons/km (corresponding to a radius of just less than 40 metres), and are approximately the average of the seven national regressions referred to earlier. It is pointed out that in Table 4.1, 360 gons/km is the equivalent of a radius of approximately 176.8 metres and 180 gons/km is equivalent to 353.6 metres.

Three criteria of consistency are offered as shown in Table 4.2, as follows.

- Comparison between design speed and driving behaviour as manifested by variations in operating speed
- Comparison of operating speed across successive elements
- Comparison of side friction assumed for design with that demanded at the operating speed

The process to be followed in applying these criteria is to

- Calculate the CCR_s of all the curves along the length of the road
- Calculate the $V85$ for tangents and all the curves using the formulae shown earlier
- Calculate T_{min} and T_{max} for the tangents between each pair of curves
- Calculate the difference between side friction demanded and that provided in terms of the design speed of the road

With the completion of these calculations the consistency of design can be assessed.

T_{min} and T_{max} are calculated as follows:

$$T_{min} = \frac{\left|(V85_1)^2 - (V85_2)^2\right|}{2 \cdot 3.6^2 \cdot a}$$

Table 4.2 Criteria of consistency

Design class CCR (gons/km)	*Speed difference (km/h)*	*Quality of design*
Criterion 1		
$\lvert CCR_2 - CCR_D \rvert \leq 180$	$\lvert V85_2 - V85_D \rvert < 10$ km/h	Good
$180 < \lvert CCR_2 - CCR_D \rvert < 360$	10 km/h $< \lvert V85_2 - V85_D \rvert < 20$ km/h	Tolerable
$360 < \lvert CCR_2 - CCR_D \rvert$	20 km/h $< \lvert V85_2 - V85_D \rvert$	Poor
Criterion 2		
$\lvert CCR_2 - CCR_1 \rvert \leq 180$	$\lvert V85_2 - V85_1 \rvert < 10$ km/h	Good
$180 < \lvert CCR_2 - CCR_1 \rvert < 360$	10 km/h $< \lvert V85_2 - V85_1 \rvert < 20$ km/h	Tolerable
$360 < \lvert CCR_2 - CCR_1 \rvert$	20 km/h $< \lvert V85_2 - V85_1 \rvert$	Poor
Design class CCR (gons/km)	*Frictional difference*	*Quality of design*
Criterion 3		
$\lvert CCR_2 \rvert \leq 180$	$+0.01 \leq f_2 - f_1$	Good
$180 < \lvert CCR_2 \rvert < 360$	$-0.04 \leq f_2 - f_1 \leq +0.01$	Tolerable
$360 < \lvert CCR_2 \rvert$	$f_2 - f_1 < -0.04$	Poor

Note: The subscripts 1 and 2 refer to the preceding curve and the curve of interest respectively and the subscript D refers to design.

where $V85_1$ and $V85_2$ = 85th percentile speeds (km/h).

$$T_{\max} = \frac{2(V85_{T_{\max}})^2 - (V85_1)^2 - (V85_2)^2}{22.03}$$

The available side friction (AASHTO, 2011a) is assumed as shown in Figure 4.2 and expressed by the relationship

$$f_{RA} = 0.925 \cdot n \cdot f_T$$

where
- f_{RA} = side friction assumed
- n = utilisation factor, which is
  - = 0.40 for hilly or mountainous topography
  - = 0.45 for flat topography
  - = 0.60 for existing or old alignments
- f_T = theoretical maximum side friction

demanded is expressed as

$$f_{RD} = \frac{V85^2}{127 \cdot R} - e$$

where
- f_{RD} = side friction demanded
- R = radius of curve (m)
- e = superelevation rate (%/100)
- $V85^2$ = eighty-fifth percentile speed

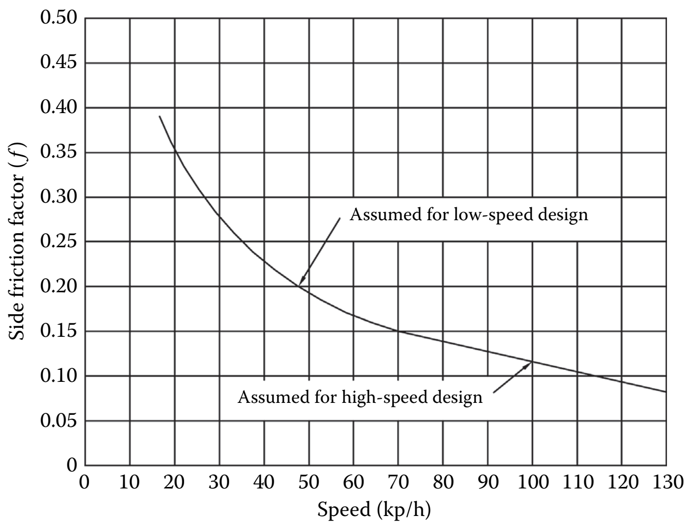


Figure 4.2 Available side friction.

Where the length of the tangent falls between T_{min} and T_{max}, the $V85$ on the curve of interest is recalculated on the basis of the $V85$ of the preceding curve plus or minus the acceleration on the tangent from $V85_2^2 = V85_1^2 + 7.2aT$,

where

a = acceleration (m/s^2) taken as ±0.85 m/s^2 for acceleration and deceleration respectively
T = tangent length (m) with $V85$ expressed in km/h

The Case 1 length is considered to be non-independent because, in going from a shorter to a longer radius curve, acceleration to the higher speed will continue on the following curve. The other two cases are both regarded as independent because they involve speeds higher than those on the adjacent curves.

The CCRs of the curves along the road are calculated as the first step towards evaluating consistency of design. The procedure to follow is thus as follows.

- Calculate the CCR_s of all the curves along the length of the road
- Calculate the $V85$ for tangents and all the curves
- Calculate T_{min} and T_{max} for the tangents between each pair of curves
- Calculate the difference between side friction demanded and that provided in terms of the design speed of the road

With the completion of these calculations the consistency of design can be assessed.

If the tangent length is shorter than T_{min}, Case 1 applies; the tangent is ignored and the comparison is then between the $V85$ speeds of the two curves. If the tangent length is greater than T_{max}, comparison is between the $V85$ of the preceding tangent and that of the curve. For a tangent, CCR = 1.

In the case of a consistent design, all three criteria will have been met in the category 'Good'. Where some or maybe all criteria fall into the group 'Tolerable', the design would need to be revised to improve the rating to 'Good' as, otherwise, it would be considered to be barely acceptable. Values falling in the category of Poor are unacceptable and call for redesign. Where one set of values falls outside the category achieved by the others, it will be necessary either to eliminate the mismatch or to consider some or other ameliorating measures, such as traffic signing or other warning to the driver of an impending change in circumstances.

4.6 THE DESIGN DOMAIN

4.6.1 Overview

The concept of design domain emphasises cost-effective designs rather than a design that simply meets 'standards'. It offers a range of values of standards from which designers select design criteria. This range is illustrated in Figure 4.3. In the case of standards such as gradient and deceleration, it is the low values that are the more acceptable whereas for radius of horizontal curvature or K value of vertical curvature, the higher values are to be preferred.

The high value is typically dictated by the fact that a further increase in value will occur, while implying an increase in construction cost is not warranted by an increase in benefit. An example is lane width. A lane width of 3.7 metres seems to be the maximum useful value, as greater widths do not reflect any corresponding increase in capacity. There is also some evidence suggesting that crash rates are higher on 3.7 metre wide lanes than on lanes that are 3.3 metres wide.

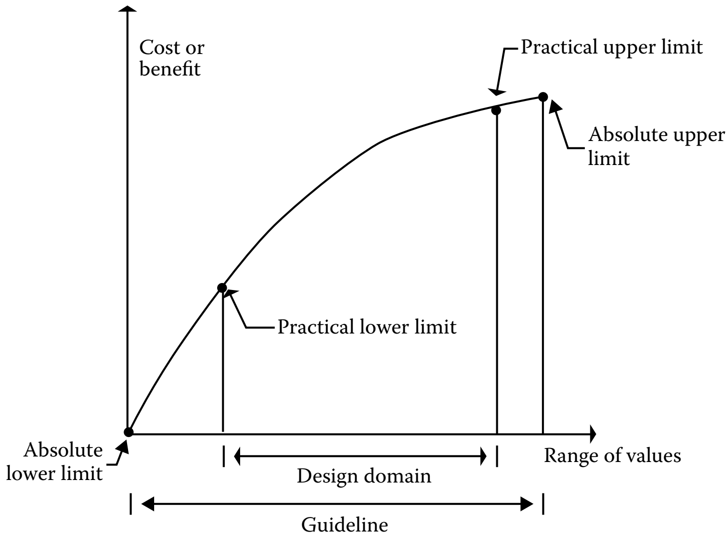


Figure 4.3 The design domain. (From Transportation Association of Canada [TAC]. *Geometric design guide for Canadian roads*. Ottawa, 1999.)

The low value would generally have a lower construction cost than the higher values of the criterion and be the minimum that would be required in terms of vehicle dynamics or driver preferences. Radius of horizontal curvature and stopping sight distance are examples of low minimum values, both of which relate to the friction available between a vehicle's tyres and the road surface. In this case, the curve of cost or benefit versus standard would be the mirror image of Figure 4.3.

Values of standards at the higher end of the range would be adopted for roads with high traffic volumes or where the higher standards can be achieved with little additional outlay. Standards at the lower end of the range would have to be accepted in the presence of physical constraints such as the topography of mountain passes or dense urban development.

The figure shows four possible values that a standard could adopt, as follows.

- The absolute lower limit
- The desirable lower limit
- The desirable upper limit
- The absolute upper limit

The desirable limits are those that would have their foundation in driver preferences or perceptions, that is, a human factors approach to design, whereas the absolute limits are those dictated by the limits of the interaction between the vehicle and the road. An absolute minimum value of horizontal radius would be that radius at which the demand for side force friction just matches the side force friction provided. However, the values of minimum horizontal radius historically adopted for design include a substantial safety factor so that these aren't true minima.

Cost or benefits are compared to the value of standard adopted. An analysis leading to selection of shoulder width, for example, would require consideration of

- Construction cost
- Maintenance cost

- Road user cost
- Environmental impact
- Safety
- Mobility

There are thus six curves of benefit compared to range of values that have to be considered. Unfortunately, these are unique to every design project and it is thus not possible to determine definitive values of standards that will apply to all cases. It is, however, recommended that designers should at least attempt to go through the exercise of determining realistic values of standards for application to their specific project.

Values of standards quoted elsewhere in this Handbook tend to be based on possibly only one of these characteristics, with that characteristic typically being safety. They also conform to international usage. Many road authorities or agencies have developed design manuals that address their own unique circumstances. Regrettably, these manuals tend to be prescriptive in nature and do not allow the designer any leeway in deciding what values of standards should be applied to a specific design.

Inexperienced designers are inclined to consult the relevant manual and then proceed to execute a design using minimum values of standards throughout. As a general rule, the use of minimum values, particularly combinations of minimum standards, should be avoided if at all possible. It could be found, if the acceptable minima relating to the horizontal and vertical alignment and the cross-section are all applied at one point, that the crash record at that point would be unacceptable. If it is necessary to accept a minimum value of one standard, it may be necessary to increase one of the others. For example, if the radius of a horizontal curve is very small, it may be necessary to increase the lane width on the curve to allow for the off-tracking that occurs between the front and rear wheels of a vehicle. Off-tracking is particularly marked in the case of articulated vehicles.

4.6.2 3R projects

There are few, if any, road authorities that have sufficient funding to undertake all that is required of them by the road user. In countries where the road network is essentially complete, many of the links in the road network may already have achieved their design life or possibly even be well beyond it. The activities of the road authorities would thus tend to focus on rehabilitation of the existing road network rather than on green fields design and construction.

Rehabilitation is principally of the design layers of the pavement and reference is made to 3R projects. These are aimed at extending the life of the road through maintenance of the pavement layers at the levels of

- Resurfacing, which is the provision of a new wearing course to the existing surface with this wearing course being anything from a fog spray with bituminous emulsion to an asphalt overlay.
- Restoration addressing failed sections of the roads where failure may be limited to potholes requiring repair or extend to several metres.
- Rehabilitation where the entire base course layer, and possibly the lower layers as well, would be scarified, reshaped and compacted and a new wearing course constructed over the rehabilitated layer(s).
- Reference is sometimes made to 4R projects with the fourth R referring to reconstruction. In this case, the existing road is written off in its entirety and its only use is as a source of construction material. Construction is from the ground up, including the bulk fill.

Whereas the design life of a pavement is typically of the order of 20 years, road geometry is long lived. For example, alignments built by the Romans are still in service today. It is hardly surprising that geometric improvements would take second place to the urgency of pavement improvements. Consideration is given to geometric improvements only in the cases of rehabilitation and reconstruction. Unfortunately, opportunities for safety improvements sometimes have to be foregone because of the overriding importance of pavement improvement. Improvement of the pavement takes place from end to end of the road whereas geometric improvements are limited to possibly isolated substandard horizontal and vertical curves. Limited funding forces maintenance engineers to have to make choices between improving the pavement from end to end of the road or to divert some funding to the improvement of local substandard geometry and risk leaving some of the inadequate pavement untouched.

The arguments in favour of accepting substandard geometry can be telling. For example, it may be found that although a crest curve provides inadequate sight distance, the crash record of the curve is within acceptable limits. Expectancy suggests that driving is in accordance with what is seen.

In fact, if 'a road is obviously dangerous, it becomes safe'. As an illustration of this contention, Bain's Kloof in the Western Cape Province of South Africa is a narrow mountain pass with very short radii of curvature. With increasing traffic volumes, National Route N1 was relocated to go through du Toit's Kloof. The cross-section was wider and radii of curvature conformed to a design speed of 60 km/h. The wider cross-section apparently encouraged drivers to select too high an operating speed because the fatality rate on du Toit's Kloof in the first year of its operation exceeded the fatalities over the preceding 80 years on Bain's Kloof. du Toit's Kloof was built in the 1940s so that the first years of travel over Bain's Kloof were by ox wagon and then by motor cars capable of only low top speeds.

Unfortunately, as stated earlier, law courts incline to the view that a road designed to standards is safe and, by extension, a road not to standard is unsafe. A road authority deliberately accepting substandard geometry runs the risk of unfavourable litigation. It must be able to demonstrate in a court of law that it has 'applied its mind' in arriving at a decision to accept the substandard geometry.

4.6.3 The extended design domain

The concept of the extended design domain (EDD), illustrated in Figure 4.4, addresses this issue. EDD is a range of values below the lower bound of the normal design domain and thus below the minimum values currently used in most road design guidelines. The scope to use lower minimum values comes when the normal design domain contains considerable latitude. This is usually due to some obsolete technical foundation or unnecessarily conservative assumptions, or a combination of these. The use of 85th percentile values is widespread in the field of geometric design, but the combination of several 85th percentile values results in a standard that is extremely conservative and well within the capabilities of the 85th percentile driver. Basically, excessive safety margins accepted in the past are being deployed in the search for cost-effective solutions to problems of pavement restoration. Credit must go to the Queensland Department of Transport and Main Roads (2004) for the development of the concept of the EDD.

EDD is applied to consideration of sight distance and hence to horizontal and vertical curvature as well as to intersections. Sight distance is based the parameters of driver eye height and object height. The United Kingdom, Australia and southern Africa have adopted a driver eye height of 1.05 metres and the United States has opted for a driver eye height of 3.5 feet (1.08 metres). The object height depends on circumstances as illustrated in Table 4.3.

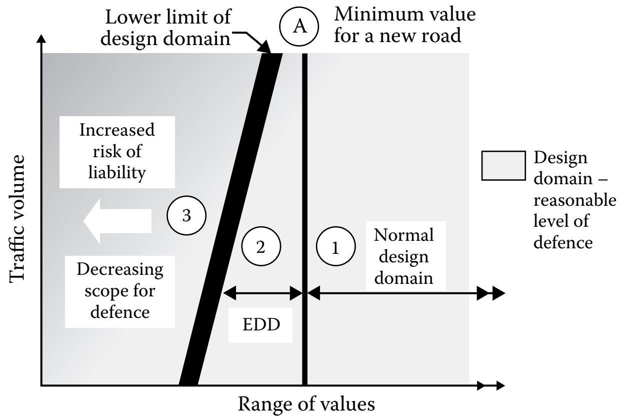


Figure 4.4 Conceptual diagram of extended design domain. (From Queensland Department of Transport and Main Roads. *Road planning and design manual*. Brisbane, 2004.)

Table 4.3 Application of object height

Object height (m)	*Applicability*
0.00	Pavement markings in critical locations Risk of washouts or large potholes
0.15	Risk of fallen trees or rocks Risk of debris from trucks or other cars
0.60	Stopping behind a stopped vehicle based on vehicle tail or brake light
1.08	Passing sight distance Intersection sight distance

Source: Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited, Pretoria, 2002.

The 85th percentile height of a passenger car is 1.33 metres. On the basis that there should be enough of the object to be visible for the driver to perceive it as a hazard, some countries have adopted a height of 1.15 metres for passing and intersection sight distance. The height adopted by the United States (AASHTO, 2004) is 1.08 metres, which allows for 250 mm as the height that needs to be visible for another driver to perceive the object as a vehicle.

The application of the extended design domain has led to the introduction of yet another object height being 0.40 metres, which, not to be confused with Hauer's (2000) height of a dead dog of 0.15 metre, apparently is the height of a dead kangaroo (Cox and Arndt, 2005)!

The extended domain is based on the need for a manoeuvre capability if stopping sight distance is not available for an object with a height of 0.15 metres (0.20 metres in Australia). The distance remaining after the initial reaction time has expired allows for a lateral movement and diverge angle similar to the exit taper on an interchange. It is also possible, without heavy braking, to achieve a 20 km/h reduction in speed before reaching the hazard.

Normal braking, that is, the rate of deceleration which passenger would find comfortable, is of the order of 3.0 m/s^2. Stopping distance as defined in the literature assumes fairly smooth tyres on a wet polished surface and average braking capability and a range of frictional coefficients dependent on speed. The values of stopping distance, without the presence of the current substantial safety factor, thus equate to a panic stop. On a dry surface, a panic stop could employ a deceleration rate double the comfortable norm, that is, 6 m/s^2. This is a rate consciously used by only a few drivers and matches the intent of the stopping sight distance model. On a wet surface, the required deceleration rate should not exceed 4.5 m/s^2.

Manoeuvring requires sufficient in which to manoeuvre. The overall width of a passenger car is in the range of 1.8 to 2.1 metres. This width should be available for the passenger car to pass the object without stopping. If there isn't sufficient sight distance to allow for a stop before an object with a nominal height of 0.40 metres, the shoulder and the fill slope (at a batter of less than 1*V*:4*H*) should provide a width of at least 1.5 metres. This is on the assumption that the object does not fill the entire width of the lane. It is also possible that the driver may elect to manoeuvre into the opposing lane. It would not be practical to provide the sight distance required to accommodate this eventuality because that would be the presence of passing sight distance across the full length of the road.

4.6.4 Design exceptions, variances and waivers

In recognition of the possible need in the case of rehabilitation projects to adopt standards less than those employed on new construction, many road authorities have adopted a formalised procedure for the determination and approval of the standards actually adopted (Texas Department of Transportation, 2010). Although any standard employed in design could become the subject of litigation, substandard road features almost invite litigation. Where a need for entering the EDD is realised, the designer is required to bring this to the attention of the relevant authority and provide a rationale for the adoption of substandard criteria.

Procedures vary between road authorities and there is no standardised nomenclature in existence. In general, exceptions refer to critical design elements and variances to noncritical elements. The waiver is the document that describes the procedure adopted in the evaluation of the particular exception or variance and the motivation for adoption of a value lower than the norm. The motivation could include the cost implications of reconstruction to values currently accepted and also the historic crash rate prevailing at the site where the relaxation of standard is being considered.

Critical elements are those with the potential to result in serious crashes and include:

- Design speed
- Stopping and passing sight distance
- Horizontal and vertical curvature
- Superelevation rates
- Lane and shoulder widths
- Camber and cross-slopes
- Horizontal and vertical clearances
- Bridge widths and structural capacity

The information that is normally required to be provided in support of a request for a design exception includes

- The precise location of the design feature that does not meet the standard specified for the normal design domain, including a location sketch map
- The geometric standard proposed for consideration of the design exception
- The gap between the proposed and the normal standard and an explanation of why the normal standard cannot be achieved

- Estimates of the cost of acquisition of the road reserve and construction for provision of the design feature to the standard specified for the normal design domain compared that of the requested design exception
- Estimates of base year and design year traffic volumes including an estimate of the percentage of truck traffic
- Crash history
- Any safety enhancements or mitigations that will be included if the feature is not constructed to specified standards

Noncritical elements such as median widths or channelising islands and various other features of intersections and interchanges, which are potentially less hazardous than the critical elements, are the subject of design variances. The information required is as described earlier.

The extent of the risk to which the department is likely to be exposed in the event of litigation is less in the case of the variance than for the exception but both are exercises in risk management. The point at issue is that the department must be able to demonstrate in a court of law that it has 'exercised its mind' in the granting of the exception or variance.

Chapter 5

Design control

CONTENTS

5.1 INTRODUCTION

A road is a three-dimensional object that supports the transportation of goods and people in a safe, convenient and economical manner with a minimum of environmental side effects. This movement takes place in a range of vehicles of various sizes with differing performance levels and guided by individuals who also demonstrate a wide range of skills levels and characteristics.

The design process must therefore consider three variables – the driver, the vehicle and the road – as well as the interaction between them and the limitations of each. The main feature of all these characteristics is that the designer has no control over them. They are thus inputs to the model, the output of which is the design of all the elements of the road, from the horizontal and vertical alignment to the cross-section and the intersections which are the nodes in the links that form the road network.

5.1.1 Aspects of speed

In this chapter, the driver and the design vehicle are discussed. Knowledge on these two components of the design process leads to consideration of various aspects of speed. These aspects include

- Average speed
  - Space mean speed
  - Time mean speed
- Design speed
- Operating speed
- Running speed
- Desired speed
- Posted speed

5.1.2 Aspects of sight distance

With speed in all its forms known, it is possible to develop criteria of the most fundamental criterion of safe design, which is sight distance. With the variety of operational aspects of safety in design, it follows that, like speed, sight distance can and does assume many forms. These include

- Manoeuvring sight distance
- Stopping sight distance including the effects on it of
  - Horizontal curvature
  - Vertical curvature comprising crest and sag curves
  - Gradient and
  - Combinations thereof
- Headlight sight distance
- Passing sight distance

- No-passing (or barrier) sight distance
- Decision sight distance
- Headlight sight distance
- Passing sight distance and
- Decision (also known as anticipatory) sight distance

These are all comprehensively dealt with in the sections that follow.

The design process has to consider two levels of operation of vehicles on the road. The human factors discussed in Chapter 4 and the various issues of speed and sight distance characterise the first of these two levels at the microscopic level. This relates to the behaviour and characteristics of individual drivers and vehicles, individual particles in a moving stream if you will. The second is the macroscopic level in which the entire stream is considered, analogous to hydraulic flow in a pipe.

5.1.3 Aspects of traffic

Proper design must, of necessity, consider the totality of flow and the various aspects thereof. These include

- Traffic flow volume and rates both in the present and in future thus including consideration of
  - The design hour and
  - The design year
- Traffic composition in the sense of the percentages of the various types of vehicle in the traffic stream
- The directional split in traffic volumes in the opposing directions of flow

The relationship between flow and delay and also between flow and hour of the year require consideration.

Traffic flow theory addresses both the microscopic and macroscopic aspects of flow and is dealt with in broad outline in Chapter 21.

5.2 THE DRIVER

The driver was discussed in some depth in Chapter 4. Although most of the driver's characteristics relate to the way in which the various components of the road network slot together to create a harmonious whole, there are two prime characteristics that impact the calculation the values of individual design standards. These are reaction time and driver eye height and they dictate the vitally important sight distance standards.

Reaction time can vary from the 0.15 to 0.20 second of the fighter pilot and the Formula 1 racing driver to the 3.5 (and more) seconds of the elderly driver. Elderly drivers are classified into two groups: 70 to 79 and the 80 to 89 age groups. In countries with aging populations, these two groups form a not insignificant proportion of the total.

Smiley et al. (2008) undertook a comparison of the actions of these two groups with those of the middle-aged group with respect to driver actions leading to crashes. Older drivers (older than 70 years) had significantly fewer rear end crashes than the very old (older than 80 years) and the middle-aged groups. Members of both older groups were less likely to have run-off-road crashes at intersections than those of the middle-age group. As age increased, so did the proportion of crashes as a result of failing to yield right of way. With respect to errors, drivers aged 80 years and older made significantly more search and detection errors (inadequate search, inattention, distraction, overload, obstruction or other) than the other two age groups combined. Inadequate search errors increased significantly with age, from 27 per cent for middle age to 65 per cent for older age drivers. The old and very old groups had significantly fewer distraction errors (11 and 9 per cent, respectively) than the younger group (27 per cent). Very old drivers made significantly more evaluation errors than the other two groups combined. Approximately 90 per cent of these errors were misjudging other vehicle's actions rather than intersection design (i.e. misunderstood lane designations or right-of-way). Both groups of elderly drivers made significantly fewer unintended course errors and vehicle action errors (where the vehicle does not respond owing to poor weather or vehicle malfunction) than the middle group. In part, the performance of the elderly drivers is reflected in their reaction times.

The 85th percentile reaction time that has been adopted internationally is 2.5 seconds. This allows for conditions more complex than those in the laboratory. It also makes some provision for the elderly driver. It does not allow for the most complex conditions, such as those found, for example, at signalised intersections and the on-ramps at freeway interchanges. A further allowance is added to the basic 2.5 seconds to make provision for these conditions and is discussed in Section 5.4.5.

Driver eye height has decreased over the years as passenger car heights have been decreased. The internationally accepted eye height shows a modest variation between 1.05 metres and 1.10 metres above the road surface, with the American preferred eye height being 3.5 feet, which converts to 1.08 metres. In this document, standards are calculated on the basis of a driver eye height of 1.05 metres. For large trucks, the driver eye height varies internationally from 1.8 metres to 2.4 metres. For design purposes the American driver height is 7.6 feet above the road surface, corresponding to a metric height of 2.33 metres. In this document, an eye height of 2.0 metres, as adopted in the United Kingdom, is used. For interest, the driver eye height adopted by Australia is 2.4 metres and, by southern Africa, 1.8 metres. Truck driver eye height is of particular importance in assessing sight distance as obscured by an overhead object such as a bridge deck.

5.3 THE DESIGN VEHICLE

There are two major groups of vehicles that have to be accommodated on the road network: motorised and nonmotorised vehicles. The characteristics common to both is that they occupy space and have performance or operational characteristics. Unlikely though it may seem, the pedestrian, occupying a physical space and having a speed of movement, also qualifies as a 'design vehicle'.

5.3.1 Motorised vehicles

Vehicles are classed into different groups based on their function and hence physical dimensions and characteristics. Vehicles fall into four distinct groups:

- Passenger cars
- Busses
- Trucks
- Recreational vehicles

The motorised design vehicle does not exist as something that can be bought. It is a composite of the many parameters that define a vehicle. In each group, the design vehicle is usually a composite of the 85th percentile values of each of the parameters. The reason for not designing roads for the 100th percentile vehicle is that it would simply be uneconomic and not economical of space to accommodate, for example, the vulgarly ostentatious stretch limo. The dimensions of the vehicle are its

- Length
- Width, overall (excluding wing mirrors) and track width (between the outsides of the tyres)
- Height
- Ground clearance
- Wheel base
- Front overhang
- Rear overhang

The dimensions of various design vehicles are shown in Table 5.1.

The operational characteristics of a design vehicle are its

- Power-to-weight ratio
- Braking capability
- Deceleration
- Minimum radius turning circle
- Swept path

The power-to-weight ratio, particularly of trucks, is used to derive the truck speed on gradients, which is an input into the decision regarding the provision of an auxiliary lane. This lane is variously referred to as a crawler lane or a climbing lane and sometimes, incorrectly, as an overtaking or passing lane. The last mentioned is totally different in its function. A climbing lane seeks to balance the level of service on the climbing grade with that on level grades whereas the passing lane improves the capacity of the road and hence the levels of service also. As an extreme example, the four-lane road could be described as a two-lane road with continuous passing lanes in each direction.

The braking capability of most vehicles is such that, without the application of antilock braking systems, the wheels could lock. Since the application of the anti-lock braking system to trucks, the difference in stopping ability between trucks and passenger cars has virtually fallen away. Trucks with conventional braking systems require longer stopping distances from a given speed than do passenger cars. The American Association of State Highways and Transportation Officials (AASHTO, 2004) suggests that the truck driver is able to see the vertical features of the obstruction from substantially further back because of the higher driver eye height. There is, however, evidence to suggest that the sight distance advantage provided by the higher driver eye level in trucks does not always compensate for their inferior braking. Some reasons for the longer truck braking distances include

- Poor braking characteristics of empty trucks. The problem relates to the suspension and tyres that are designed for maximum efficiency under load
- Uneven load between axles
- Propensity of truck drivers not to obey posted speed limits
- Inefficient brakes of articulated trucks

Maximum deceleration rates depend on the roughness and wetness of the road surface in combination with the condition of the vehicle's tyres. On a dry, rough surface, a vehicle can decelerate at a maximum of about 8 m/s^2 and drivers sometimes elect to apply a decelerate rate

Table 5.1 Typical international design vehicle dimensions (m)

Design vehicle parameter	*United States*	*Portugal*	*Germany*	*Southern Africa*	*Australia*
Passenger car (P)					
Height	1.3	–	1.51	1.3	
Width	2.1	1.80	1.76	1.8	1.9
Length	5.8	4.80	4.74	4.8	5.0
Wheelbase	3.4	2.80	2.70	3.1	2.9
Front overhang	0.9	1.00	0.94	0.7	0.9
Rear overhang	1.5	1.00	1.10	1.0	1.2
Single-unit truck (SU)					
Height	4.1	–	4.0	4.1	
Width	2.4	2.6	2.6	2.5	2.5
Length	9.2	10.40	12.0	9.1	8.8
Wheelbase	6.1	5.36		6.1	5.0
Front overhang	1.2	1.40		1.2	1.5
Rear overhang	1.8	3.64		1.8	2.3
Bus (S-BUS 12)					
Height	3.2	–		4.6	
Width	2.4	2.47		2.6	2.5
Length	12.2	12.10		12.3	12.5
Wheelbase	6.1	6.19		7.6	6.6
Front overhang	2.1	2.72		2.1	2.4
Rear overhang	4.0	3.19		2.6	3.5
Articulated bus (A-BUS)					
Height	3.4	–			
Width	2.6	2.55			2.5
Length	18.3	17.99			19.0
Wheelbase	6.7	5.79 + 2.33 + 5.35			6.2 + 2.7 + 5.0
Front overhang	2.6	2.72			2.6
Rear overhang	3.1	1.8			2.5
Semitrailer (WB-19)					
Height	4.1	–		4.1	
Width	2.6	2.62		2.5	
Length	20.9	10.40	18.8	17.0	
Wheelbase	6.6 + 12.3	5.36		6.9 + 9.4	
Front overhang	1.2	1.40		0.9	
Rear overhang	0.8	3.64		0.6	
Semitrailer (WB-20)					
Height	4.1	–			
Width	2.6	2.6			2.5
Length	22.4	17.1			25.0
Wheelbase	6.6 + 13.4	4.1 + 8.0			5.4 + 13.9
Front overhang	1.2	1.40			1.6
Rear overhang	1.4 – 0.8	3.6			4.5

Table 5.1 (Continued) Typical international design vehicle dimensions (m)

Design vehicle parameter	*United States*	*Portugal*	*Germany*	*Southern Africa*	*Australia*
Interlink (WB22) (B-double)					
Height	4.1	2.70		4.1	
Width	2.6	18.10		2.5	2.5
Length	22.0	5.46 + 6.34 + 5.65		22	25.0
Wheelbase	5.3 + 12.2 + 12.2	1.40		4.0 + 6.9 + 8.2	4.5 + 8.0 + 8.7
Front overhang	0.7	0.40		0.9	1.6
Rear overhang	0.9	0.6		0.6	1.4

of 6.0 m/s^2. For normal braking that the passengers in a car would find comfortable, only half of this rate, that is, 3 m/s^2, is recommended and this value has been used in this document.

The minimum turning circle radius is applicable only to crawl speeds and certainly not faster than about 15 km/h. It is applied to intersection design and also to off-street parking areas. The swept path of the vehicle, also referred to as the wall-to-wall path, is defined by the path of its outer front overhang and its inner rear wheel. The wheel track is defined by the path of the outer front wheel and the inner rear wheel. In designing an intersection, commercially available software packages should be used to define the kerb lines of the intersection and its various channelising islands. Two illustrative turning templates are shown in Figures 5.1 and 5.2.

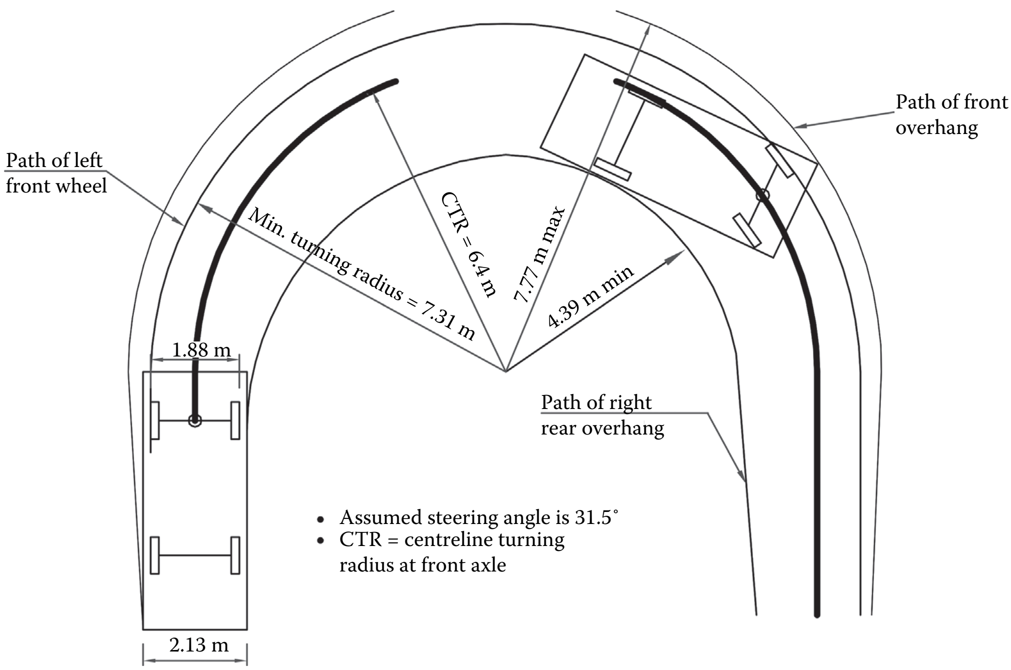


Figure 5.1 Minimum turning path for a passenger car. (From American Association of State Highways and Transportation Officials. *A policy on geometric design of highways and streets*. Washington, DC, 2011a.)

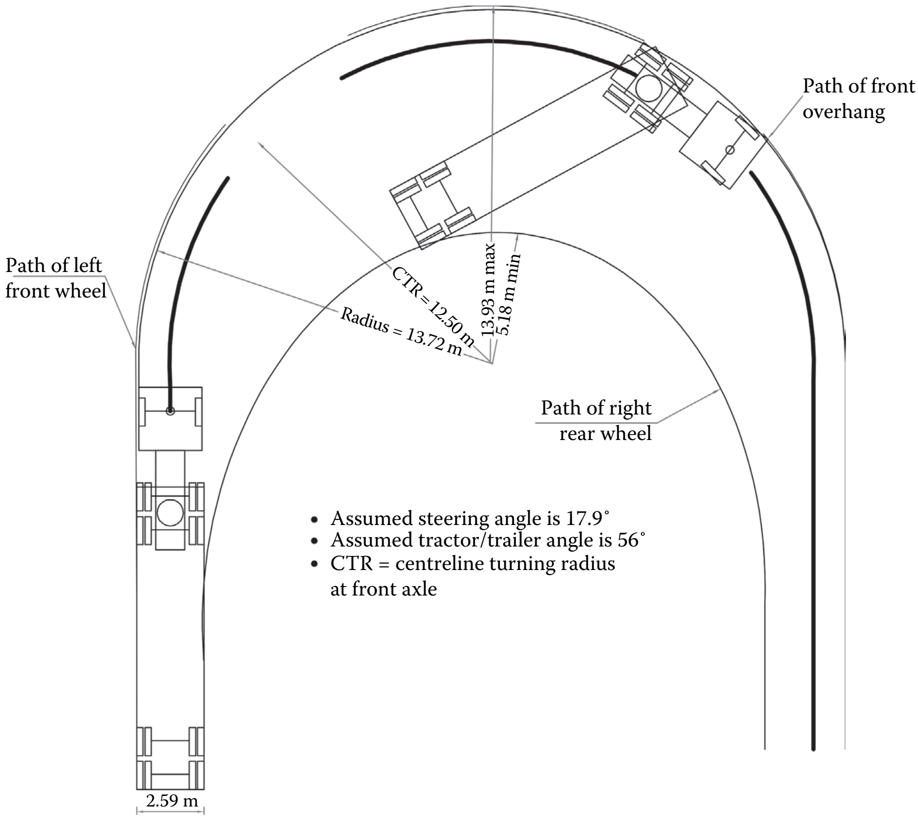


Figure 5.2 Minimum turning path for a WB-15 semitrailer. (From American Association of State Highways and Transportation Officials. *A policy on geometric design of highways and streets*. Washington, DC, 2004.)

5.3.2 Nonmotorised vehicles

Nonmotorised vehicles include

- Cyclists
- People pushing prams
- Pedestrians
- People confined to wheel chairs
- Animal-drawn vehicles

5.3.2.1 Cyclists

Cyclists very often have to take their chances by occupying the lanes used by motorised vehicles. Paved shoulders or widening of the travelled lane by 1 metre would improve their safety. They would, however, be better served by having lanes demarcated exclusively for their use and even better by being provided with cycle paths that are removed from the normal carriageway. A path that is 2.4 metres wide would allow for two cyclists riding abreast or passing each other either in the same or opposite directions.

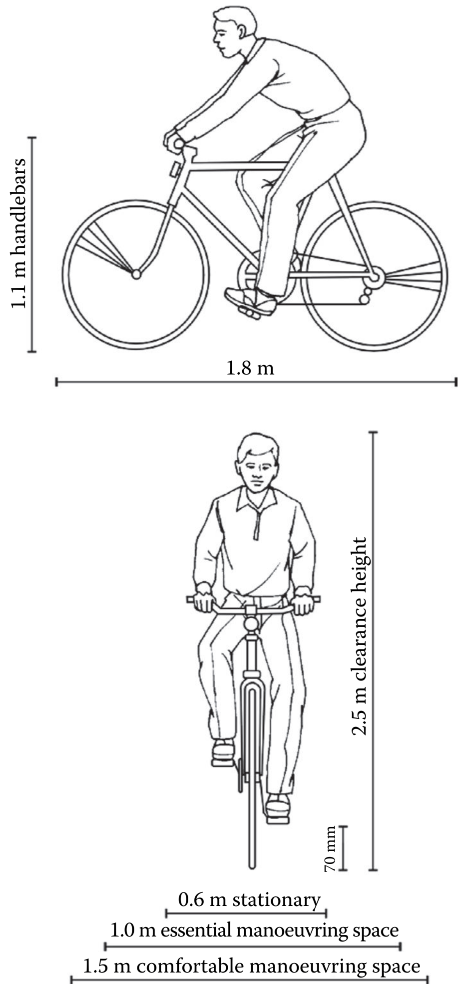


Figure 5.3 Dimensions of a cyclist. (From South African Department of Transport. *Pedestrian and bicycle facility guidelines*. Pretoria, 2003.)

The spatial requirements of a cyclist are illustrated in Figure 5.3.

Cycle paths should have a crossfall of 2 per cent for drainage. They are usually not super-elevated but, on curves the crossfall could be towards the inside of the curve. Small rocks on a gravelled surface can cause a painful fall so that cycle paths should be surfaced and maintained to provide a clean smooth riding surface. Where cyclists are required to cross a dual carriageway, the median should be sufficiently wide to accommodate the length of the bicycle.

5.3.2.2 Pedestrians

If others are standing too close to them, people feel as though their 'space is being invaded'. This space does not have fixed dimensions but varies with circumstances. For example, in a crowded elevator, people do not object to others standing very close to them, whereas on a public street, a person this close would probably be interpreted as having other than desirable intentions. Various levels of service are illustrated in Figure 5.4.

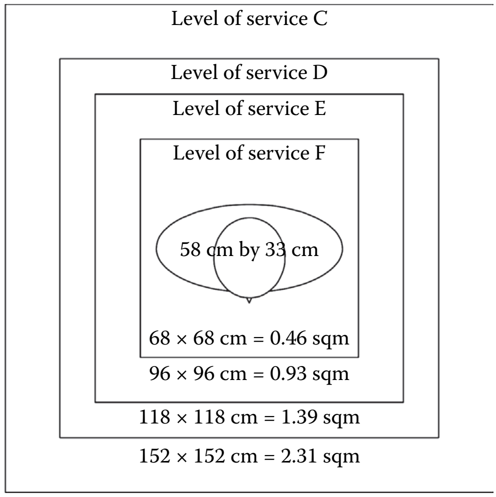


Figure 5.4 Spatial requirements of pedestrians. (From South African Department of Transport. *Pedestrian and bicycle facility guidelines*. Pretoria, 2003.)

Pedestrians have an 85th percentile speed of about 1.3 m/s. However, at an average of about 1.0 m/s elderly pedestrians generally walk slower than this. Near hospitals and retirement villages or homes, this slower walking speed should be selected for design. This is of importance in the design of pedestrian crossings. A four-lane undivided road could have a width of about 15 metres. An elderly pedestrian would thus require 15 seconds to cross it. If a road had sufficient traffic volumes to warrant this width, the likelihood of having gaps of this magnitude with any sort of frequency would be low. Providing a median island at the pedestrian crossing may be an option but would require careful design to ensure that motorists were not surprised by the sudden appearance of an island in their travel path. A safer option would probably be to signalise the crossing.

5.3.2.3 People confined to wheelchairs

A wheelchair occupies a space of 750 mm × 1200 mm. A minimum width of 1.8 metres would be required for two wheelchairs to pass each other or to make a 180° turn in a circular motion. Ramps should be provided at pedestrian crossings. Median islands should also be provided with ramps, although a cut through at grade would be more convenient, particularly in the case of narrow median islands.

People in wheelchairs are often pushed by others and, even if they are self-propelled, usually can manage to maintain a comfortable walking speed.

5.3.2.4 Animal-drawn vehicles

Animal-drawn vehicles are essentially a manifestation of Third World countries and the animals concerned are usually donkeys or water buffaloes. The carts are typically homemade two-wheeled vehicles, with the wheels and axle being salvaged from a scrapped passenger car. This provides the clue that the 'design width' of these vehicles would be 1.8 metres, as in southern Africa. The smart coach drawn by a high stepping horse does not feature and travel, at best, is at walking speeds. Adding animal-drawn vehicles to a traffic stream of fast moving passengers cars and trucks is a recipe for disaster but, without specific provision being made for them, that is precisely where these vehicles would be.

Limited space makes it difficult to make separate provision for animal-drawn vehicles in urban areas. If they are present in any numbers, about all that would be possible would be to appropriate some of the cross-section from the pedestrians to widen the travel lanes above the normal maximum of 3.7 metres. If the road has a four-lane cross-section, some width could also be taken from the central lanes and allocated to the outside lanes. The cross-section could then have outer lanes that are 4.1 metres wide with 3.3 metre wide inner lanes.

On the outskirts of towns, where space is not at such a premium, a 2 metre wide shoulder could be provided. It should, for preference, be unsurfaced or gravelled to ensure that motorised vehicles do not use it.

5.4 SPEED PARAMETERS

One of the most fundamental inputs into the design process is speed of travel. It determines

- The sight distance required for manoeuvring, stopping and passing
- Minimum radius of horizontal curvature
- Minimum *K*-value of crest curves

It also suggests desirable lane widths.

There are numerous forms of speed measurement:

- Average speed
- Design speed
- Operating speed
- Running speed
- Desired speed
- Posted speed

Of these, design speed and operating speed find application in geometric design. Average speed is used in various analytical processes; the differences between running speed and desired speed provide an indication of the level of congestion on the road; and posted speed is a legal limitation of travel speed.

The various forms of speed measurement are discussed in the sections that follow.

5.4.1 Average speed

Average speed is one of the fundamental characteristics of traffic flow and is significant to the road user in terms of travel time, economy, safety and service. A distinction is drawn between microscopic mean speed and macroscopic mean speed. The former is the average of the individual speeds of a sample of vehicles and is also referred to as 'time mean speed'. The latter, also referred to as 'space mean speed' or 'harmonic mean speed', refers to the average travel times across the length of the roadway. It is the sum of the distances covered by all the individual vehicles measured divided by the sum of all their travel times. These relationships are shown in the formulae that follow.

Space mean speed is applied to the determination of delay experienced by the traffic stream and is a measure of the quality of the service provided by the road. Delay features as an element of road user cost in economic analyses.

Microscopic mean speed is calculated as

$$V = \frac{1}{n}\sum_{i=1}^{n}\frac{L_i}{t_i}$$

where
n = number of vehicles
L_i = length of roadway traversed by the ith vehicle
t_i = travel time of the ith vehicle

Macroscopic mean speed is

$$U = \frac{\sum L_i}{\sum t_i}$$

with the symbols as before.

The relationship between microscopic and macroscopic speeds is given (van As and Joubert, 1993) as

$$V = U + \frac{S^2}{U}$$

with

$$S^2 = \frac{\sum K_j(U_j - U)^2}{\sum K_j}$$

where
V = microscopic mean speed
U = macroscopic mean speed
S = standard deviation of macroscopic speed
U_j and K_j = macroscopic speed and density of a substream, j, in which all vehicles all travel at the same speed U_j

5.4.2 Design speed

Design speed used to be defined as the maximum speed at which a road section could be traversed when conditions are so favourable that only the geometry of the road prevailed. As discussed in Chapter 4, design speed constituted the first attempt to achieve consistency of design. Unfortunately, many vehicles, principally passenger cars, exceed the design speed by a considerable margin, with the 85th percentile speed sometimes being 10 km/h higher than the design speed. It should, however, be noted that trucks exceeding the design speed by as little as 5 km/h could roll over on a minimum radius curve because of their high centre of gravity.

Design speed is now simply defined as the speed selected for design, which actually is no definition at all. In practice, design speed is no longer a speed but simply a grouping of design standards with a common base in Newtonian physics. Design speed specifies the minimum value of the various design standards so that a road with very high standards but including

Table 5.2 Typical design speeds

Road type	*Design speed (km/h)*
Limited-access roads	
Freeways in urban areas	90–130
Freeways in rural areas	110–130
Expressways in rural areas	80–110
Conventional roads	
Rural	
Flat terrain	90–120
Hilly terrain	80–100
Urban	
Arterial streets	60–100
Arterial streets with extensive development	50–70
Local residential streets	40–50
Access loop or crescent	40
Cul-de-sac	30

one curve with a design speed of 80 km/h could be described, completely accurately but totally misleadingly, as an 80 km/h road. Not surprisingly, the correlation between design speed and operating speed is low except in the presence of curve radii or *K*-values that are close to the limit for the selected design speed (Fitzpatrick et al., 2003).

As a measure of consistency, the values of all standards should, in theory, be as close as possible to those dictated by the design speed. Unfortunately, as suggested previously, this tacitly directs the design towards being at minimum values throughout. The general approach should be to design to as high a standard as possible without losing sight of the principle of consistency of design. Provided the principle of consistency is correctly applied, design speed will offer a first approximation of consistent design. Final selection of an appropriate design speed is based on a feedback loop aimed at eliminating operating speed inconsistencies. The initially assumed design speed should be logical with respect to the topography, anticipated operating speed, the adjacent land use and the functional classification of the road.

Typical design speeds applied to the various types of roads in urban and rural areas are listed in Table 5.2.

5.4.3 Operating speed

As is the case with so many of the geometric design standards where the performance of the drivers or their vehicles can assume a continuous range of values, the operating speed is defined as the 85th percentile of the distribution of observed speeds.

It is not economical to accommodate ALL speeds achieved on a road because drivers of high-performance vehicles, regardless of the possibility of prosecution, achieve very high speeds, sometimes in excess of 200 km/h. A curve showing the percentile distribution of speeds demonstrates a flattening off at the 85th percentile value. For this reason, this percentile value has been adopted internationally for design.

5.4.4 Running speed

Running speed is the quotient of the distance travelled and the travel time by a single vehicle. Average running speed can either be time mean speed or space mean speed as defined in Section 5.4.2.

5.4.5 Desired speed

In rural areas, the desired speed is the speed selected by the majority of drivers when not constrained by other vehicles. The difference between travel times at desired and actual running speeds constitutes delay, which features in economic assessments of the viability of a road alignment.

Desired speeds arise from a perception by the driver of the environment being traversed and can be modified by the concept of the 'self-explaining' road.

5.4.6 Posted speed

Posted speed limits are considered to be one of the more effective ways of enhancing road safety. They are basically intended to reduce the speed differential between vehicles because the frequency and severity of crashes are directly related to this speed differential. They are, however, generally unpopular for being set either too high or too low. To be effective, speed limits have to be actively enforced and there is a general belief that traffic calming may be a more suitable option. The only weakness of this argument is that it can apply only to local street systems and is definitely not to be considered in the case of freeways or arterials.

Posted speed limits are usually enshrined in legislation but with regional and local authorities generally having the right to vary these for specific situations. Figure 5.5 shows speed limits as applied by the various states of the United States.

Southern Africa applies a speed limit of 120 km/h to all rural roads with a blanket limit of 80 km/h for trucks and 60 km/h to urban streets. Because of a lack of effective policing, the truck speed limit is largely disregarded. Local authorities may vary the speed limits. Speed limits are often set to 70 or 80 km/h for higher order roads in the urban network and less than 60 km/h for very constrained situations.

Australian states and territories also use two default speed limits. These apply automatically in the absence of posted speed restriction signage. The two default speed limits are

- Within built-up areas, 50 kilometres per hour (31 mph), except for the Northern Territory, which remains at 60 kilometres per hour (37 mph)
- Outside built-up areas, 100 kilometres per hour (62 mph); two exceptions are Western Australia and the Northern Territory at 110 kilometres per hour (68 mph)

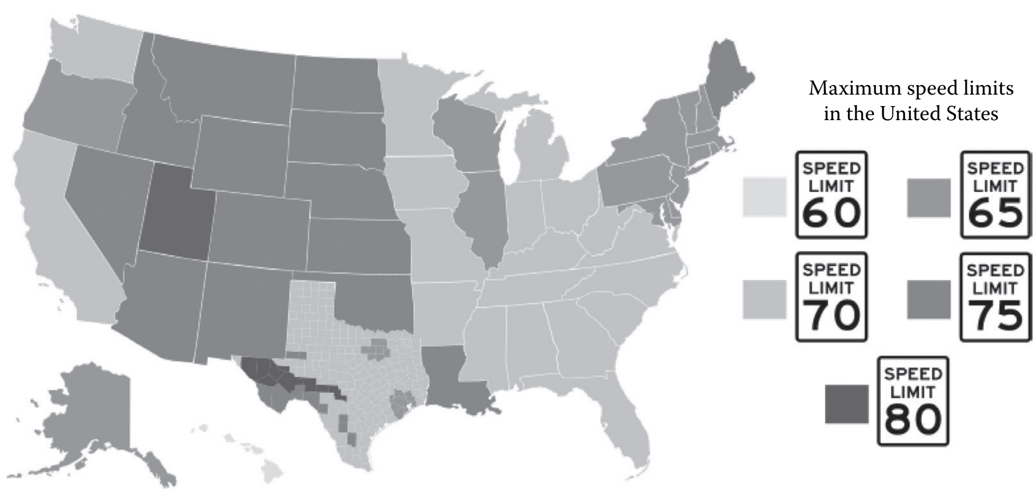


Figure 5.5 Speed limits in the United States.

Table 5.3 Speed limits in the United Kingdom

	Single carriageway		*Dual carriageway*		*Freeway*	
Vehicle type	*mph*	*km/h*	*mph*	*km/h*	*mph*	*km/h*
Passenger car	60	97	70	110	70	110
Recreation vehicle	50	80	60	97	60	97
Bus or coach	50	80	60	97	70	110
Truck below 7.5 tonnes	50	80	60	97	70	110
Truck above 7.5 tonnes	40	64	50	80	60	97

European countries have modest differences between the posted limits they impose but, in general, these are

- Urban areas 50 km/h
- Rural areas 90 or 100 km/h
- Freeways 120 or 130 km/h

Interestingly, Germany does not have any restriction on speed on its Autobahn system but, in the event of a crash at a speed higher than 130 km/h, the driver is automatically guilty, whether or not he or she was at fault. This serves as a very effective modifier of driver behaviour. On six-lane autobahnen, speeds are typically of the order of 140 km/h.

Speed limits in the United Kingdom vary not only by class of road but also by type of vehicle. These speed limits are shown in Table 5.3. Arbitrarily setting low speed limits on roads with obviously high geometric design standards will result in the credibility of the speed limit being reduced and consequently largely ignored.

There should be a fair level of agreement between posted speed limits and operating speeds. However, although it is accepted that operating speeds can exceed design speeds, blindly setting posted speeds to match operating speeds is not considered good practice. Design speed was previously defined as the maximum safe speed at which a road section could be traversed. A posted speed higher than design speed could be interpreted as an encouragement by the authority towards dangerous driving.

5.5 SIGHT DISTANCE

5.5.1 Introduction

Sight distance is the most fundamental of all the parameters of geometric design. Just about all modern passenger cars have the ability to go very fast and modern road design supports this ability. However, the ability to stop in sufficient time to avoid a crash is perhaps a more important feature of the driver/vehicle/road system. Without adequate sight distance, every journey would be fraught with danger or alternatively would be undertaken at a snail's pace.

The driving task is complex, as it has three components:

- Control – involving the tasks of steering, speed control and, in non-automatic cars, gear changing
- Guidance – which is the process of locating the vehicle relative to the road surface and other vehicles
- Navigation, involving trip planning and route following

These three activities are undertaken, for the most part, almost at a subconscious level but there are times when high levels of concentration are called for and stress levels are also high. For example, the process of entering a freeway requires the driver to

- Accelerate to the speed of vehicles on the freeway
- Locate a gap in the freeway traffic into which the merging manoeuvre can take place
- Maintain a gap between his or her vehicle and the leading vehicle on the ramp and then maintain the gap to the leading vehicle on the freeway, which may or may not have been the leader on the ramp

These activities are required to be executed virtually simultaneously. Another, and possibly more complex, task is overtaking another vehicle on a two-lane two-way road. In overtaking, drivers must judge

- The speed and acceleration potential of their vehicle taking account of the gradient of the road where the manoeuvre is to be executed
- The speed of the lead vehicle
- The speed and rate of closure of a vehicle travelling in the opposing direction
- The size of the gap in the traffic stream at the moment of initiation of the passing manoeuvre

Highway design has the greatest effect on the guidance task and an appreciation of this component of driving is needed if the designer is to be sympathetic to driver performance.

There are many activities required of the driver and it follows that there are many different forms of sight distance that the designer has to ensure are available as and when they are required. These include

- Manoeuvring sight distance
- Stopping sight distance (SSD)
- Headlight sight distance
- Passing sight distance (PSD)
- No-passing (or barrier) sight distance
- Intersection sight distance (ISD)
- Decision sight distance

These various forms of sight distance are discussed in the sections that follow.

5.5.2 The measurement of sight distance

Sight distance is measured as the distance between the driver's eyes and the object that has to be seen. Reference is made in the literature to a 'grazing' ray as a description of the line of sight. This line is from the observer's eye to grazing an obstacle to sight such as the crest of a hill and on to grazing the top of the object of interest. This creates the impression that the line of sight is to an object that is just out of sight. There has to be some or other additional height of object that has to be seen for the driver to be aware of the existence of the object and this is implicit in the discussion that follows. An example is that, while the height of a passenger car is about 1.30 metres, sight distance to a passenger car as the object of interest is measured to a height of 1.15 metres.

Typical heights of object and their application in the measurement of sight distance are shown in Table 5.4.

The object height of 0.4 metres in the table is applicable to manoeuvring sight distance and the Australian explanation, probably facetious and referring to Ezra Hauer's question

Table 5.4 Object height

Object height (m)	*Applicability*
0.0	Pavement markings
0.15	Small object in road
0.40	Large object in road
0.60	Vehicle tail or break light
1.15	Top of passenger car for Passing sight distance Intersection sight distance

Source: Burrell RC et al., *Geometric design guidelines*. South African National Roads Agency Limited, Pretoria, 2002.

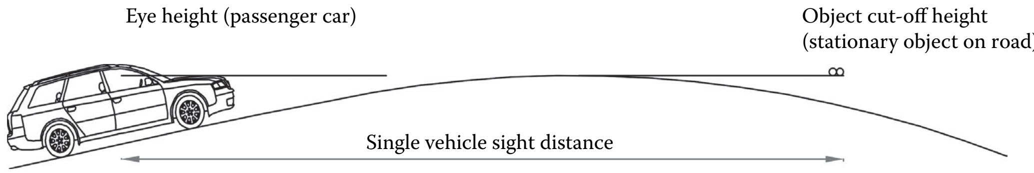


Figure 5.6 Sight distance on a crest curve. (From Queensland Department of Transport and Main Roads. *Road planning and design manual.* Brisbane, 2002.)

of 'How high is a dead dog?', is a dead kangaroo! The point is that many vehicles would be able to straddle an object that is 0.15 metres high, whereas all vehicles would have to manoeuvre around one that is 0.4 metre high.

The measurement of sight distance is illustrated in Figure 5.6. Driver eye height is discussed in Section 5.2. When reference is made to the various applications of sight distance, the driver's reaction time also has to be considered. As stated previously, for design purposes the 85th percentile reaction time of 2.5 seconds is often used although a range of reaction times is employed in the case of decision sight distance.

5.5.3 Manoeuvring sight distance

It isn't always necessary or even desirable that a vehicle should have to stop for an obstacle in its path. The preferred option is that the driver should be able to steer around the object either at unreduced speed or with only a relatively modest reduction in speed. This is because speed differentials in a traffic stream are potentially dangerous. A vehicle stopping for an object in its path is at risk of being hit by a following vehicle if that driver is at all inattentive. The availability of manoeuvring sight distance ensures that speed differentials remain low. Although manoeuvring sight distance is not a recommended design parameter, it is a useful measure of geometry in extremely limited circumstances.

Manoeuvring requires that there be space into which a vehicle can be steered to avoid the offending object. A surfaced shoulder that is 2.0 metres wide would suffice for this purpose.

Manoeuvring sight distance is the sum of distance travelled during the driver's reaction time and the distance required to manoeuvre around the object, expressed as

$$\mathrm{MSD} = \frac{vt}{3.6} + s$$

Table 5.5 Manoeuvre sight distance

Design speed (km/h)	*Manoeuvre sight distance (m)*
60	60
70	75
80	95
90	120
100	155
>100	Do not use

Table 5.6 Decision sight distance

Design speed (km/h)	*Avoidance manoeuvre*				
	A	*B*	*C*	*D*	*E*
50	70	155	145	170	195
60	95	195	170	205	235
70	115	325	200	235	275
80	140	280	230	270	315
90	170	325	270	315	360
100	200	370	315	355	100
110	235	420	330	380	430
120	265	470	360	415	470
130	305	525	390	450	510

Source: Fambro DB et al., *Determination of stopping sight distances*. NCHRP Report 400, Transportation Research Board, Washington, DC, 1997.

Note: Manoeuvre A: Stop on rural road: t = 3.0 seconds; Manoeuvre B: Stop on urban road: t = 9.1 seconds; Manoeuvre C: Speed/direction change on rural road: t varies between 10.2 and 11.2 seconds; Manoeuvre D: Speed/direction change on suburban road: t varies between 12.2 and 12.9 seconds; Manoeuvre E: Speed/direction change on urban road: t varies between 14.0 and 14.5 seconds.

where

MSD = manoeuvre sight distance
v = initial speed of vehicle (km/h)
t = reaction time (seconds)
s = evasive action distance

The initial speed of the vehicles is assumed to be the design speed of the road, and the driver's reaction time, as stated previously, is taken as 2.5 seconds. The distance, s, has to be assumed. Australian experience suggests that the evasive action distance should have the values shown in Table 5.5. The distance travelled during reaction time must be added to s to establish the total manoeuvre sight distance.

It should be noted that manoeuvre sight distance has values similar to those for decision sight distance shown in Table 5.6 where the required manoeuvre is a stop on a rural road. The object height is 0.40 metre and the driver's eye height is 1.05 metres above the road surface.

5.5.4 Stopping sight distance

The ability to stop is the most basic of all the possible safety measures that can be brought to bear on the safety of a moving vehicle. It follows that stopping sight distance has to be provided at all points along the road.

Stopping sight distance has two components. These are the distances traversed during

- The driver's reaction time, which is the time that elapses between the time that the circumstance requiring the stop is first manifested and the application of the brakes
- Braking time, which is the time taken from the first application of the brakes to the vehicle having been bought to a stop

Stopping sight distance is thus calculated as

$$\mathrm{SSD} = v \quad t + \frac{v}{2a}$$

where
SSD = stopping sight distance (m)
v = design speed (m/s)
t = time (seconds)
a = deceleration rate (m/s^2)

Speed is, however, more conveniently expressed in km/h. The above relationship is thus recast as

$$\mathrm{SSD} = v \quad 0.278t + 0.039\frac{v}{a}$$

with the variables being as before.

The rate of deceleration for comfortable stopping is widely accepted as being 3 m/s^2. American research has shown that some drivers regularly accept 6 m/s^2 as an acceptable rate of deceleration. This is higher than the rate possible on a wet pavement, namely about 4 m/s^2, and approaching the maximum available on a dry surface with good tyres, namely about 8 m/s^2. In Table 5.7, desirable stopping sight distance is based on 3 m/s^2 and absolute minimum stopping sight distance, corresponding to a panic stop, on 6 m/s^2.

Table 5.7 Stopping sight distance

	Stopping sight distance (m)			
	Desirable		*Absolute minimum*	
Design speed (km/h)	*Calculated*	*Design*	*Calculated*	*Design*
50	67.2	65	51.0	50
60	88.5	90	65.1	65
70	112.3	110	80.5	80
80	138.8	140	97.2	100
90	167.8	170	115.2	115
100	199.5	200	134.5	135
110	233.7	230	155.1	155
120	270.6	270	177.0	175
130	310.1	310	200.2	200

5.5.5 The effect of horizontal curvature on stopping sight distance

Horizontal curvature can reduce sight distance because of an object such as a cut face or retaining wall or vegetation obscuring the line of sight. Required stopping sight distance may, however, increase because some of the available friction between the vehicle's tyres and the road is required to maintain the vehicle in a circular path. Furthermore, drivers are reluctant to brake sharply on curves. The reason could be that, because of the centrifugal force that they are already experiencing by traversing the curved path, they are reluctant to add to the forces working on their bodies. This is a facet of human factors that has yet to be researched.

If it is not possible to remove the obscuring object, a radius of curvature higher than the minimum would have to be accepted to ensure that stopping sight distance is achieved.

The obscuring object could also be a safety barrier on the inside of the curve. Because these barriers are normally located on the shoulder and close to the shoulder breakpoint, the option of increasing the offset is not available. It would be necessary to select a higher radius of curvature to address this circumstance also. With a fixed object, the higher radius of curvature required to provide the desired stopping sight distance will result in one or both of the tangents to the curve having to be relocated to move the curve farther away from the obstruction.

The method of measurement of the effect of horizontal curvature on sight distance is illustrated in Figure 5.7.

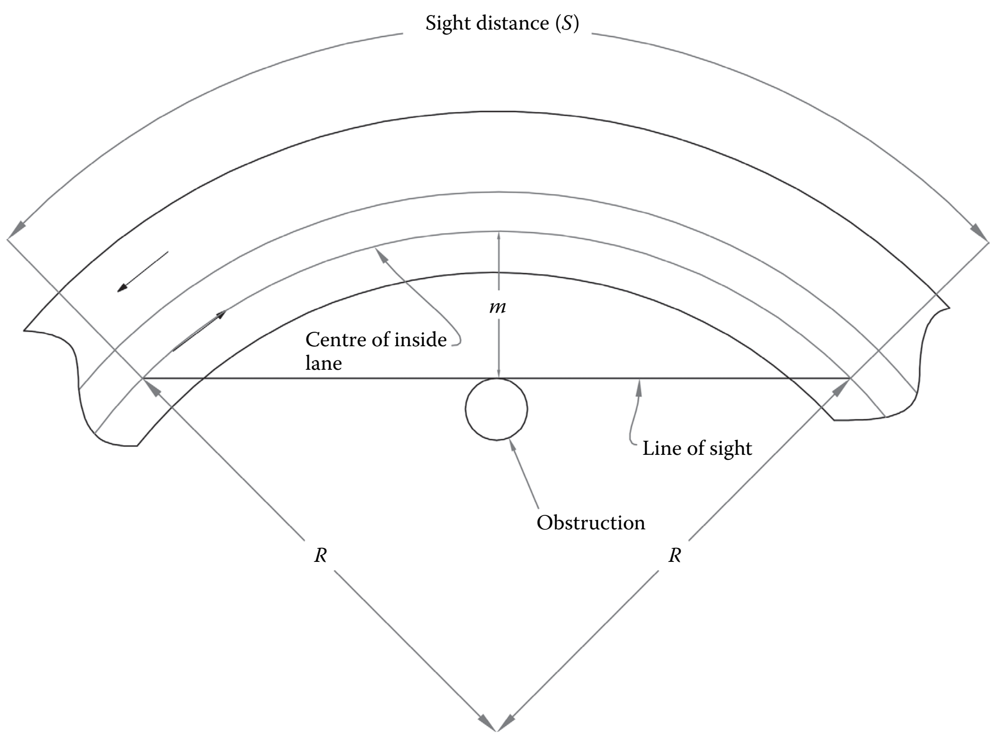


Figure 5.7 The measurement of stopping sight distance on a curve. (From Iowa Department of Transport. *Geometric design manual*. Des Moines, 2004.)

Many manuals quote the relationship between the radius of curvature and the offset to an obscuring object as being (AASHTO, 2004)

$$m = R \;\; 1-\cos\frac{28.65S}{R}$$

where
m = offset (m)
R = radius (m)
S = stopping sight distance (m)

This creates the impression that although the radius of the curve is fixed, there is a choice available for the value of the offset. Unfortunately this is not always the case. The pleas of environmentalists notwithstanding, a tree can always be chopped down even if it is 300 years old. A high rise building as the offending object may be more difficult to remove and the designer will have no choice but to increase the radius of the curvature to a value that will provide the required stopping sight distance. Using the relationship offered above to establish the required radius will require a trial-and-error approach or a hill-climbing algorithm. It is more convenient to recast the relationship as (New Jersey Department of Transportation, 2011)

$$R = \frac{m}{2} + \frac{\frac{s}{2}^{2}}{2m}$$

with the symbols having the same meaning as previously.

The radius for a range of offsets from 4 metres to 12 metres can be read off the curves shown in Figure 5.8.

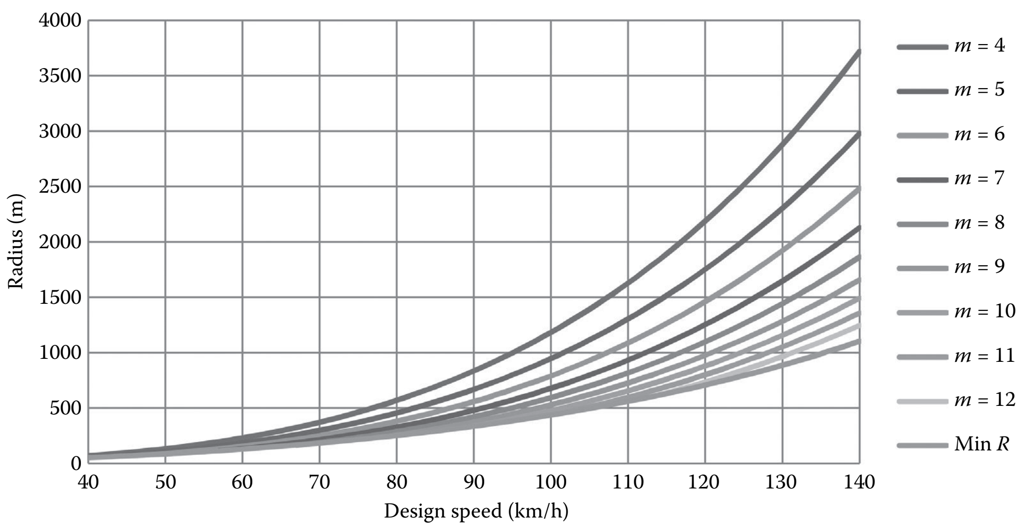


Figure 5.8 Radii for stopping sight distance on horizontal curves. (From Burrell RC et al., *Geometric design guidelines*. South African National Roads Agency Limited, Pretoria, 2002.)

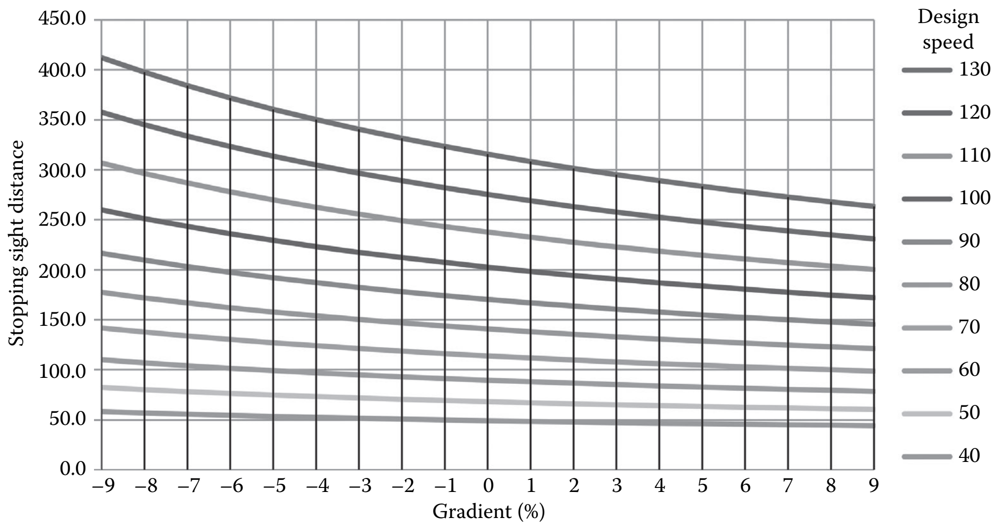


Figure 5.9 Stopping sight distance on gradients.

5.5.6 The effect of gradient on stopping sight distance

Vehicles can stop quicker on upgrades and require longer stopping distances on downgrades. To some extent this variation is nullified by drivers' preferences for a rate of deceleration of about 3 m/s^2, regardless of whether they are travelling up-hill or down. However, should drivers' preferences be for a level of braking effort rather than for a rate of deceleration, the correction for gradient would have to be taken into account. Similarly, it must be taken into account in relation to absolute minimum stopping sight distances with a deceleration rate of 0.6 m/s^2 plus or minus the allowance for the effect of gradient.

The portion of the relationship for stopping sight distance relating to braking distance has to be modified to

$$d = \frac{v^2}{254 \quad \frac{a}{9.81} \pm G}$$

where
 d = braking distance (m)
 G = gradient (m/m) (alternatively percentage gradient divided by 100)

The stopping sight distance modified to take cognisance of gradient is shown in Figure 5.9.

5.5.7 The effect of a combination of horizontal curvature and gradient on stopping sight distance

In rugged terrain, such as on mountain passes, short radius curves compounded with steep gradients constitute a worst case scenario for stopping sight distance. The presence of the gradient could cause the selection of a radius based purely on horizontal forces as shown in Table 5.7 to be an under design. To accommodate this situation, the stopping sight distance as modified by the gradient should be read off Figure 5.9. The revised stopping sight distance should then be applied to one or other of the two relationships offered for the value either for the value of m, the distance to an obscuring object from the centre of the inside lane, or the value of R given a fixed value for m as shown in the relationship preceding Figure 5.8.

5.5.8 Headlight sight distance

Sag curves do not present any problems of sight distance during the hours of daylight or on well-lit streets at night. Minimum curvature on sag curves is thus much lower than that on crest curves and the main criterion of sag curves is passenger comfort. Headlight sight distance therefore applies to driving at night on an unlit road with particular reference to sag curves. The values attached to headlight sight distance are, in fact, those of stopping sight distance. The driver's eye height of 1.05 metres is replaced by the height of the headlight as the point of departure for the line of sight. The headlight is assumed to be mounted on the vehicle at a point 0.60 metre above the road surface and the cone of light is taken to have a divergence of 1° from the central axis of the vehicle.

The object height could be any of 0.0 metre, 0.15 metre or 0.4 metre as the taillights of a leading vehicle would, in all probability, be lit, hence reducing the need for consideration of an object height of 0.6 metre. General practice is that sight distance is measured from the vehicle to the point where the upward side of the light beam intersects the road surface, that is, an object height of 0.0 metre. The object height of 0.4 metre, in addition to being that of a large object on the road surface, is also the height of the bottom of the headlight beam at a distance equal to stopping sight distance.

Headlight sight distance is illustrated in Figures 5.10 and 5.11.

Headlight sight distance is used in the determination of the K-value of sag curves, with this parameter of vertical curvature being discussed in Chapter 7. With bright headlights, it is possible to see about 100 metres ahead and with dipped beams about 50 metres. These distances correspond to design speeds of about 70 km/h and 40 km/h respectively. It follows that drivers tend to overdrive the available sight distance by a considerable margin.

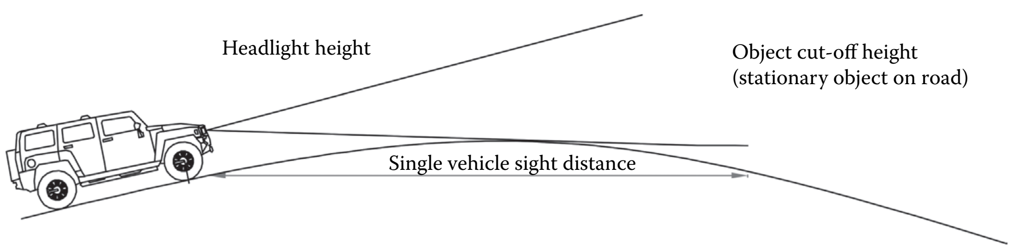


Figure 5.10 Headlight sight distance on crest curve. (From Queensland Department of Transport and Main Roads. *Road planning and design manual*. Brisbane, 2002.)

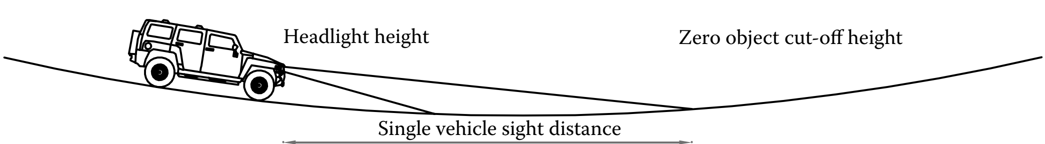


Figure 5.11 Headlight sight distance on sag curve. (From Queensland Department of Transport and Main Roads. *Road planning and design manual*. Brisbane, 2002.)

5.5.9 Passing sight distance

Most rural roads and many urban streets have a two-lane two-way cross-section and vehicles overtake slower vehicles by driving temporarily on the lane allocated to the opposing stream of traffic. If this manoeuvre is to be completed safely, the driver has to be able to see a sufficient distance ahead to ensure that the manoeuvre can be completed before meeting an opposing vehicle.

The passing manoeuvre is the most demanding of all driving operations. It requires the driver to assess the ability of his or her vehicle to accelerate on the gradient where the passing manoeuvre is to be executed, the speed of the vehicle to be passed and the speed and location of an opposing vehicle at the start of the manoeuvre. The gaps between the vehicle and the passed vehicle preceding and following the execution of the manoeuvre also have to be controlled. In addition to being a demanding task, passing is a potentially dangerous manoeuvre because the closing speed between the passing and the opposing vehicles could be 240 km/h or approximately 70 m/s so that there is little room for error.

The passing manoeuvre can be broken down into four discrete steps as shown in Figure 5.12. These are

- d_1 – distance traversed during perception and reaction time and during the initial acceleration to the point of encroachment on the left lane
- d_2 – distance travelled while the passing vehicle occupies the left lane
- d_3 – distance between the passing vehicle at the end of its manoeuvre and the opposing vehicle
- d_4 – distance traversed by an opposing vehicle for two-thirds of the time the passing vehicle occupies the left lane, or two-thirds of d_2 above

Two possibilities can arise: that the passing manoeuvre can either be completed or, having been initiated, is aborted. The point of no return is where the time required to abort the

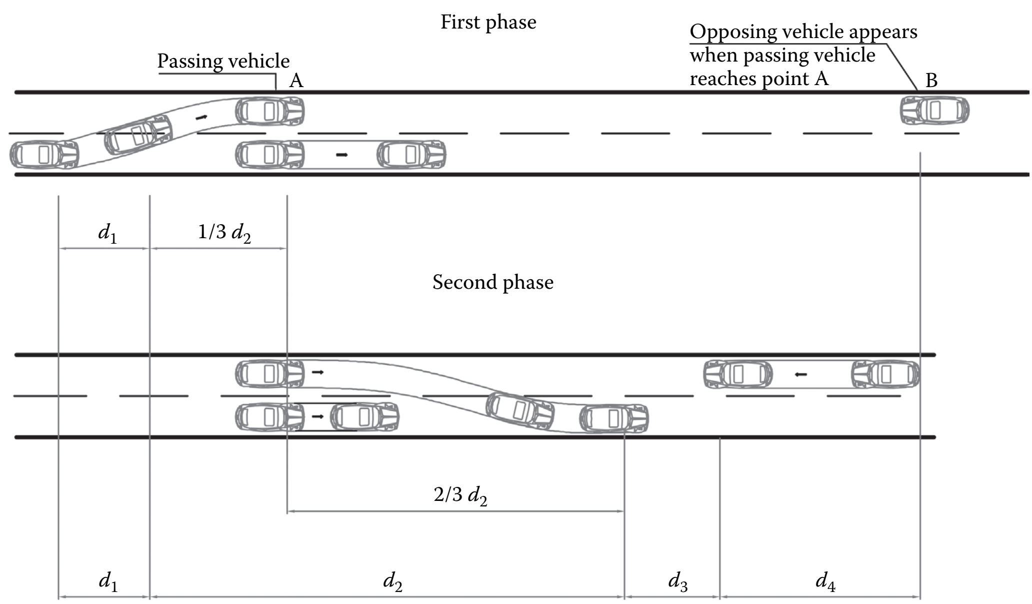


Figure 5.12 The passing manoeuvre. (From Harwood DW et al., *Passing sight distance criteria*. NCHRP Report 605, Transportation Research Board, Washington, DC, 2007.)

manoeuvre is equal to the time taken to complete the manoeuvre. Once past this point, the driver is committed to completing the passing manoeuvre.

Because of the complexity of the task, numerous variables are involved. The assumptions regarding the magnitude of these variables (Harwood et al., 2007) are the following:

- The passing and opposing vehicles are both travelling at the design speed of the road.
- The passed vehicle maintains a constant speed.
- The speed differential between the passing and the passed vehicles is 19 km/h (12 mph).
- The passing vehicle has sufficient acceleration capability to reach the specified speed differential relative to the passed vehicle by the time it reaches the critical position, which generally occurs about 40 per cent of the way through the passing manoeuvre.
- The lengths of the passing and passed vehicles is that determined for the American design passenger car being 5.8 metres.
- Aborting the passing manoeuvre will be based on a reaction time of 1 second.
- In the case of aborting the manoeuvre, a deceleration rate of 3.4 m/s^2 (which is the rate adopted by the United States for the calculation of stopping sight distance) will be applied.
- For either an aborted or a completed passing manoeuvre, the headway between the passing and the passed vehicles will be 1 second.
- The minimum gap between the passing and the opposing vehicles at the end of the passing manoeuvre will be the spatial equivalent of 1 second of travel time at the design speed with the end of the manoeuvre being signified by the passing vehicle being completely inside its own lane.

The number of assumptions that have to be made testify to the complexity of the manoeuvre. Values of the aforementioned variables were derived from comprehensive field tests. Passing sight distances calculated on the basis of these assumptions for the range of design speeds from 30 km/h to 130 km/h are shown in Table 5.8.

Passing sight distance is based on a driver eye height of 1.05 metres. The assumed object height is 1.15 metres on the basis of a passenger car having an 85th percentile height of 1.30 metres.

Table 5.8 Passing sight distance

	Assumed speeds (km/h)		
Design speed (km/h)	*Passed vehicle*	*Passing vehicle*	*Passing sight distance (m)*
30	11	30	120
40	21	40	140
50	31	50	160
60	41	60	180
70	51	70	210
80	61	80	245
90	71	90	280
100	81	110	320
110	91	110	355
120	101	120	395
130	110	130	440

Not surprisingly, passing sight distances quoted by various authorities display a considerable range of values. Figure 5.13 illustrates the range of passing sight distances recommended by

- The American Association of State Highway Officials
- The United Kingdom Highway Agency
- The Southern African Development Community
- The Australian body Austroads Limited

The no-passing marking distance recommended in the Federal Highway Administration's *Manual on Uniform Traffic Control Devices* is, by way of comparison, also shown in this figure (Federal Highway Administration, 2009).

The passing sight distances illustrated in Figure 5.13 generally tend to run parallel to each other with the Australian passing sight distance being based on a different model (Austroads, 2009a). It points to a variable not addressed by the other national standards, specifically the judgment of the passing driver and the risks this driver is prepared to take. Various vehicles are offered as the overtaken vehicle including

- A semitrailer
- A 'B-double' combination vehicle
- Two types of road train

The 'B-double' is referred as a B-train in many countries and as an interlink in southern Africa. It comprises two trailers linked together by a fifth wheel coupling, which is typically located above the rear wheels of the lead trailer. This area of the lead trailer is often referred to in North America as the 'bridge'. The combination vehicle can be up to 26 metres long and passing it requires significantly more effort than passing a passenger car. It is the vehicle

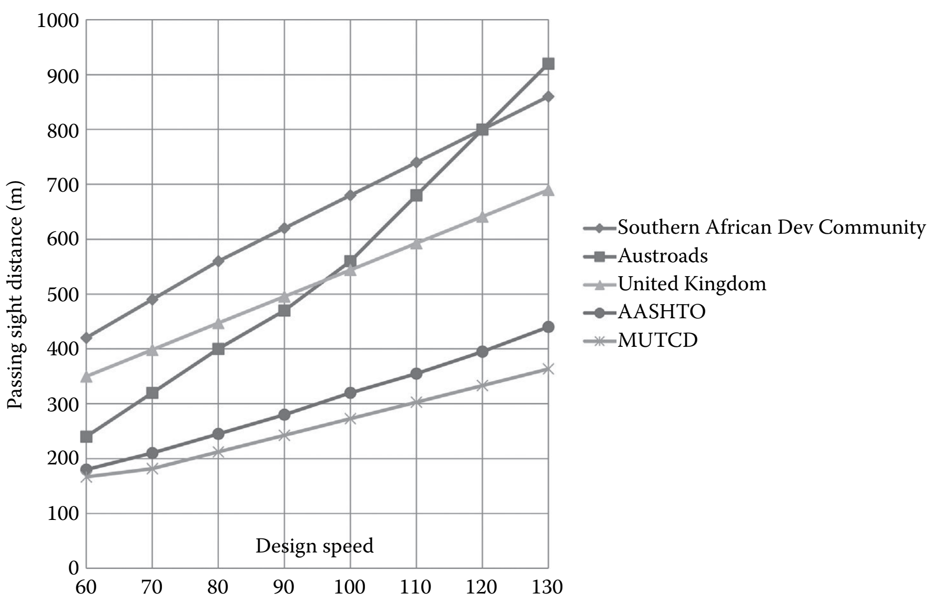


Figure 5.13 International range of passing sight distances.

Figure 5.14 The B-train. (From Austroads. *Guide to road design, Part 3: Geometric design*. Sydney, 2009a.)

selected in Figure 5.11 as the passed vehicle, as the road trains are not typically found in other countries. This vehicle is illustrated in Figure 5.14.

The Australian approach to passing sight distance is a start towards a human factors approach. It refers to an 'establishment sight distance' and a 'continuation sight distance'. These have the significance of

Establishment: A minimum sight distance that is adequate to encourage a given proportion of drivers to commence an overtaking manoeuvre. It is called the overtaking establishment sight distance (OED) because it establishes a length of road as a potential overtaking zone. It is the distance required for most drivers of passenger cars to overtake other vehicles.

Continuation: A sight distance that, if maintained for some length of road after the OED has become available, will enable a driver either to complete or to abandon a manoeuvre already commenced with safety. This is called the overtaking continuation sight distance (OCD). After the OED first becomes available an overtaking zone is assumed to extend as long as the shorter distance remains available (Austroads, 2009b).

As discussed in Chapter 10, intersection sight distance was previously based on an elaborate model with many variables and this has now been replaced by the human factors approach of gap acceptance. It is possible that, in times to come, the currently very elaborate model of passing sight distance will also be replaced by another version of gap acceptance. This is research that is waiting to be done.

5.5.10 No-passing (or barrier) sight distance

No-passing sight distance refers to a road marking indicating that a passing manoeuvre would not be desirable when sight distance is restricted to this or a lesser distance. In some countries, a passing manoeuvre may not be initiated where it occurs and, in others, the manoeuvre must be completed before the marking has been reached. Apart from the safety implications, passing in a no-passing zone is also an offence.

No-passing sight distance is measured from a driver eye height of 1.05 metres to an object height of 1.15 metres.

Australian practice refers to an intermediate sight distance that is equivalent to no-passing sight distance elsewhere and is similar to the no-passing sight distance applied in southern Africa, although, in the Australian case, it is applied only to narrow, essentially one-lane, roads. New Zealand has a similar sight distance but describes it as being the minimum acceptable passing sight distance. This distance is twice stopping sight distance. In effect, two vehicles on opposing trajectories in a common lane should be able to stop without hitting each other. American practice is defined in the Federal Highway Administration's

Table 5.9 No-passing (barrier) sight distance

Design speed (km/h)	*2 × SSD (m)*	*MUTCD*
50	100	150
60	130	170
70	160	200
80	200	240
90	230	270
100	270	300
110	310	330
120	350	360
130	400	370

Note: MUTCD, Federal Highway Administration's *Manual on uniform traffic control devices*; SSD, sight stopping distance.

Manual on Uniform Traffic Control Devices (MUTCD) as being the sight distance available between two points that are 3.5 feet (1.06 metres) above the road surface.

British practice is to define full overtaking sight distance (FOSD). No-passing sight distance as demarcated by road markings is everywhere else. No-passing road markings would commence on a curve to the right and on a crest curve where the sight distance is half of FOSD. On a curve to the left, the markings would commence at a point that is FOSD/4 metres in advance of the tangent point (or centre of transition curve). In all cases, the termination of the no-passing road marking is at the point where FOSD is available once more.

No-passing sight distances as defined by MUTCD and by the South African Department of Transport and Austroads as twice stopping sight distance are listed in Table 5.9.

5.5.11 Decision sight distance

Stopping sight distance is sufficient to allow a reasonably alert driver to bring the vehicle to a stop in an unhurried manner, that is, at a deceleration rate of 3 m/s^2. This presumes that the driver can perceive the need to stop in a timeous manner. However, when the environment is filled with visual clutter, or when decisions have to be made in quick succession, such as at a busy intersection, reaction time effectively slows down because human beings are single-stream data processors and more information has to be processed to arrive at a decision. Designers must take this into account.

Decision sight distance is defined as the distance needed for a driver to detect an unexpected or otherwise difficult-to-perceive information source. It also allows time for the driver to complete evasive manoeuvres at unaltered or reduced speed instead of stopping. Decision sight distance is thus substantially longer than stopping sight distance.

The object triggering the need for a manoeuvre could be a road marking indicating a need to change lanes such as from a through lane at an intersection to one or other of the turning lanes. For this reason, decision sight distance is measured to the road surface, an object height of zero metres.

Because of the additional time made available to the driver, the availability of decision sight distance at critical points along the road should be checked. Decision points should be located such that a sufficient time elapses between successive decisions or, if this is not possible, additional traffic control measures should be brought to bear on the driving task to provide the driver with advance warning of the situation to be encountered.

The required manoeuvre could be to bring the vehicle to a stop in either a rural or an urban environment or to change direction or speed in a rural or a suburban or an urban environment

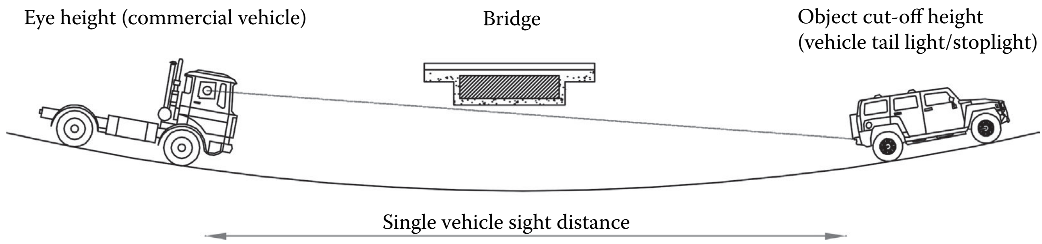


Figure 5.15 Sight line under a structure. (From Queensland Department of Transport and Main Roads. *Road planning and design manual*. Brisbane, 2002.)

as shown in Table 5.6. It will be noted that reaction time increase as traffic densities increase from rural to urban hence making these manoeuvres more complex to execute.

The similarity between the American decision sight distance and the Australian manoeuvring sight distance has already been mentioned. In Table 5.5 the suggestion is made that manoeuvring sight distance should not be applied to design speeds greater than 100 km/h. This suggestion has distinct merit, as a high-speed swerve around an object on the road could be distinctly dangerous. Although values are offered for design speeds up to 130 km/h for manoeuvres C, D and E, it is possible that the generous reaction times would allow for perhaps a quicker actual reaction time and a longer breaking period so that the final manoeuvre around the offending object could happen at a speed significantly lower than the design speed.

5.5.12 The effect of sag curves under structures

A bridge deck or a similar structure crossing the road could have the effect of obscuring the sight distance on a sag curve, as shown in Figure 5.15. In other cases, the additional eye height afforded the truck driver is a benefit but, in the case of sight distance under a structure, it constitutes a worst case. The driver eye height is thus taken as being 2.0 metres and the object height is usually selected as 0.60 metre, that is, the height of the tail light of the leading vehicle. The sight distance under the structure should at least be stopping sight distance or greater.

5.6 TRAFFIC

5.6.1 Introduction

Geometric design requires that attention be paid both to microscopic and macroscopic issues. Microscopic issues relate to the behaviour of individual drivers and the limitations of the performance of their vehicles. Macroscopic issues refer to the behaviour of the traffic stream as a whole. If one borrows an analogy from hydraulics, microscopic issues relate to the individual particles in the stream whereas macroscopic issues refer to the total flow in the pipe. The pipe diameter defines the capacity of the system, with the quantity of the flow in the pipe being determined by the flow depth.

Traffic characteristics address the macroscopic issues of interest. These include

- Traffic flow
- The composition of the traffic stream
- The directional split between opposing flows

They are discussed further in the following sections.

5.6.2 Flow, present and anticipated

The volume of traffic dictates the cross-section of the facility to be provided. It is the major determinant of whether a road should have a two-way two-lane cross-section, or be a freeway with two or more lanes in each direction. On existing roads that are in need of upgrading, present-day volumes can be assessed by means of traffic counts to address current restraints on capacity. There is, however, no point in upgrading a road merely to serve existing traffic volumes. This would result in the road already running at capacity levels as soon as it is opened to traffic with no provision for future growth in traffic volumes.

It is necessary to develop a prediction of traffic volumes that are likely to occur sometime in the future. Reference is thus made to the design year and the design hour.

The design year is often chosen as being 20 years after the opening of the road to traffic. In the case of a very expensive project, a longer period may have to be selected in terms of benefit/cost ratio analyses. On the other hand, if the project involves only the provision of a climbing lane on an existing road on which the design life has already partially expired, it may be prudent to accept a design life equal to the remaining lifespan of the road. The selection of the prediction period thus requires some thought.

5.6.3 Composition

The majority of vehicles on the road are passenger cars. The heavy vehicles present in the traffic stream are quoted as being a percentage of the total flow and typically range from 5 to 15 per cent of the total. Heavy vehicles are single-unit trucks and buses, semitrailers, referred to in some countries as articulated vehicles, and combination vehicles, referred to variously as WB22, B-doubles or interlinks. These vehicles constitute a major impediment to the movement of the lighter and faster passenger cars as passing them is the equivalent of a manoeuvre in which a platoon of anything up to three cars are passed. Their power/weight ratio is such that they lose speed on upgrades steeper than 3 per cent and, as a safety measure, downgrades are usually essayed at speeds similar those achievable on the equivalent upgrades. Safety posters put this rather succinctly as ‘Go down at the speed at which you can go up’.

Traffic streams comprising the entire spectrum of different vehicle types are inconvenient to analyse. For this reason, the *Highway Capacity Manual* adopts the initial step of converting all the vehicles in the stream to passenger car equivalents (PCE) or passenger car units (PCU) and subsequent analyses are conducted on the basis of this homogeneous traffic stream. The conversion to the homogeneous stream is based on a stratified traffic count separating the passenger cars from the heavy vehicles, then multiplying the number of heavy vehicles by the appropriate PCEs before adding them back into the traffic stream.

It must be noted that the number of PCEs per truck vary according circumstances. For example, on upgrades the PCEs for trucks can vary from 1.5 to over 15.0 depending on the gradient and length of the grade as well as on the directional flow rate.

5.6.4 Directional split

In rural areas, the flows in the opposing directions tend to be similar in magnitude, with the difference between them seldom being more than a 60:40 split.

In urban areas, the split is dramatic with traffic inbound towards the central business district (CBD) in the morning peak periodically being brought to a standstill because of congestion while the outbound traffic lanes are virtually deserted. In short, level of service (LOS) E in one direction is matched by LOS A in the reverse direction. In the evening peak, the heavy flow is outbound and similar in magnitude to the morning peak flow. For this reason, the directional split results in the flow sometimes being referred to as tidal. The directional split has been observed to be as high as 90:10, although this is rare.

The design must be split into two processes both of which cater for the heavier of the two flows, one for the morning peak and the other for the evening peak. In the case of the 90:10 split referred to, the design must thus be able to accommodate 90 per cent of the total hourly flow in each direction.

Chapter 6

Horizontal alignment

CONTENTS

6.1 INTRODUCTION

The horizontal alignment comprises a series of horizontal curves and the tangents between them. It is the most long-lived of the various elements of the road, with the alignment of some roads built by the Roman Empire still in use today. It is also the first element to be considered in the design sequence of horizontal alignment, vertical alignment and cross-section.

The design sequence, however, is not to be construed as being a linear process. It is, in fact characterised by ongoing feedback and refinement because each element of the design can and does impact on the others. The vertical alignment could generate a need to realign the horizontal alignment to avoid undesirably steep gradients. The location of an intersection may be so critical that everything else has to be realigned to accommodate it, although, customarily, the horizontal and vertical alignments would jointly dictate where an intersection could be located. Generously sized cross-sectional elements would create the expectation that high speeds could safely be maintained. The standards selected for the horizontal and vertical alignments would have to match this expectation. In short, the various elements and dimensions of the design have to be in harmony with each other.

Reference was made in Chapter 4 to consistency of design. This relates principally to the horizontal alignment, specifically to the sequence of horizontal curve radii and the magnitude of successive curves as well as to the spacing between them.

There are two ways of approaching design of the horizontal alignment. These are fully discussed in Chapter 9. The easier of the two methods is to locate the tangents follows by selection of the radii of the curves linking them. The other is to select the curves first and follow this by incorporating the tangents. The first-mentioned is likely to result in long tangents connected by curves that are relatively short compared to the lengths of the tangents. The second method tends to lead to long curves with high radii and short linking tangents and generally results in a visually more pleasing alignment. In flat terrain, the alignment would comprise very long tangents so that a curvilinear approach would be inappropriate.

6.2 GENERAL PRINCIPLES FOR HORIZONTAL ALIGNMENT

6.2.1 Introduction

The factors having an impact on the design of the horizontal alignment include

- Physical constraints such as the general shape of the topography, including the presence of watercourses, land use, and man-made features. Geophysical conditions such as collapsing sands, expansive clays and so on should also be considered.
- The effect the road may have on the environment such as its effect on ecologically sensitive areas, adjacent land use and community impacts.
- Cost of land acquisition, construction and maintenance.
- Road user costs.
- Safety on the basis of human factor considerations, context sensitive design and consistency of alignment.
- Highway classification and design policies.

In addition to these fundamental inputs into selection of the horizontal alignment, sight must not be lost of the fact that geometric design is an exercise in three dimensions. The horizontal alignment, the profile and the cross-section have to be properly harmonised if the prime objective of safe, economical and convenient transportation of people and goods is to be realised.

General principles are not calculated standards. They should rather be considered as constituting best practice. As such, they are not theoretically derived but have to be borne in mind to ensure smoothly flowing traffic streams.

6.2.2 Planning

The shortest distance between two points is a straight line. Ideally, the horizontal alignment should thus be as directional as possible between its ends. The design of the horizontal alignment starts with planning, specifically route location.

As a rule of thumb approach to route location, as based on location of the tangents followed by selection of the radii of the curves linking them, the two ends of the proposed route should be joined by a straight line. This provisional location should then be studied to identify features requiring a departure from it. Starting at one of the ends, called for convenience the origin, it is necessary to establish the *minimum* deviation to the left and to the right of the original route required to avoid

- A topographic restraint
- Alienation of portions of an existing property
- Damage to some or other area of value to the local community

The origin should then be joined to these two points, which should then be joined by further straight lines to the far end of the original line, the destination. The original line is then abandoned.

The process is then repeated for the two new straight lines and two new points requiring deviation from each of the two new lines identified. These points are then linked to the destination so that there are now four possible routes between origin and destination. The process is repeated until there is no further need for new points of deviation on any of the routes now identified for consideration. The benefit of this process is that it eliminates the temptation of only one route being identified between origin and destination. Choices can be made only between alternatives and an absence of alternatives may result in a less than optimum route location being accepted.

The process is also to be repeated from what was originally the destination and proceeding back to the original origin starting with the straight line linking them. With the completion of this process, there may be 20 or more possible routes between the two points. Engineering judgment would be sufficient to eliminate many of the alternatives. With experience, many would be rejected even before completing their further progress towards the destination. The remaining routes still in contention should then be subjected to a full systems analysis process combining economic and utility analyses as discussed in Chapter 16.

Preserving the direction of the alignment suggests that the deflection angle between two successive tangents should be as small as possible. As stated previously, what is required is the *minimum* deviation. Some fine-tuning of the selection of the location of the point of intersection (PI) of the two tangents is thus called for. In the process of pursuing the minimum deflection angle, it is possible that the PI may become too close to the obstacle being avoided in the sense of the limitation that this places on the selection of the radius of horizontal curvature. A large radius may cause the road to pass through the obstacle that is to be avoided, thus defeating the objective being sought. This problem would not arise if it is the curve that is being located in the first instance.

In exactly the same way that a continuous feedback process is required in the design of the various elements of the design, feedback is also required between the location of successive tangents and selection of the radius of the curves connecting them. Minimum values of radius of curvature should be avoided as far as possible and be used only in the most limiting of circumstances.

6.2.3 Safety

Sharp curves should not be located at the end of long tangents (AASHTO, 2011a). Speeds tend to increase gradually on long tangents and drivers would not necessarily be aware of their speed in approaching a sharp curve. This could result in a run-off-the-road (ROR) crash. If the topography forces the selection of short radius curvature, these curves should be preceded by a series of successively sharper curves. A design speed of 120 km/h could be followed by a series of design speed reduction steps of 10 km/h each until the desired design speed is achieved. If space permits, each section of reduced speed should include at least two curves at the minimum radius for that speed. This is to reinforce the realisation that the previous minimum curve was not just an isolated situation.

It is difficult for drivers to assess the sharpness of a curve in the absence of features such as cut faces, trees and shrubbery or buildings projecting above the road surface. It is thus not desirable to locate sharp curves on high fills.

Compound curves can make it easy for the designer to fit the road within the prevailing topographic constraints. Ease of design must, however, be sacrificed in the service of road safety. In terms of human factors design, drivers expect that the curve they are negotiating will maintain a constant radius. A reduction in radius would probably require the driver to brake. Braking on curves is generally applied in a gingerly fashion and the speed reduction achieved may be a case of too little, too late.

The only place where compound curves are acceptable to drivers is on loop ramps at interchanges. The limitations on the length and radii of successive curves are discussed in Chapter 11.

From a safety point of view, the goal should be to use the highest possible value of radius. There is, however, a caveat to this insofar as two-way two-lane roads require tangents long enough to allow for safe overtaking manoeuvres. Generally, the longer the curves, the shorter are the tangent between them. A compromise should therefore be sought between passing sight distance and curve radius.

On a two-lane road, curves to the left enhance sight distance because the driver of a vehicle following a truck can see past the truck without having to move closer to the opposing lane to see approaching vehicles. Unfortunately, a curve to the left for the one direction of flow is a curve to the right for the opposing vehicles. The right-hand curve requires that a driver wishing to overtake would have to venture fairly far to the left, possibly even into the opposing lane to check for a passing opportunity. On minimum radius curves, the probability of a head-on crash is too high to be disregarded.

As an alternative, moving to the right-hand side of the lane and possibly even onto the shoulder may make it possible to see past the right-hand side of the truck. In either case, the overtaking manoeuvre would have to commence at a greater distance behind the truck, generating a need for a longer passing sight distance than would otherwise be the case. In the case of a leading passenger car, drivers can often see through the leading vehicle so that the restriction of sight distance is less severe.

The curve to the left vis-à-vis the curve to the right is an issue in the consideration of the percentage passing sight distance. Obviously, the percentage passing sight distance has to be assessed for both directions. The problem of lack of passing sight distance on curves to the right can be minimised either by shortening the radius of curvature, hence increasing the tangent lengths on either side of the curve, or by increasing the radius of curvature. This serves to demonstrate that it is only in the classroom that there is one 'correct' answer to any problem.

Impediment to sight distance caused by trucks is an issue only where there is a high percentage of truck traffic on the road. Trucks normally constitute about 5 per cent of the traffic stream. At this level it is not necessary to make adjustments to the alignment to improve passing sight distance.

6.2.4 Aesthetics of horizontal alignment

Aesthetics are discussed in depth in Chapter 9. In that chapter, the focus is largely on the combination of horizontal and vertical alignment features. It should be noted that horizontal curves should not be so short that they creation the impression of a kink in the road's location. For a design speed of 120 km/h, the minimum radius curve need only be about 60 metres long for a 5° deflection angle. Practical experience indicates, however, that for aesthetic reasons curves should be at least 150 metres long for this deflection angle and that every 1° reduction in deflection angle should be accompanied by a 30-metre increase in curve length. As a function of design speed, the minimum curve length in metres should be three times the design speed in kilometres per hour on major roads and six times the design speed in the case of freeways.

6.3 TANGENTS

6.3.1 Introduction

Tangents are defined by their length and their bearing, sometimes referred to as their heading. These characteristics are discussed further in the sections that follow.

6.3.2 Lengths of tangents

In flat terrain, there are few constraints on the length of tangents. They can be anything up to 30 or even 50 kilometres long.

There are three conditions that have to be considered in determination of the maximum length of tangents:

- Approaching headlights creating a problem of dazzle with the growing popularity of high intensity discharge (HID) or bi-xenon headlights exacerbating the problem, as they can be up to three times brighter than conventional halogen headlights
- Driver boredom resulting in the possibility of the driver falling asleep at the wheel
- The safety implications of speed variations resulting from drivers accelerating after negotiating a short radius curve and then decelerating again at the next curve

6.3.2.1 Dazzle

When opposing vehicles are separated by 60 seconds of driving time, dazzle is not yet a problem but it does have a certain amount of nuisance value. At travel speeds of 120 km/h, the vehicles will be 4 kilometres apart. By the time that the separation has dropped to 30 seconds, or 2 kilometres, sight distance is starting to be impeded, that is, dazzle becomes a problem.

On two-lane roads little can be done to combat dazzle by reducing the length of tangent simply because the closer the opposing vehicles are to each other the worse the dazzle becomes. It is suggested that limiting the length of tangents to, say, 2 kilometres, would at least reduce the time of exposure to dazzle. On divided highways and streets, shrubbery on the median would eliminate dazzle provided that it grows to higher than about 1 metre.

6.3.2.2 Driver boredom

In the United Kingdom, it has been suggested that two million drivers a year fall asleep at the wheel. According to the Harvard Medical School, 250,000 American drivers fall asleep at the wheel every day. Drowsiness can be as result of exhaustion but it is suggested that boredom also has plays a role. Very long tangents require little effort from the driver and control of the vehicle is at a subconscious level. The constant sound of the engine and the tyres on the road and the stroboscopic effect of the centrelines disappearing under the bonnet of the car all conspire to cause the driver to fall asleep in what is actually a light hypnotic trance. For this reason, drivers who are ostensibly asleep can actually travel for considerable distances before the inevitable ROR crash occurs.

African experience (SATCC, 1997) suggests that crash rates at the minimum tangent length (being sufficient only to roll the superelevation from a curve in one direction to the curve in the opposite direction) and at a tangent length of about 25 kilometres are similar. Between these two extremes, the crash rate follows a roughly parabolic curve, with its low point at a tangent length of about 10 to 12 kilometres. It is suggested that drivers apparently need to actually do something at about 5- to 6-minute intervals simply to stay awake and that a maximum tangent length of 10 to 12 kilometres should for preference not be exceeded.

6.3.2.3 Speed variations

On long tangents operating speeds of 120 to 140 km/h are common and these seem to be the generally desired speeds of travel. Restricting the length of the tangent to effect some or other reduction in speed, if the design speed of the horizontal curves is 120 km/h, is an exercise in futility.

If the terrain is such that a reduced design speed is necessary, the length of tangent has to be reduced. This is because, if the tangents are still long, there will be a tendency to accelerate at the end of one curve towards the generally desired speed and then to decelerate at the start of the next. As a general rule, as stated previously, short curves are to be avoided at the ends of long tangents because drivers lose the sensation of the speed they're travelling at and may enter such curves at too high a speed.

The oscillation in speed level is potentially hazardous because a following driver may not expect the drop in speed of the leading vehicle at the entry to the next curve. A driver may also not decelerate sufficiently and thus be forced into a situation of breaking sharply on the curve with the possible loss of control that this implies.

It has been found that, at speeds of 100 km/h or less, drivers will tend to maintain a steady speed if the tangent length, in metres, is shorter than 10 to 20 times the design speed in kilometres per hour. For example, at a speed of 80 km/h, tangents should desirably not be longer than 800 metres and definitely not longer than 1600 metres.

6.3.3 Bearing of tangents

An east–west orientation of tangents should, if possible, be avoided. At sunrise and sunset, the sun would dazzle the driver. A route alignment comprising a series of tangents zigzagging around a mean east–west heading may be an option worth pursuing. Not to appear forced, the tangents should have a reasonable length of, say, 2 to 5 kilometres and the bearing should be at least 10° off due east/west.

At other times of the day, a combination of bearing and gradient could also be a problem. For example, In South Africa, few people leave Ladybrand to travel to Bloemfontein at 16:00 h simply because the length, bearing and gradient of the first tangent on leaving Ladybrand has

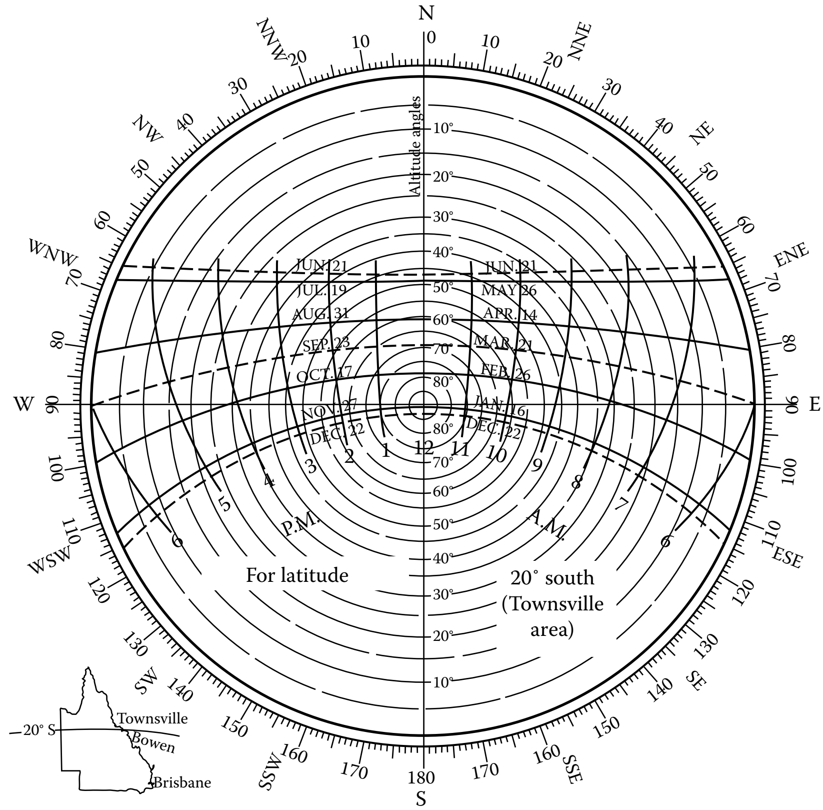


Figure 6.1 Solar chart. (From Queensland Department of Transport and Main Roads. *Chapter 12, Road planning and design manual.* Brisbane, 2002.)

the sun on the centreline for an uncomfortably long period of time. A Solar Chart (Queensland Department of Transport and Main Roads, 2001), as illustrated in Figure 6.1, would provide an indication of combinations of bearing and gradients that, if possible, should be avoided.

The centre of the diagram represents the observer's position. The heavy curved lines represent the sun's path for selected dates and, in the preceding example, latitude 20° S and are crossed by lines indicating hours. To find the sun's position for the required conditions, select the point where the appropriate lines intersect. The sun's altitude (in degrees above the horizontal plane) is shown by the relation of this point to the concentric circular lines within the diagram. The direction of the sun's rays is shown by a line drawn through this point from the outer graduated circle towards the centre.

The University of Oregon Solar Radiation Monitoring Laboratory is a source of solar charts for any location on the earth's surface and its website is http://www.solardat.uoregon.edu/SunChartProgram.html.

6.4 CURVES

6.4.1 Introduction

The majority of crashes on rural roads occur on curves, and the crash rate increases with reducing radius of curvature. Many of these are run-off-the-road (ROR) single-vehicle crashes, typically attributed to selection of a speed inappropriate to the radius. Small radii of curvature are often associated with large angles of deviation and hence reduced sight distance. As a general principle, designers should seek to use large radii of curvature and small angles of deviation, with the latter being determined in the early stages of route location.

In this section, minimum radii for a range of design speeds are shown. A change of direction from a tangent to a short radius curve requires a fairly sharp movement of the steering wheel and this may not always be successfully executed. A spiral curve preceding or following a short radius curve would ease the transition from an infinite radius (i.e. tangent) to a short radius and vice versa. These curves are useful for the development of superelevation, which is discussed in depth in the text that follows and superelevation rates are offered for minimum and above-minimum radii of curvature. Roads with low design speeds often have horizontal curves following each other in quick succession, and dealing with sequences of curves is discussed in Section 6.4.8.

6.4.2 Length of curves

6.4.2.1 Minimum length

Very small deflection angles do not require that horizontal curves be used. However, as the size of the deflection angle increases, the change of direction will become visible as a kink and a curve has to be inserted between the intersecting tangents. The maximum deflection angle without the application of a circular curve is 1° for two-lane roads. For minor roads, a maximum deflection angle of 1° 30′ does not have to be provided with a circular curve. For multi-lane roads, the maximum deflection angle without a circular curve is 0° 30′. On high-order roads – freeways and expressways – all changes of direction should be by means of circular curves regardless of the extent of the deflection angle (Transit New Zealand, 2002).

Where curves are required, their minimum length is determined by aesthetics. The eye of the beholder suggests that this is very subjective but, in general terms, a minimum length of curve of 150 metres would be adequate to avoid the appearance of a kink. It is suggested that, on expressways, a minimum length of 200 metres would suffice. On freeways, this should be increased to 300 metres. For deflection angles of less than 5° 0′, the length of the curve should be increased from 150 metres by 30 metres for each 1° decrease in the deflection angle (SATCC, 1997).

6.4.2.2 Maximum length

A long curve, particularly if it is of near minimum radius, may cause tracking problems. These are experienced by vehicles travelling at speeds markedly different from the design speed of the road. Instead of following a smoothly curving path, drivers start making a series of relatively small corrections that can increase in size until the curve is being negotiated as a series of short paths alternately of longer and shorter radius than the actual radius of the curve. Although usually not critical, these do suggest that the driver is less than relaxed behind the wheel.

The chief complication introduced by the long curve is its possible effect on passing opportunities. On a two-lane road the principal obstruction to sight distance often is the vehicle to be overtaken. On a left-hand curve, passing sight distance is enhanced by the presence of the curve. This is partly because of the ability to see easily past the leading vehicle and partly because the superelevation results in a slightly higher eye height increasing the extent of the available sight distance. On right-hand curves, the overtaking path (being an outside path) is longer than that on a left-hand curve (an inside path). The overtaking manoeuvre on the right-hand curve also has start from further back behind the leading vehicle than that on the left-hand curve. This is because the overtaking vehicle has to move out towards if not beyond the centreline of the road to be able to check the availability of passing sight distance.

Consequently, the maximum length of a curve should not exceed about 1000 metres, with the preferred maximum length being 800 metres. On curves with very large radii, that is, greater than 3000 metres, the limitation on maximum curve length is no longer applicable.

6.4.3 Minimum radii of curvature

The minimum radius of curvature is a function of the speed at which it is to be traversed. This is defined by Newtonian physics but the coefficients of friction applied in design are derived from a human factors approach of values that drivers and their passengers find comfortable rather than from limits of adhesion that, if exceeded, result in a vehicle sliding out of control. The vector forces in play are illustrated in Figure 6.2.

The equation for f, the lateral or side friction factor, is

$$e + f = \frac{V^2}{127R}$$

where

e = superelevation (taken as positive where the slope is downward towards the centre of the curve)
V = speed of vehicle (km/h)
R = radius of curvature (m)

Slightly recast, the minimum radius is calculated from

$$R_{\min} = \frac{V^2}{127(0.01e_{\max} + f_{\max})}$$

with the values of $f_{\max}$ shown in Figure 4.2 (which is repeated for convenience as Figure 6.3) adopted for calculation. Minimum radii for design speeds of 40 to 130 km/h and values of $e_{\max}$ of 4 per cent to 12 per cent are shown in Table 6.1.

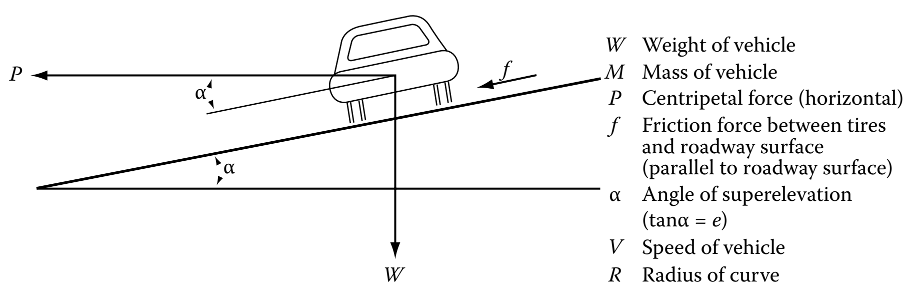


Figure 6.2 Forces involved in driving on a curve. (From Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited [SANRAL]. Pretoria, 2002.)

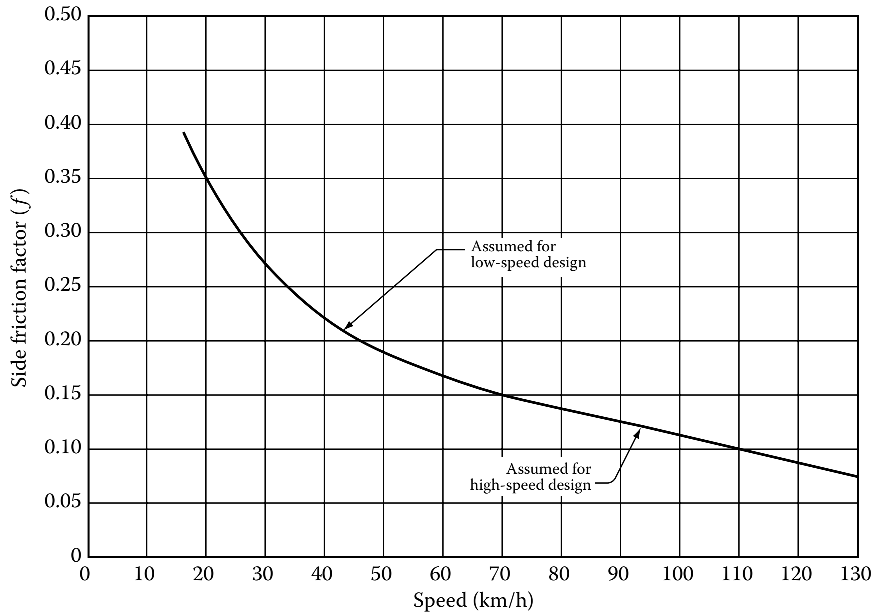


Figure 6.3 Values of f_{max}. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on geometric design of highways and streets*. Washington, DC, 2001.)

Table 6.1 Minimum values of radii for limiting values of e and f

Design speed (km/h)	*Minimum radius (m) using e_{max} of (%)*				
	4	*6*	*8*	*10*	*12*
40	50	45	40	40	35
50	85	80	75	70	65
60	135	125	115	105	100
70	205	185	170	155	145
80	280	250	230	210	195
90	375	335	305	275	255
100	490	435	395	360	330
110	N/A	560	500	450	415
120	N/A	750	665	595	240
130	N/A	950	830	740	665

6.4.4 The accuracy of setting out

The design process terminates with a host of drawings containing bearings and tangent lengths, levels, radii of curvature, and so on. To convert these data from lines on paper to lines on the ground, it would be useful to provide the surveyor with a list of coordinates, setting out data, in fact.

At this point, comment could perhaps be made about the 'purity of science' versus its alternative, the engineering approach of 'close enough is good enough'. A bearing can be calculated to within fractions of a second; a tangent length or gradeline level in nanometres is a mathematical possibility. Observations using a total station can match this accuracy. However, there are limits to the accuracy that can be achieved during construction.

An expert grader operator, whose eye level would be close to 3 metres above the road, can barely manage to cut the base course to a tolerance closer than about 10 millimetres. It is suggested that any effort directed towards ensuring that levels are accurate to less than this are a waste of time. The caveat is that cumulative errors can occur and caution should be exercised to ensure that the final level at any point is at least as accurate as the tolerance of construction. Concrete or asphalt pavers guided by piano wire or laser can, of course, achieve high levels of accuracy but, in terms of safety or comfort of the road users, this probably is not really necessary.

The situation is similar with regard to lengths. A closing error of 150 millimetres on the end of a circular would, provided it is on the new tangent and at the correct bearing, is less than the contact area of a tyre on the road and, as such, would, although it would no doubt offend the purist, be of no consequence to the driver. Once again, the possibility of cumulative errors should not be ignored.

In the case of drainage, gradients flatter than 0.5 per cent are becoming problematic and a very special effort would have to be put into ensuring an acceptable level of accuracy at 0.3 per cent. The problem is the likelihood of flat spots or even slight upgrades in the drain profile could arise from inaccuracy in the levelling process. This could result in local damming and water being diverted out of the drain and flowing onto the road.

6.4.4.1 The circular curve

As stated previously, the basic equation for a circular curve is

$$R = \frac{V^2}{127(e+f)}$$

where
- R = radius (m)
- V = speed (km/h)
- e = superelevation rate (decimal)
- f = side friction factor (decimal)

The elements of a circular curve are illustrated in Figure 6.4.

The symbols shown in this figure have the following significance:

- Δ = deflection angle (degrees)
- T = tangent distance (m) = distance from point of intersection (PI) to the beginning of the curve (BC), that is, point of curvature (PC) × tangent curve (TC) or to the end of curve (EC), that is, point of tangency (PT) × curve tangent (CT)
- L = length of curve = distance along curve from BC to EC
- R = radius of curvature (m)
- m = external distance (m) = PI to midpoint of curve
- C = centre of curve
- LC = long chord (m) = BC to EC
- M = middle ordinate (m) = midpoint of arc to midpoint of long chord
- a = length of arc from BC to any point on the curve (m)
- c = length of chord from BC to any on the curve (m)
- ϕ = deflection angle from BC to any point on curve (degrees)
- t = distance along the tangent from BC to any point on the curve (m)
- o = tangent offset to any point on curve

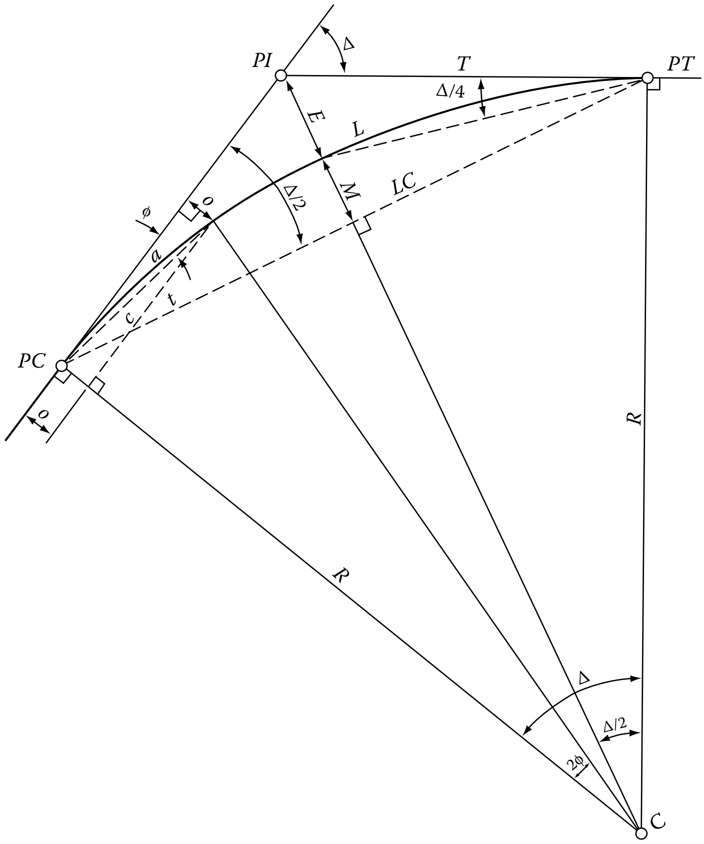


Figure 6.4 Elements of a circular curve. (From Caulfield B. *Topic 8: Horizontal curves*. Trinity College, Dublin, 2012.)

Curve Formulae

$$T = R \tan (\Delta/2)$$

$$c = 2R \sin \frac{90a}{\pi R}$$

$$L = \frac{}{360} 2\pi$$

$$\cos\phi = \frac{R - o}{2R}$$

$$m = \frac{R}{\cos \frac{\Delta}{2}} - R = T \tan \Delta/4$$

$$t = R \sin 2\phi$$

$$LC = 2R \sin \frac{\Delta}{2} = 2T \cos \frac{\Delta}{2}$$

$$o = c \sin \phi$$

$$M = R \ 1 - \cos \frac{\Delta}{2} = E \cos \frac{\Delta}{2}$$

$$o = R - \sqrt{R^2 - t^2}$$

$$a = \frac{\Delta \pi R}{90}$$

$$o = R(1 - \cos 2\phi)$$

For the purposes of setting out of curves on the open road, that is, other than at intersections or interchanges, the first step that generally would be required is the calculation of the bearing of the tangents on either side of the curve. This is by means of a join or traverse between successive points of intersection (*PI*), from which the angle of deflection, Δ (sometimes referred as the deviation angle), is calculated. The selection of the curve radius, *R*, makes it possible to calculate the tangent length, *T*, of the curve and hence the coordinates of the *BC* or end of curve, *EC*, by means of a polar in the backward direction in the case of the *BC* and in the forward direction for the *EC*. The join and the polar are discussed in Chapter 22.

The other formulae offered in the preceding list may be required in more complex cases such as in the setting out of interchange ramps or intersections.

6.4.5 Transition curves

Although horizontal alignments frequently have tangents joining directly onto circular curves, it is physically impossible for drivers to rotate their steering wheel instantaneously from steering straight ahead to following the curve of the road. They drive along a spiral path that provides a transition from the tangent to a curve with the radius as selected by the designer.

If it is decided that a transition curve should be provided, the properties of the transition curve would have to include

- Its being tangential to the preceding straight
- An infinite radius (i.e. zero curvature) at the tangent end of transition (*TS*)
- A linear decrease in radius from infinity to the radius of the circular curve that it precedes
- A radius at the circular curve end of the transition (*SC*) equal to that of the circular curve
- Joining the circular arc tangentially

Transition curves are discussed in detail in Chapter 22.

The only curve that possesses these properties is the clothoid. An alternative curve that can be used is the cubic parabola. This is not a true spiral because the change in radius along the transition is not proportional to the distance travelled along it. Its principal attraction is that it is easier to set out than the clothoid.

The radii of curvature below which transition curves should be considered are shown in Table 6.2. On curves with large radii, the spiral followed by the vehicle entering a circular

Table 6.2 Radii below which transition curves may be considered

Design speed (km/h)	*Radius (m)*
40	95
50	150
60	215
70	290
80	380
90	480
100	590
110	715
120	850
130	1000

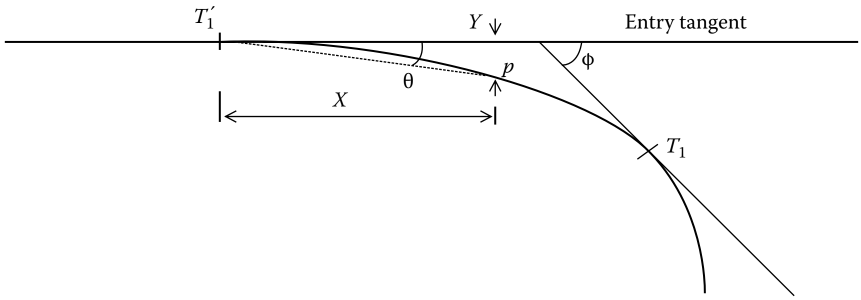


Figure 6.5 Coordinates of the transition curve.

curve is accommodated within the width of the travelled lane. To achieve this condition on shorter radii, it may be necessary to provide widening of the road to allow for the off-tracking of large trucks and semitrailers. This curve widening can conveniently be contained within the length of the transition curve so that the outer edge of the lane follows the natural path template followed by vehicles on entry into a curve.

Transition curves provide an aesthetically pleasing aspect to the road, and this is more important than being merely pretty pictures. Driving along a road with transition curves can be a pleasurable experience so that a driver on such a road would be relaxed and more readily able to respond to any situation that may arise. By way of contrast, a driver who is tense is more likely to respond inappropriately to unexpected situations.

Calculation of the angle between the line joining T_1' and p would require the use of tables of standard data or the use of a computer. Setting out of transition curves is often done by the use of distances along the tangent and offsets as illustrated in Figure 6.5. A series expansion for the clothoids is

$$Y = \frac{l^3}{6RL} - \frac{l^7}{336(RL)^3} + \frac{l^{11}}{42{,}240(RL)^5} - \ldots$$

where
- l = distance along the curve (m)
- R = radius of circular curve (m)
- L = length of transition curve

Only the first term of the expansion needs to be considered because at minimum values of R (where the transition is most likely to be applied) and maximum length, L, of the transition, the second term would suggest a reduction in length of the offset of the order of 1 millimetre, which could, it is suggested, be ignored. The points of interest of the transition curve are illustrated in Figure 6.5.

6.4.6 Superelevation

It was found in the 1920s (Collins and Hart, 1935) that, with the higher speeds of passenger cars, drivers were starting to use the inside lane for negotiating curves regardless of their direction of travel. This was to take the advantage of the camber towards the inside of the curve. To remove this potentially dangerous situation, superelevation was introduced with the slope of the inside lane being extended to the full width of the travelled way. Theoretically, there is no need to apply any more superelevation than this extension of the inner camber except that the radius of curvature required at high design speeds would be so high that location of the horizontal alignment would become problematical. Steeper superelevations up to a maximum of 12 per cent would enable the use of significantly shorter radii, thus easing the problems of route location. Experience in South Africa suggests that superelevation is a powerful tool in the armoury of the geometric designer.

Values of 4 per cent or 6 per cent for e_{max} are the preferred options for urban streets. This is because the cross-section of the road should be as close as possible to the natural ground level in support of ready access from adjacent properties to the road. It also facilitates drainage of the surrounding properties. For this reason the cross-section should preferably be slightly below the natural ground level. High values of superelevation in urban areas could result in fill heights or cut depths that could be sufficient to make access from adjacent properties difficult. Furthermore, at high values of superelevation, superelevation development could require a distance of the order of 100 metres or more. Closely spaced intersections would make it difficult to adequately develop superelevation.

On rural roads, maximum superelevation rates are normally in the range of 6 to 10 per cent with a maximum of 12 per cent partly because of practical problems of construction. Where icing conditions are likely to occur, the maximum level of superelevation should not exceed 8 per cent.

The superelevation tables (Tables 6.4 to 6.8) include values for e_{max} of 12 per cent. This is the maximum practical value of e_{max} but, as a general rule, e_{max} should not exceed 10 per cent, particularly where steep gradients are likely to be encountered. It has been observed that trucks with high loads, such as bales of dry fodder, can lose their loads while attempting to negotiate a 12 per cent superelevation at a crawl speed. As a personal observation, it should be noted that a 12 per cent superelevation also has the appearance of the wall of death, usually found only in fun fairs.

As a further problem of the 12 per cent superelevation rate, light rain after a long dry spell dramatically reduces side friction, particularly where the road surface is polluted by rubber and oil spills. It has been said that storm water runoff after a long dry spell is far more polluted than anything found in a foul water sewer. It should be borne in mind that a combination of a 12 per cent superelevation and an 8 per cent longitudinal gradient would provide a resultant of nearly 14.5 per cent at angle of 60° to the centreline of the road.

Whatever value selected for e_{max}, it should be applied consistently an on area-wide basis. Its selection is the determinant of the superelevations applied to all curves with radii above the minimum. It follows that variations in e_{max} would lead to variations in e for curves of the same radius. From a human factors perspective, it is likely that drivers select their speed for entry into a curve on the basis of the appearance of the radius of the curve, with the superelevation provided being of lesser importance in the decision-making process. A lack of consistency in superelevation rate would thus result in differences in the demand for side friction and, in extreme cases, the demand may be sufficiently high to cause discomfort to the occupants of the vehicle even if not actually exceeding the supply of side friction.

Recommended values of e_{max} are listed in Table 6.3.

AASHTO (2011) describes four different ways of distributing values of e and f for curves of above minimum radii:

- Superelevation and side friction are simultaneously increased in direct proportion to the inverse of the radius of the curve.
- For a vehicle travelling at design speed, first f and then e are increased in inverse proportion to the radius of the curve.

Table 6.3 Design contexts for recommended values of e_{max}

Design context	*Recommended value of* e_{max} *(%)*
Minor urban roads	4–6
Rural roads	6–8
High-speed urban roads	8–10
Freeways and arterials	8–10
Not recommended	>10

- For a vehicle travelling at design speed, first *e* and then *f* are increased in inverse proportion to the radius of the curve. A further distribution is also described as being similar except that it is based on average running speed rather than design speed.
- Superelevation and side friction are simultaneously increased in a curvilinear relationship with the inverse of the radius of the curve.

The first method of distributing *e* and *f* is illustrated in Figure 6.6.

This form of distribution is strongly related to the human factors approach to design, as it finds resonance in drivers' preference for travelling at constant speeds. It results in side friction factors ranging in a linear fashion from zero on tangents to f_{max} on the minimum radius curve.

The second form of distribution is very useful in urban areas because superelevation is applied only after the side friction has achieved its maximum value, with this value of side friction being available for all curves with a smaller radius. As pointed out, spatial limitations can make it difficult to provide sufficient superelevation in terms of the other forms of distribution. Fully utilising side friction before application of superelevation becomes necessary is thus a useful device under constrained circumstances.

Superelevation rates for above minimum radii and e_{max} = 4 per cent, 6 per cent, 8 per cent, 10 per cent and 12 per cent are shown in Tables 6.4 to 6.8 respectively.

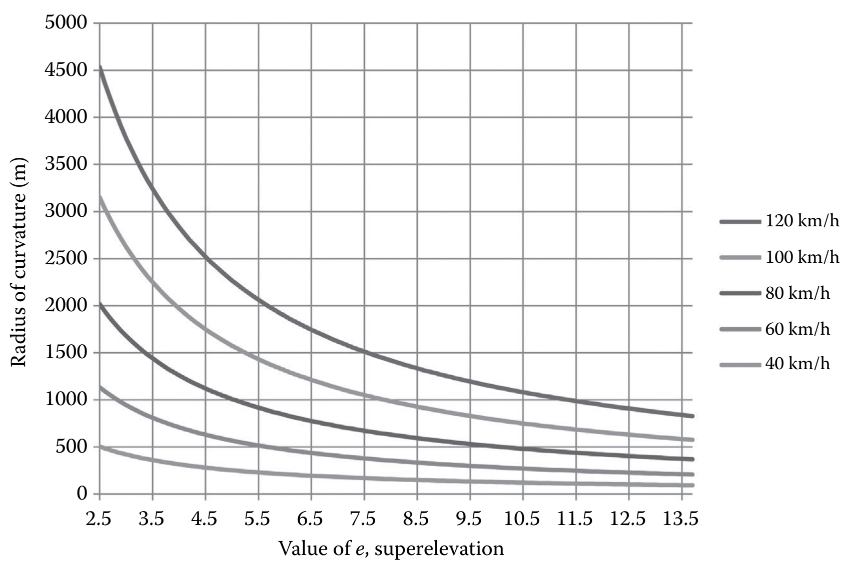


Figure 6.6 Distribution of e.

Table 6.4 Superelevation rates for above minimum radii and e_{max} = 4%

Radius (m)	e *for design speed (km/h) of*									
	40.000	*50.000*	*60.000*	*70.000*	*80.000*	*90.000*	*100.000*	*110.000*	*120.000*	*130.000*
4000.000										NC
3500.000									NC	RC
3000.000								NC	RC	RC
2500.000							NC	RC	RC	0.027
2000.000						NC	RC	RC	0.028	0.031
1750.000						RC	RC	0.027	0.031	0.034
1450.000					NC	RC	0.027	0.030	0.034	0.038
1200.000				NC	RC	0.027	0.030	0.034	0.037	0.040
1000.000				RC	RC	0.029	0.033	0.037	0.040	
850.000			NC	RC	0.028	0.032	0.036	0.039		
750.000			RC	0.026	0.030	0.034	0.037	0.040		
600.000		NC	RC	0.029	0.033	0.037	0.040			
550.000		RC	0.026	0.030	0.034	0.038				
450.000	NC	RC	0.028	0.033	0.037	0.400				
375.000	RC	0.026	0.031	0.035	0.039					
300.000	RC	0.028	0.034	0.038	0.400					
250.000	0.026	0.031	0.036	0.040						
200.000	0.028	0.034	0.039							
175.000	0.030	0.035	0.040							
150.000	0.031	0.037								
130.000	0.033	0.039								
110.000	0.035	0.040								
95.000	0.037									
80.000	0.039									
65.000	0.040									

Note: NC, normal camber; RC, reverse camber.

6.4.7 Superelevation development

Superelevation development has two components: tangent runout and superelevation runoff. Tangent runout involves the rotation of the outside lane(s) of the cross-section from the normal camber, usually 2.5 per cent, to a zero crossfall. Superelevation runoff then continues this rotation until a crossfall equal to the slope of the normal camber across the full width of the travelled way is achieved. From this point further, the entire width of the travelled way is rotated until the full superelevation appropriate to the design speed and radius of curvature is achieved. The process is illustrated in Figure 6.7 for the case of rotation around the centreline.

The axis of rotation can, in fact, be located anywhere across the cross-section or even outside it. Selection of its location is dependent on the constraints under which the superelevation has to be developed. This is particularly so in the case of superelevation development in urban areas. These constraints could involve issues of drainage, aesthetics or fitting the cross-section to the topography. The problem to be solved is largely one of the location of the road edges relative to the ground line. In the case of a two-lane road, the axis of rotation would typically be located on the centreline. Other standard locations are the inside and outside edges of the travelled way.

Table 6.5 Superelevation rates for above minimum radii and e_{max} = 6%

Radius (m)	e *for design speed (km/h) of*									
	40	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*	*130*
4000										NC
3500									NC	RC
3000								NC	RC	RC
2500							NC	RC	RC	0.030
2000						NC	RC	RCR	0.033	0.037
1750					NC	RC	RC	0.031	0.036	0.040
1450					RC	RC	0.027	0.037	0.043	0.049
1200				NC	RC	0.027	0.038	0.043	0.050	0.056
1000				RC	0.026	0.037	0.043	0.049	0.056	0.060
850			NC	RC	0.028	0.041	0.048	0.054	0.059	
750			RC	0.026	0.030	0.045	0.052	0.058	0.060	
600		NC	RC	0.029	0.044	0.051	0.057	0.061		
550		RC	0.026	0.030	0.047	0.053	0.059	0.060		
450	NC	RC	0.028	0.045	0.052	0.058	0.060			
375	RC	0.026	0.042	0.049	0.056	0.060				
300	RC	0.028	0.046	0.054	0.060					
250	0.035	0.042	0.050	0.058						
200	0.039	0.047	0.056	0.061						
175	0.041	0.050	0.058	0.060						
150	0.044	0.053	0.060							
130	0.047	0.056								
110	0.050	0.059								
95	0.053	0.060								
80	0.056	0.060								
65	0.060									

Note: NC, normal camber; RC, reverse camber.

The rotation of dual carriageways often takes place around the outer edge of the median island so that the median shoulders rotate in concert with the travelled lanes. However, as in the case of the two-way road, no hard and fast rules can be laid down concerning the selection of the location of the axis of rotation. It could be at the centreline of the median, at the edge of the median or even having the entire cross-section rotating as a unit around one of other of the outer edges of the cross-section. Each case would have to be considered on its own merits.

Figure 6.7 shows the development of superelevation as a sequence of straight lines. It is not necessary in practice to go to the refinement of replacing the instantaneous changes of gradient by vertical curves because, being short, these curves would be invisible to the driver even in the unlikely event of actually looking at them. Furthermore, the construction process would be hard put to actually build the instantaneous changes of gradient and a smoothly flowing transition would result automatically.

6.4.7.1 Length of superelevation development

The length of the superelevation development is based on aesthetics. Too short a development would show up as a sharp discontinuity in an otherwise smoothly flowing travelled way edge. On the other hand, a very long development would raise the spectre of potential flat spots on

Table 6.6 Superelevation rates for above minimum radii and e_{max} = 8%

Radius (m)	e *for design speed (km/h) of* 40	50	60	70	80	90	100	110	120	130
4000									NC	NC
3500								NC	RC	RC
3000							NC	RC	RC	0.027
2500						NC	RC	RC	0.029	0.032
2000					NC	RC	RC	0.030	0.035	0.039
1750				NC	RC	RC	0.030	0.034	0.040	0.044
1450				RC	RC	0.025	0.035	0.041	0.047	0.053
1200			NC	RC	RC	0.035	0.042	0.048	0.056	0.063
1000			RC	0.028	0.025	0.041	0.049	0.056	0.065	0.074
850		NC	RC	0.032	0.039	0.047	0.055	0.064	0.074	0.080
750		RC	0.028	0.036	0.043	0.051	0.061	0.070	0.080	
600	NC	RC	0.034	0.043	0.052	0.061	0.071	0.080		
550	RC	0.028	0.037	0.046	0.055	0.065	0.075			
450	RC	0.033	0.043	0.053	0.063	0.073	0.080			
375	0.029	0.038	0.049	0.060	0.070	0.080				
300	0.035	0.045	0.057	0.068	0.077					
250	0.040	0.051	0.063	0.074	0.080					
200	0.046	0.058	0.070	0.080						
175	0.050	0.062	0.074							
150	0.054	0.067	0.078							
130	0.058	0.072	0.080							
110	0.063	0.076								
95	0.067	0.080								
80	0.072									
65	0.080									

Note: NC, normal camber; RC, reverse camber.

the road surface and the associated drainage problems. Experience has shown that relative gradients of 0.35 per cent for a design speed of 130 km/h and 0.80 per cent for 20 km/h result in visually acceptable lengths of superelevation development. Interpolating between these values for the intervening design speeds produces the values of relative gradient shown in Table 6.9.

Using these relative gradients, the length of the superelevation runoff is calculated from

$$L = \frac{wne_{\mathrm{d}}}{\Delta} b$$

where

L = length of superelevation runoff (m)
w = width of one traffic lane (m)
n = number of lanes rotated
e_{d} = superelevation rate (%)
Δ = relative gradient (%)
b = adjustment factor for number of lanes rotated

Table 6.7 Superelevation rates for above minimum radii and e_{max} = 10%

Radius (m)	e *for design speed (km/h) of* 40	50	60	70	80	90	100	110	120	130
8000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
7000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
6000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
5000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
4000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
3500	NC	NC	NC	NC	NC	NC	NC	NC	RC	RC
3000	NC	NC	NC	NC	NC	NC	NC	RC	0.024	0.027
2500	NC	NC	NC	NC	NC	NC	RC	0.025	0.029	0.032
2000	NC	NC	NC	NC	NC	NC	0.027	0.031	0.036	0.040
1750	NC	NC	NC	NC	NC	RC	0.030	0.035	0.041	0.046
1450	NC	NC	NC	NC	RC	0.030	0.036	0.042	0.049	0.055
1200	NC	NC	NC	RC	0.029	0.035	0.043	0.050	0.058	0.066
1000	NC	NC	RC	0.028	0.035	0.042	0.050	0.058	0.068	0.081
850	NC	NC	0.026	0.033	0.040	0.048	0.058	0.067	0.084	0.091
750	NC	RC	0.029	0.037	0.044	0.053	0.064	0.083	0.089	0.100
600	NC	0.026	0.035	0.044	0.053	0.064	0.084	0.090	0.100	
550	RC	0.028	0.037	0.047	0.057	0.081	0.087	0.094		
450	0.026	0.033	0.044	0.078	0.081	0.086	0.093	0.100		
375	0.030	0.039	0.050	0.081	0.085	0.091	0.100			
300	0.037	0.046	0.059	0.084	0.091	0.100				
250	0.043	0.053	0.082	0.088	0.096					
200	0.051	0.081	0.086	0.094	0.100					
175	0.056	0.083	0.089	0.098						
150	0.062	0.085	0.093	0.100						
130	0.068	0.088	0.097							
110	0.088	0.091	0.100							
95	0.091	0.094								
80	0.095	0.100								
65	0.100									

The adjustment factor, *b*, arises as a practical consideration because, without it, the length of the superelevation would double, treble or more depending on the number of lanes rotated. There may simply be not enough space to accommodate these lengths of runoff development. On a purely empirical basis, it is recommended that the calculated lengths be adjusted downwards by the adjustment factors offered in Table 6.10.

6.4.7.2 Location of superelevation development

Where a transition curve is used, superelevation runoff should be contained within the length of the transition curve. Given the relationship above for length of runoff, it follows that this also defines the length of the transition curve, enabling the calculation of the rest of the setting out data of the transition curve.

At curve radii above those for which transition curves would be provided, the two extremes of runoff location are for

Table 6.8 Superelevation rates for above minimum radii and e_{max} = 12%

Radius (m)	e *for design speed (km/h) of*									
	40	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*	*130*
8000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
7000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
6000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
5000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
4000	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
3500	NC	NC	NC	NC	NC	NC	NC	NC	RC	RC
3000	NC	NC	NC	NC	NC	NC	NC	RC	0.024	0.027
2500	NC	NC	NC	NC	NC	NC	RC	0.025	0.029	0.032
2000	NC	NC	NC	NC	NC	NC	0.027	0.031	0.036	0.040
1750	NC	NC	NC	NC	NC	RC	0.030	0.035	0.041	0.046
1450	NC	NC	NC	NC	RC	0.030	0.036	0.042	0.049	0.055
1200	NC	NC	NC	RC	0.029	0.035	0.043	0.050	0.058	0.066
1000	NC	NC	RC	0.028	0.035	0.042	0.050	0.058	0.068	0.081
850	NC	NC	0.026	0.033	0.040	0.048	0.058	0.067	0.084	0.091
750	NC	RC	0.029	0.037	0.044	0.053	0.064	0.083	0.089	0.100
600	NC	0.026	0.035	0.044	0.053	0.064	0.084	0.090	0.100	
550	RC	0.028	0.037	0.047	0.057	0.081	0.087	0.094		
450	0.026	0.033	0.044	0.078	0.081	0.086	0.093	0.100		
375	0.030	0.039	0.050	0.081	0.085	0.091	0.100			
300	0.037	0.046	0.059	0.084	0.091	0.100				
250	0.043	0.053	0.082	0.088	0.096					
200	0.051	0.081	0.086	0.094	0.100					
175	0.056	0.083	0.089	0.098						
150	0.062	0.085	0.093	0.100						
130	0.068	0.088	0.097							
110	0.088	0.091	0.100							
95	0.091	0.094								
80	0.095	0.100								
65	0.100									

- The full superelevation to be available at the start of the circular curve
- Only tangent runout to be achieved at the start of the circular curve

Neither of these extremes is acceptable, as they would create high levels of lateral acceleration. The first suggests that a vehicle could be on a tangent with a crossfall of, say, 12 per cent, which obviously would be difficult, if not impossible, to negotiate. The second would have the vehicle in the outside lane entering the curve on a zero crossfall with the loss of control that this would almost certainly engender.

The ideal is to have some of the superelevation runoff taking place on the tangent with the balance located beyond the start of the curve. Practical experience has found that having two-thirds of the runoff on the tangent and one-third on the curve is a reasonable compromise. The reason for the acceptability of the compromise is that the vehicle entering the circular curve is, in fact, following a transition curve that is contained within the width of the lane. This transition curve is generally split between the tangent and the curve in the ratio of 2:1.

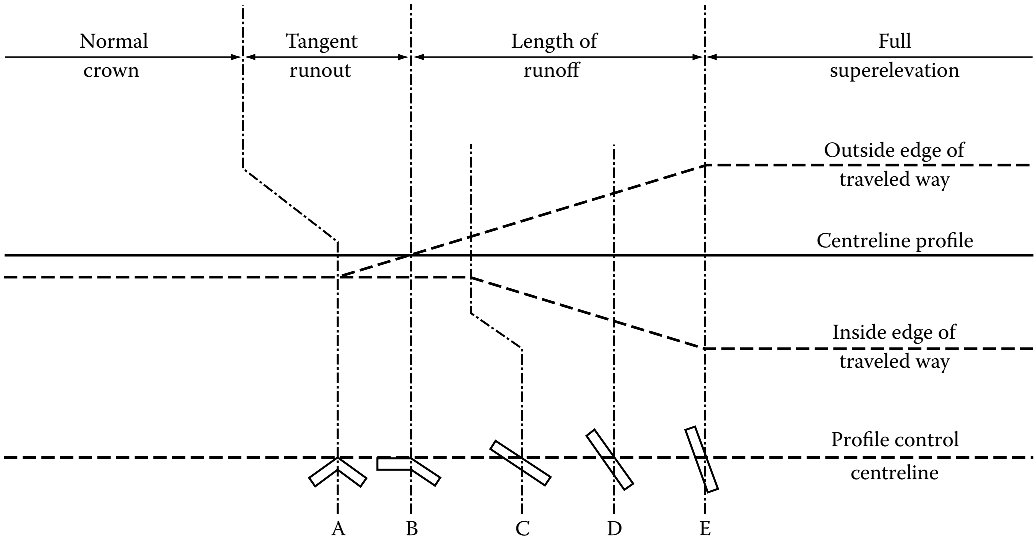


Figure 6.7 Superelevation development. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on geometric design of highways and streets*. Washington, DC, 2001.)

Table 6.9 Relative gradients of superelevation development

Design speed (km/h)	*Maximum relative gradient (%)*	*Equivalent max relative slope 1*
20	0.8	125
30	0.76	132
40	0.72	139
50	0.68	148
60	0.64	157
70	0.60	168
80	0.55	180
90	0.51	195
100	0.47	212
110	0.43	232
120	0.39	256
130	0.35	286

Table 6.10 Adjustment factor for number of lanes rotated

Number of lanes rotated, n	*Adjustment factor,* b	*Length increase relative to one lane rotated*
1.0	1.0	1.00
1.5	0.83	1.25
2.0	0.75	1.50
2.5	0.70	1.75
3.0	0.67	2.00
3.5	0.64	2.25

Note: For other values of n, use the equation $b = [1 + 0.5(n - 1)]/n$.

If for any reason space precludes providing the 2:1 split in the location of the runoff, it is possible to change this location but it is recommended that the change should not be more than 10 per cent in either direction.

6.4.8 Curve widening

In the case of narrow lane widths, tight radii of curvature may make it impossible for vehicles to negotiate these curves without encroaching either on the shoulder or the adjacent lane. The problem is that the rear wheels of the vehicle track inside the path of the front wheels. Articulated vehicles with multiple axles and pivot points are particularly prone to marked off-tracking. The wheel path is the area encompassed by the path of the outer front wheel and the inner rear wheels.

The swept area is defined by the paths of the outermost front point and the innermost rear point of the vehicle body. The swept area is sometimes also referred to as the wall-to-wall path because of its application in the layout of parking garages. Even if the wheel path is contained within the width of the lane, it is possible that the swept area may still encroach on either the shoulder or the adjacent lane. The possible presence of pedestrians and cyclists on the shoulder suggests that it is the swept area and not the wheel path that ultimately defines what the lane width should be.

The lane width on the curve should thus make allowance for

- The track width, U
- The front overhang of the vehicle, F_A
- An empirical clearance between the outer wheel path and the edge of the travelled way, C
- A further clearance, Z, to allow for the complexity of negotiating the curve

The width of the travelled way can be calculated as (AASHTO, 2011)

$$W_C = N(U + C) + F_A(N - 1) + Z$$

where
N = the number of lanes to be widened

The various elements of the relationship are calculated as follows:

1. The track width is calculated as

$$U = u + R - (R^2 - \Sigma l_i^2)^{0.5}$$

where
U = track width on curve (m)
u = track width on tangent (m)
R = radius of turn (m)
l_i = wheel base of design vehicle between successive axles and pivot points (m)

Strictly speaking, the radius, R, should be the radius of the path of the midpoint of the front axle. However, it is close enough to apply the radius of the road centreline to the equation.

2. The front overhang is the distance from the front axle of the vehicle to the furthest projection of the vehicle body in front of the front axle. In the case of the turning vehicle, the width of the front overhang is defined as the radial distance between the

path followed by the outer front edge of the vehicle and the tyre path of the outer front wheel. The width of the front overhang is calculated as

$$FA = [R^2 + A(2L + A)]^{0.5} - R$$

where
FA = width of front overhang (m)
R = radius of curve (m)
A = front overhang (m)
L = wheel base of single unit or tractor (m)

3. It is necessary to provide an allowance, C, for lateral clearance between the edge of the roadway and the nearest wheel path, and for the body clearance between passing vehicles.
   Typical values of C are
   - 0.60 metre for a travelled way width of 6.0 metres;
   - 0.75 metre for a travelled way width of 6.6 metres; and
   - 0.90 metre for a travelled way width of 7.4 m.
4. The allowance, Z, is provided to accommodate the difficulty of manoeuvring on a curve and the variation in driver operation. This additional width is an empirical value that varies with the speed of traffic and the radius of the curve. It is expressed as

$$Z = 0.1 \ \frac{V}{R^{0.5}}$$

where
V = design speed of the road (km/h)

The rear overhang is the distance from the rear axle of the vehicle to the furthest projection of the vehicle body behind the rear axle. The rear overhang can be substantial and, in the case of a bus could be as much as 2 metres. There is a fairly widespread belief that the rear overhang can swing out beyond the track width but this is not normally the case with the radii normally encountered along a road. The width of the rear overhang is the radial distance between the outside edge of the inner rearmost tyre and the inside edge of the vehicle body. In the case of a passenger car this distance is typically less than 0.15 metre. The width of truck bodies is usually the same as the wheelbase width so that the width of the rear overhang is zero. The rear overhang is normally disregarded in the calculation of lane width.

As a general rule, values of curve widening, ($W_C - W$), where W is the width of the travelled way on tangent sections, that are less than 0.6 metre are disregarded. Lane widening is thus generally not applied to curves with a radius greater than 300 metres, regardless of the design speed or the lane width.

In negotiating a curve, drivers use the inside edge of the curve as an aiming line; they aim towards the inside of the curve as opposed to aiming away from the outside. The argument then becomes whether the widening should be applied to the inside of the curve or added to the outside. Adding the widening to the outside of the curve has the advantage that the aiming line maintains a smooth flowing appearance as opposed to the jagged edge resulting from the sudden addition and dropping of the extra width. In addition to maintaining the aesthetic integrity of the design, the quality of guidance offered to the driver is maintained. On the other hand, if the transition length equals that of the superelevation runoff, the illusion of a transition curve providing the widening on the inside of the curve could be created. This would support guidance to the driver in addition to any aesthetic benefits it may possess. In short, there is no hard and fast rule as to the side of the curve on which widening should be provided.

General practice is to provide the widening on the inside of the curve with the transition being spread across the length of the superelevation development. Narrow slivers usually have to be built by hand and are correspondingly expensive to construct in addition to having considerable nuisance value. The recommended construction process is thus that the design layers are built to the full width of the pavement as required by curve widening commencing at the point at which the widening transition is initiated. The hand application is thus limited to the surfacing. Widening could be split with each lane being widened individually. This, however, is an unnecessary refinement with a significant increase in the nuisance value attached to curve widening.

General construction practice is to provide paving in single-lane widths. This has the practical benefit of joints in the surfacing being located in the same position as the lane markings, in short where traffic loading and hence rutting is least likely to occur. Joints are areas of weakness and discontinuity where the ingress of water into the design layers of the road is an ever-present possibility. Protecting them from traffic loading aids significantly in maintaining the integrity of the paved layer.

6.4.9 Successive curves

The preceding discussion of horizontal curvature has tacitly dealt with isolated curves, that is, remote from each other and with no interaction between them. Interaction could be visually or physically. Visual interaction arises when two curves are visible to the approaching driver at the same time. Physical interaction arises from the curves being so close to each other that normal superelevation development is compromised.

Three combinations of successive curves are possible.

- The reverse or S-curve where the two curves are in the opposite directions
- The broken-back or flat-back curve where the successive curves are in the same direction
- The compound curve where the successive curves abut each other and are in the same direction but with differing radii

The reverse curve matches driver expectations, which are that a curve to the left will be followed by a curve to the right or vice versa. In terms of the merits of curvilinear alignment, the reverse curve is aesthetically pleasing, particularly so if the radii of the successive curves are similar. By way of contrast, the same cannot be said of the broken-back curve. It is generally considered to be an eyesore of note. It has been said of the broken-back curve that the designer didn't go around enough the first time! Unfortunately, topographic or other restraints may make it impossible to avoid the use of a broken back curve but a certain amount of effort put into its avoidance is to be encouraged. The flexibility of the compound curve may tempt the designer into using it to negotiate broken terrain. The temptation should, however, be firmly resisted. The human factors approach indicates that, once drivers enter a curve, they expect that its radius will remain unaltered across the length of the curve.

There is only one place where driver expectations support a compound curve and that is on the loop ramps at interchanges. Compound curves may also find application on turning roadways at at-grade intersections but, other than possibly noting that maintaining position on the turning roadway is comfortable, drivers would probably be unaware of the existence of the compound curvature of the roadway.

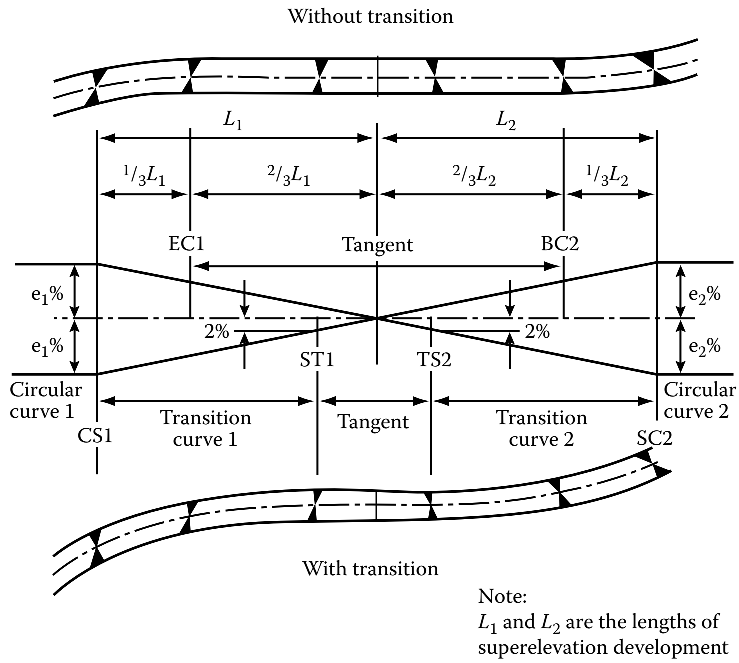


Figure 6.8 Superelevation development in a reverse curve. (From Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited [SANRAL]. Pretoria, 2002.)

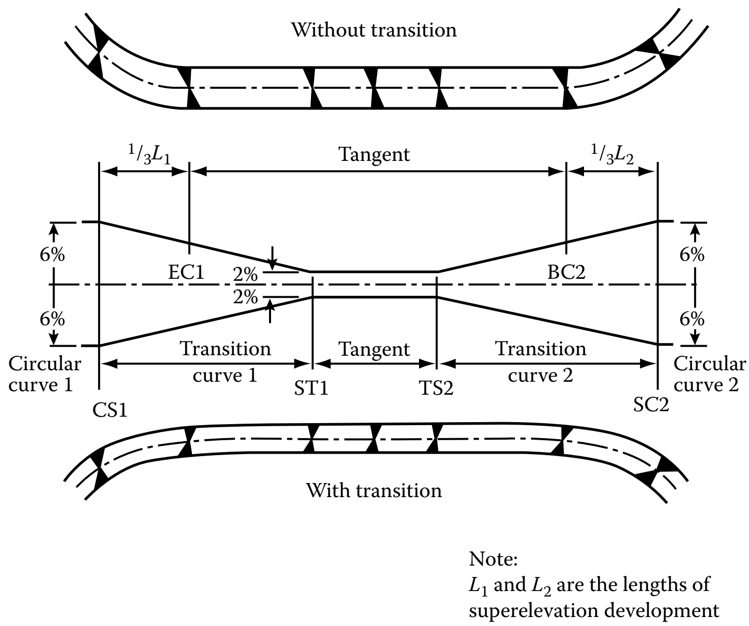


Figure 6.9 Superelevation development in a broken-back curve. (From Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited [SANRAL]. Pretoria, 2002.)

Failure is a hen and egg phenomenon. In other engineering disciplines, failure is a measurable quality responsive to the laws of physics, for example, a beam failing under load. In geometric design, there is a tendency similarly for road users to be dealt with as inanimate matter, that is, as if it were possible to represent them by fixed parameters. In fact, drivers respond according to what they see as modified by historic expectancy. Numerous examples abound. Locating the off-ramp at an interchange on the right-hand lane constitutes practical design. Drivers have now come to expect to see the off-ramp to their right. Designing a loop ramp as a spiral makes it possible to reduce substantially the area required to accommodate it. Drivers now expect to enter a loop ramp and decelerate to the speed of the smallest radius, accelerating thereafter towards the design speed of the next through road. And this is the only situation in which the use of a compound curve can be entertained.

The compound curve is, in fact, a surrogate for the spiral and came into existence purely to avoid the inconvenience of setting it out. The compound curve is discussed in detail in Chapter 12.

The reverse and the broken-back curves comprise two curves with a short length of tangent between them. The question is one of determining how short the tangent can become without negatively impacting on the superelevation development. It is suggested that the shortest tangent is the one that makes it possible to roll the superelevation over from that of the first curve to that of the second curve in a continuous fashion as illustrated in Figure 6.8 applying the appropriate relative slope. As shown, the relative slope remains constant from end to end of the roll over.

In the case of the broken-back curve, it is suggested that the superelevation runoff terminate at the point where a straight crossfall of reverse camber is achieved, as shown in Figure 6.9. The term broken-back is usually not applied if the intervening tangent is of a reasonable length, say about 500 metres. Even then, if both curves are simultaneously visible from some distance ahead, the general appearance is not good.

Chapter 7

Vertical alignment

CONTENTS

7.1 INTRODUCTION

The vertical alignment, or gradeline, of the road comprises grades and the vertical curves between them. The grade is to the vertical alignment as the tangent is to the horizontal alignment. The gradient is the steepness of the grade. It is usually expressed in percentage form, being the number of metres of rise or fall over a horizontal distance of 100 metres. Although in various texts the terms 'grade' and 'gradient' are used almost interchangeably, in this book the term 'grade' is used throughout as the reference to the straight component of the grade line or profile and the term 'gradient' as a reference to its steepness.

The gradeline defines the levels of the road immediately below the sub-base layer of the pavement. For this reason the bulk earthworks are often referred to as the subgrade.

In fitting the road to the topography, consideration has to be given to:

- Maximum gradients
- Minimum curves
- Drainage
- Cost of construction
- Aesthetics, that is, of the road itself, the abstract ribbon in space and principally comprising the coordination of the horizontal and vertical alignment
- Its relationship to the landscape being traversed

The horizontal curve, being circular, provides a constant rate of change of bearing. The vertical equivalent is the parabola that provides a constant rate of change of gradient. As such, it has a certain academic appeal, although in some countries the practice is to use the circular curve both in the horizontal and in the vertical alignments. The parabola has the benefit of ease and accuracy of calculation and provides slightly greater sight distance across the crest of curve. In practice, the differences in levels across a vertical curve between a parabola and a circle are negligible and can be ignored. In this text, only the parabola is discussed.

Factors to be considered in the development of a gradeline include

- Topography
- Road type
- Horizontal alignment
- Sight distance
- Drainage
- Heavy vehicle operational characteristics
- Appearance
- Land purchase and construction costs
- Cultural developments

The gradeline must ensure that all relevant design speed sight distance requirements are met at every point on the road alignment.

It is good design practice to make the vertical alignment design speed 10 to 15 km/h greater than the horizontal alignment design speed to provide an additional safety margin (Transit New Zealand, 2002). It is well known that drivers tend to select their speeds on the basis of the horizontal alignment and the messages they pick up from the cross-section of the road. The latter stem, for example, from wide lanes and shoulders encouraging a feeling of openness and safety at high speeds whereas narrow lanes would cause a significant reduction in speed. Obviously, the vertical alignment does not require any action from the driver. In consequence, particularly at night on an unlit road there is a tendency to overdrive the vertical alignment by a considerable margin.

7.2 GENERAL CONTROLS FOR VERTICAL ALIGNMENT

7.2.1 Sight distance

Adequate sight distance is the most important characteristic of a road. The vertical alignment must be such that at least stopping sight distance is available at every point along the road.

This has particular reference to crest curves. It also applies to sag curves where these occur on roads without the benefit of lighting. Desirably, crest curves should be designed to provide passing sight distance but, in most case, this would probably prove to be unaffordable.

The selection of values of curvature to ensure adequate sight distance is discussed in Section 7.5.

7.2.2 Topography

Although the topography has an impact on the horizontal alignment of the road, this is largely as a result of an unfavourable effect on the vertical alignment and the need to avoid excessive gradients. The effect on the vertical alignment is even more marked than that on the horizontal alignment. In general, the topography is defined in three categories: flat, rolling and mountainous. The question most frequently asked is: How does one differentiate between flat, rolling or mountainous terrain? And the answer is to be found in truck performance. If a truck can maintain a fairly steady speed, the topography is described as flat. If truck speed reductions by 15 to 20 km/h are frequent or of extended duration, the topography is described as rolling. Trucks operating at crawl speed either at frequent intervals or over extended distances cause the topography to qualify for the appellation 'mountainous'.

This is a somewhat circular definition because it is related to the speed at which trucks can move along the road and not directly to the topography itself. Given an unlimited budget, a road could be constructed to be dead flat regardless of the mountainous nature of the topography being traversed. The principal control is what constitutes a reasonable outlay for construction and maintenance cost plus road user cost.

7.2.3 Earthworks quantities and the mass-haul diagram

Construction cost is directly affected by the vertical alignment in relation to the natural topography as this defines the height of fills and the depth of cuts and hence the quantities of material that have to be moved. The calculation of quantities is invariably undertaken using commercially available software. Historically, it was a laborious process of plotting of cross-sections at some or other fixed interval and measuring of average end areas by planimeter. This, understandably, was not popular with young engineers. Designers should be aware of the intrinsic limits of accuracy due to variations of the terrain between surveyed cross-sections and variations from a truly prismoidal shape. The calculated volume may vary from the actual volume by 5 per cent to 10 per cent from these factors alone (Austroads, 2009a).

The mass-haul diagram shows the volume of material, either cut or fill, available at any point along the road. The standard notation is that cut represents a 'gain' of material for use, that is, positive, and fill a 'loss' of material, that is, negative. The mass-haul diagram shows the transition from net cut to net fill and hence the direction in which material needs to be moved indicating also the lead or haulage distance involved. Shortfalls in material indicate where additional material has to be sought and hence the possible location of borrow pits. These locations are required because it does not follow that suitable borrow material will always be available where desired. The haul distance from the borrow pits to the road will have to be added to the mass-haul diagram. Excess material requires that spoil areas be located. In rural areas, spoil material could perhaps be used in the construction of farm dams or other beneficial environmental features as compared to merely being waste tips.

A typical mass haul diagram is illustrated in Figure 7.1. It show the gradeline and, below it, the volumes of material to be moved and the direction of movement.

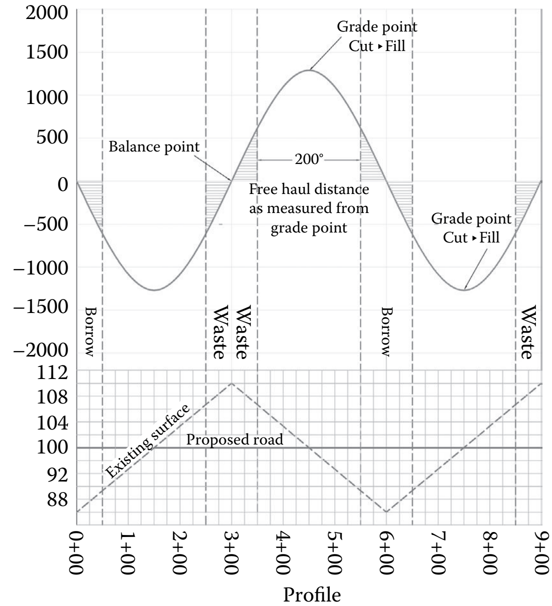


Figure 7.1 Mass-haul diagram. (From Mastergraphics, *Mass haul for the masses.* Civil 3D, Madison WI, 2008.)

Construction of the earthworks requires the provision of a stable base for the fill and the pavement in cuts. In the calculation of quantities, it is thus necessary to establish the quantities of material such as topsoil and other material not suitable for construction. This material should be removed as part of the site clearing exercise and carefully stockpiled for later use such as the trimming of cut slopes and the like. It will obviously not feature in the derivation of the mass-haul diagram. Factors that do affect the earthwork quantities include

- The compaction factor, that is, the ratio in volume between one cubic metre of *in situ* material and the same material after placement and compaction. Cohesive materials commonly occupy less volume after compaction, whereas rock may occupy more volume after excavation and placement.
- Material that is unsuitable for embankment construction including
  - Large boulders
  - Excavated hard rock which may be uneconomical to crush to a size that can be compacted
  - Any unstable or expansive material to be carted to waste
- Flattening of embankment batter slopes outside stability limits to provide for safety and maintenance requirements.
- Photogrammetric bias; it is preferable that ground survey be used for detailed design, in order to eliminate this factor and improve accuracy.

The quantities of cut include considerations of material type being hard rock, which requires blasting, soft rock, which can be ripped, and soft material, which can be picked up by a front-end loader without prior processing. The nature of the material has an obvious impact on the costs of excavation and subsequent compaction. As stated previously, when *in situ* material is disturbed and loaded, it bulks and, when placed in the fill, it is compacted. Bulking and compaction factors have to be estimated and allowed for in the achieving of a balance between cut and fill (Queensland Department of Transport and Main Roads, 2002). This information is presented on the mass-haul diagram.

Balancing the earthworks is often understood to be a volumetric balance. This is not entirely correct or even desirable. Drilling and blasting to achieve a hypothetical balance with adjacent fill volumes is significantly more expensive than raising the gradeline over the rock layer and opening a borrow pit where the shortfall in material occurs. On the other hand, it may be found that the rock layer is ideally suited for use as a crusher-run base course, chippings for bituminous surfacing or concrete aggregate. The gradeline is then adjusted accordingly but the additional material so gained is not included in the calculation of the mass haul diagram, which is exclusively related to the construction of the bulk cut and fill of the earthworks.

Side cut, where part of the cross-section is in cut (as opposed to a box cut) and part in fill, can lead to distortions of the mass-haul diagram. This is because it is theoretically, but only theoretically, possible to move material directly across the cross-section from the cut side to the fill side. About the only device that can achieve this operation is a drag-line scoop. This type of plant tends to be very large and confined to opencast mining. It is not likely to be found on a normal road construction site. A bulldozer acquires and spreads its load by moving approximately longitudinally down the road. At the end of its run, it has to turn around and repeat the operation in the reverse direction. This is for two reasons:

- The elimination of dead haul when the bulldozer is travelling without a load
- To even the wear on the steering clutches of the tractor

7.2.4 The critical length of grade

The speed reductions forced on trucks by the gradient of the road have safety implications as it has been found that the crash rate increases exponentially with increases in speed differentials. The safety slogan 'Speed kills' is not entirely correct; it is the speed differential that kills. It has been found that, with a speed reduction of 15 km/h or less, the crash rate is between 1 and 5 crashes per million kilometres of travel, whereas with a doubling of the speed differential, the crash rate increases to about 21 crashes per million kilometres.

As a general rule, a speed reduction of 15 km/h is sufficient to warrant the addition of a climbing lane to the cross-section. Reference is made to the critical length of grade, which is the distance that a truck will travel before its speed has reduced by 15 km/h. Climbing lanes are discussed in Chapter 8.

The relationship between speed and gradient is illustrated in Figure 7.2. This figure assumes a starting speed of 100 km/h. If the upgrade is preceded by a downgrade, truck drivers will accelerate to get the benefit of a higher entry speed. The vertical axis of the graph could thus be moved down to illustrate a higher entry speed, which should not, in any event, exceed about an additional 10 to 15 km/h. These speeds are hypothesized on a loaded truck with a mass to power ratio of 120 kg/kW, which may be considered to be the design vehicle for performance on a gradient.

7.2.5 Aesthetics

The profile cannot be designed in isolation. It has to be coordinated with the horizontal alignment to form a harmonious whole. In Chapter 9, reference is made to internal and external harmony. The coordination of the horizontal alignment and the vertical alignment forms part of the internal harmony, the abstract ribbon in space concept, which is sought. This is dealt with in detail in Chapter 9.

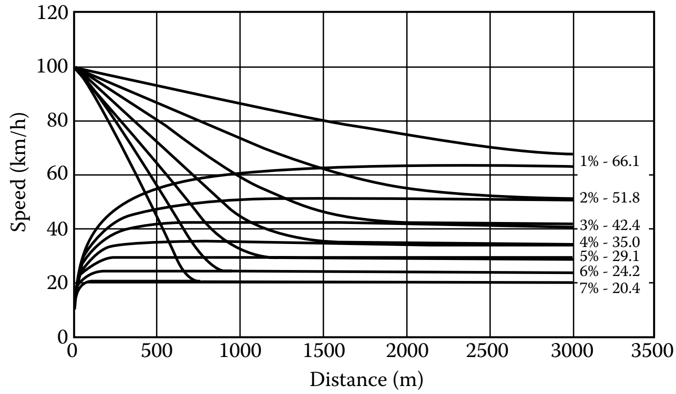


Figure 7.2 Truck speeds on grades. (From Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited [SANRAL], Pretoria, 2002.)

7.3 VERTICAL TIE POINTS AND CLEARANCES

The topography generates a host of tie points, defined as levels above which, through which or below which the profile must pass. These are

- Drainage constraints which are typical of minimum levels that the profile should exceed. These include flood levels or water tables. The design layers of the pavement lose strength when saturated. The bottom of the design layers, that is, the base course and sub-base, thus desirably should be about 500 millimetres above the 50-year flood line. With regard to clearance above the water table, geotechnical advice should be sought on capillary rise.
- Geotechnical considerations such as expansive clays, collapsing sands and saturated materials in swamps, which will also tend to provide minimum levels that the profile should exceed.
- Existing intersections and property entrances, which provide levels through which the profile should pass.
- Vertical clearances to overpasses, which dictate levels that the profile should not exceed and where a distinction is drawn between vehicular structures and lighter structures, such as pedestrian bridges, which are more prone to damage. As discussed in Chapter 5, clearances between the centreline of the road and the soffit of the deck above it are typically of the order of 5.2 metres and pedestrian bridges normally have a clearance of 5.7 metres.
- Underground service utilities and their protection requirements, which will also impact on the selection of profile level (Austroads, 2009a).

Underground services include

- Gas mains
- Water mains
- Storm water drains
- Sewerage
- Telecommunication, including cable TV, cables
- Underground electrical cables
- Road authority assets for ITS, traffic signals and street lighting

In each case, designers should consult with the relevant authority to determine the minimum clearances they require. With gas pipes, for example, clearances depend on the pressure within the pipe. The age of the service can also be relevant, particularly when lead joints have been used for water supply pipes. It is sometimes possible to obtain agreement to the use of a reduced clearance together with some form of special protection for the service such as encasement of the service in a box culvert, encasement in mass concrete or the provision of a concrete slab over the service.

Other vertical clearance considerations include

- Airport flight paths
- Grade separated railway crossings where a distinction is usually drawn between electrified lines and others
- Power lines, where a distinction is drawn between the various voltages of power supply
- Wildlife crossings
- Navigable waterways
- Irrigation canals

As illustrated, the lowest vertical clearance point need not necessarily fall under the centreline of the structure causing the obstruction. It is necessary to derive the point of minimum clearance on the basis of a matrix of points defining the horizontal and vertical alignments of the road and the bridge deck above it. As shown in Figure 7.3, the matrix could comprise as much as 25 points depending on the geometric complexity of the crossing in terms of

- The angle of skew between the two roads
- Their gradelines, including the presence of vertical curves
- The horizontal alignments of both roads, including the presence of horizontal curves and hence also superelevation
- Superelevation development on either or both roads

On the basis of most trucks having a height of 4.1 metres or less, AASHTO (2011) specified that the vertical clearance from the road surface to the soffit of the deck above it should

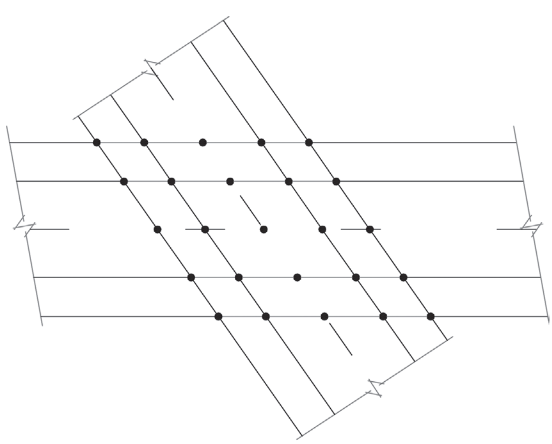


Figure 7.3 Matrix of vertical clearance points.

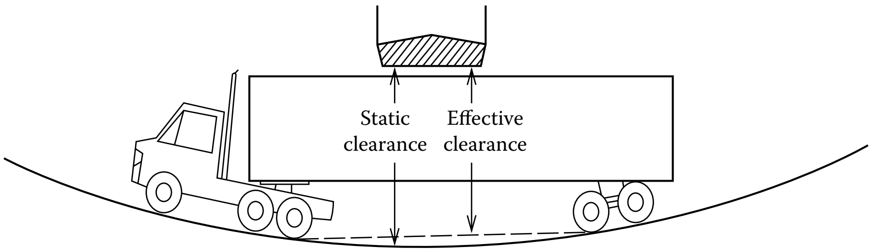


Figure 7.4 Clearance for long vehicles.

be not less than 4.3 metres. This did not allow much of a margin and the recommended clearance was subsequently increased to 4.9 metres (16 feet). A further allowance of 300 millimetres (1 foot) would make provision for future layers or resurfacing of the pavement arriving at a total clearance of 5.2 metres. Southern Africa and Australia/New Zealand also require a minimum clearance of 5.2 metres.

Lighter structures such as sign supports and pedestrian bridges should provide a vertical clearance that is at least 5.7 metres high, as required by British standards. This is because they are more likely to suffer structural failure as a result of being hit by a vehicle with the possibility of their falling on the road below.

The development of the gradeline is thus initiated by identifying these tie points and deciding whether they are discretionary or mandatory.

Where other responsible authorities are involved, the designer should obtain information on their requirements regarding these clearances.

Changes in the cross-section will have a bearing on the vertical clearance and must also be checked for compliance with the clearance standards of the road authority and affected asset owners. These changes include

- The crossfall or superelevation
- The position of the road crown
- The addition or removal of lanes
- The angle between the road and the object to be avoided

Vehicle dimensions may also play a role in the identification of tie points. As illustrated in Figure 7.4, longer vehicles passing under a structure on a sag curve may have less clearance actually provided them than the profile would suggest.

7.4 GRADES

7.4.1 Introduction

Grades have the parameters of length and gradient. Unlike the tangents of the horizontal alignment, grades can, theoretically, have any length down to zero metres. The length of zero arises from the fact that vertical curves can abut each other without having to make allowances for superelevation development. There is also no limit on the maximum length of a grade except that, as in the case of the horizontal alignment, a long grade would introduce the problem of dazzle over an extended distance. As gradients become steeper, operational constraints start to intrude, specifically with regard to the performance of heavy vehicles.

7.4.2 Maximum gradients

The higher the road is in the hierarchy of the road network, the flatter its maximum gradient should be. Streets in an urban township can be extremely steep and developers, whose sole aim is to create as many saleable properties as possible, will have no compunction in proposing gradients steeper than the ability of passenger cars to climb them.

San Francisco is well known for its steep streets. Lombard Street is internationally famous and best known for the one-way section on Russian Hill between Hyde and Leavenworth Streets, in which the roadway has eight sharp turns (or switchbacks) that have earned the street the distinction of being the most crooked street in the world. The switchback design, instituted in 1922, was born out of necessity to reduce the hill's natural 27 per cent grade, which, at that time, was too steep for most vehicles to climb. The crooked section of the street, which is about 1/4 mile (400 m) long and reserved for one-way traffic downhill, is illustrated in Figure 7.5.

It can be observed that parking on the section of Lombard Street below the switchback is at right angles to the centreline of the street to avoid the possibility of parking brakes not being able to hold vehicles on the downgrade.

The two steepest streets in San Francisco are Filbert Street, between Hyde and Leavenworth Streets, and 22nd Street, between Church and Vicksburg Streets. They have grades of 31.5 per cent. Both of these streets are one-way, going downhill.

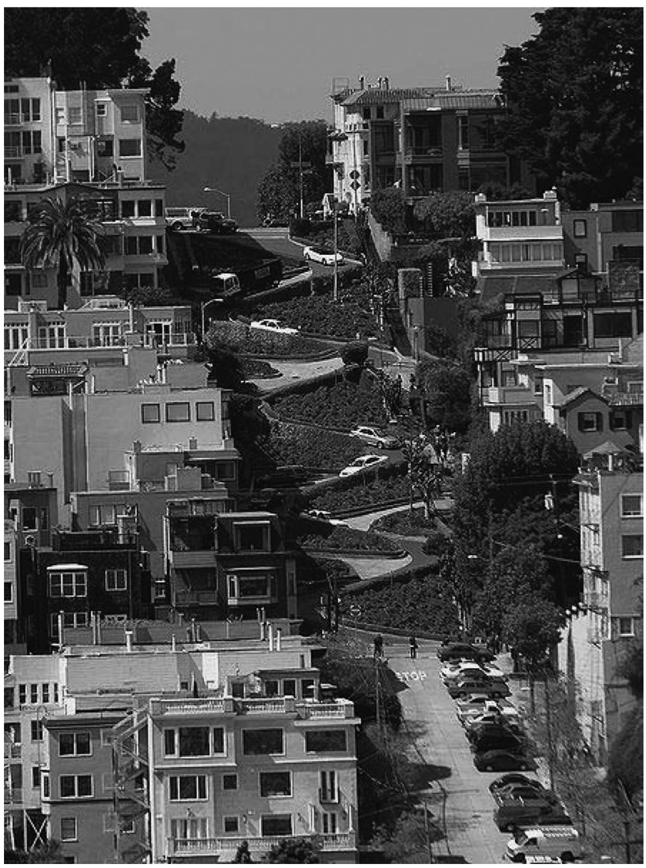


Figure 7.5 Lombard Street, San Francisco.

Some older streets in the United States as listed below have gradients steeper than 30 per cent (http://www.geographylists.com/list17y.html). These are

1. Honokaa-Waipio Road (near Waipio, HI, maximum grade 45 per cent)
2. Canton Avenue (between Coast and Hampshire, Pittsburgh, PA, 37 per cent)
3. 28th Street (between Gaffey and Peck, Los Angeles, CA, 33.3 per cent)
4. Eldred Street (west of Avenue 48, Los Angeles, CA, 33 per cent)
5. Baxter Street (between Alvarado and Allesandro, Los Angeles, CA, 32 per cent)
6. Fargo Street (between Alvarado and Allesandro, Los Angeles, CA, 32 per cent)
7. Maria Avenue (north of Chestnut, Spring Valley [near San Diego], CA, 32 per cent)
8. Dornbush Street (between Bricelyn and Vidette, Pittsburgh, PA, 31.98 per cent)

7.4.2.1 Desirable maximum gradients

Most urban street authorities now will not permit gradients steeper than 15 per cent. The United Kingdom specifies that any gradient steeper than 8 per cent will be considered to be a design exception requiring the issue of a waiver by the road authority (Highways Agency et al., 2002). Construction problems require that roads steeper than about 10 to 11 per cent should be concrete paved. Twelve-fourteen–tonne rollers are at their limits of hill-climbing ability on these gradients, so that attempts to compact the base course using these rollers would most likely be unsuccessful. They also tend to create the 'rumpled tablecloth effect', building corrugations into the road surface on stopping at the bottom of their run.

Operationally, braking on a steep downgrade on a bitumen surface will create the same effect. Truck drivers also experience significant difficulties in pulling away from rest on upgrades this steep. Typical maximum gradients are shown in Table 7.1.

In keeping with the philosophy of providing guidelines as opposed to rigid applied standards, it follows that there may be sound reason for exceeding these maximum gradients. These reasons may include

- Extremely adverse topography where gradients flatter than the maximum are not readily achievable without considerable outlay in construction costs
- Areas where short lengths of steeper gradient can achieve considerable cost savings
- Low volumes of truck traffic suggesting that an impediment to the free movement of passenger cars and other lighter vehicles is not an issue

The maximum gradients should not be adopted as a general rule and, preferably, should be used as sparingly as possible. If truck volumes exceed about 5 per cent, it may be useful to consider the provision of climbing lanes on steep gradients.

Table 7.1 Maximum gradients

Design speed (km/h)	*Gradients (%), by topography*		
	Flat	*Rolling*	*Mountainous*
60	6	7	8
80	5	6	7
100	4	5	6
120	3	4	5

Source: Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited (SANRAL), Pretoria, 2002.

7.4.2.2 *Safety on steep downgrades*

Steep upgrades mean steep downgrades in the opposite direction, and sensible truck drivers tend to essay these downgrades at about the same speed that they can achieve on the upgrade. Passing a slow-moving truck on a steep downgrade would normally not be a problem but, if passing opportunities are restricted, it may be worthwhile to consider a 'descending' lane. In addition, arrestor beds may be used to address the problem of runaway trucks. As a final safety measure, compulsory stops at the start of a long downgrade provide an opportunity to cool the brakes before the downgrade is essayed. A problem that truck drivers sometimes experience is a missed gear change whereby it is found that, having changed into neutral, it is not possible to engage the next gear. The truck then freewheels down the downgrade with a distinct possibility of loss of control and the resultant crash. The compulsory stop makes it possible to select the appropriate gear prior to the descent. Fortunately, the general adoption of automatic transmissions on trucks is phasing this problem out but there are still numerous trucks on the world's highways that have crash gearboxes, that is, transmissions without the benefit of synchromesh where the missed gear change is still a distinct possibility.

Speeds increase automatically on long downgrades and it is prudent to attempt to increase the value of geometric standards progressively to match the higher operating speeds likely to be found towards the bottom of the grade.

Steep downgrades can create problems with regard to storm water drainage in urban areas. Highly energised streams of water simply flow past drop inlets and ultimately flow across the surfaces of intersecting streets.

In rural areas, the combination of camber and a steep gradient may result in water flowing for a long distance down the road before achieving the shoulder. The possibility of vehicles hydroplaning is an ever-present danger.

Foul water sewers are usually provided within the road reserve and thus tend to have gradients similar to that of the road. On steep gradients, separation between fluid and solid waste can occur, leading to blockage of the drain.

7.4.2.3 *Maximum gradients on gravel roads*

Gravel roads are subject to scour so that gradients of more than about 6 per cent should be avoided. Where this is not possible, the grades should be paved. Paving should be provided at least from the point at which the gradient exceeds 6 percent to the point at which the gradient drops back to 6 per cent. Because the roughness coefficient of concrete is lower than that of gravel, flow down the paved section would be faster than that on a gravel surface with a gradient of 6 per cent. The downstream end of the paved section should thus be extended to a point where the gradient is such that the flow speed of the water would be that associated with a 6 per cent gradient on a gravel surface. As a further measure, the paved section should be provided with a camber or crossfall redirecting the flow of water to a side drain. The side drain should also be paved as a protection against scour on the steep gradients.

Roads are usually gravelled when traffic volumes are not high enough to warrant paving. Bitumen surfaces require the kneading action provided by traffic to maintain their flexibility. Insufficient traffic, for example, volumes of less than about 400 vehicles per day, thus suggests that paving should be concrete. It is possible that storm water flowing towards a paved section could undercut it, creating a dangerous situation. The upstream terminal of the paved section should thus be provided with a cutoff wall. Water flowing from the paved section onto the downstream gravel could be moving swiftly. Protection against scour will have to be provided either by directing the flow to the side of the road by the provision of a camber or straight crossfall or by extending the paved section to where the gradient is flat enough to cause the flow to be slower than that likely to cause scour.

7.4.3 Minimum gradients

If a road is not kerbed, the camber or crossfall can simply direct the flow of storm water over the edge of the fill or into a side drain in the case of a cut. The gradient could thus be 0 per cent, which is not uncommon in river plains and estuarine areas. If storm water is prevented from flowing over the side of the cross-section, that is, if it is trapped by the presence of kerbing, the gradient should be not less than 0.5 per cent. With stringent quality control of the levels of the road, this could be reduced to 0.3 per cent. Any attempt to ensure an adequate gradient flatter than this would be outside the limits of accuracy of construction. Furthermore, water speeds would be so low that there is a likelihood of water spreading across the road surface resulting in hydroplaning and vehicles sliding out of control. A final complication of very flat gradients is that self-cleansing velocities may not be achieved and the side drains could silt up. This could also have the effect of directing storm water onto the travelled way.

It is to be noted that the gradeline should not be designed in support of drainage but that the drainage must be adapted to suit the alignment of the road. However, a gradeline oscillating up and down at a gradient of 0.5 per cent is unsightly and unacceptable. It may, however, be possible to conceal the fluctuation in gradient by applying it at the start or end of a horizontal curve.

Flat gradients require particular attention in the presence of superelevation, as the superelevation development will result in the creation of a flat spot where water can pond. Water depths of as little as 10 millimetres can result in hydroplaning where the tread pattern on the tyres cannot remove the water from under the tyres fast enough, resulting in the bond between tyre and road being broken.

Successive curves in opposite directions can create a problem even on relatively steep gradients. It is possible that water flowing from the high side of the superelevation has not reached the low shoulder before the superelevation development in the opposing direction commences. The water then meanders down the road possibly building up sufficient depth to create hydroplaning. The pavement contours have to be checked to ensure that this condition does not arise. If it is likely, remedial measures may have to be considered such as changing the rate of superelevation development.

If the topography is such that the road has to have a gradient less than 0.5 per cent, channel grading can be applied. This applies to the channel defining the edge of the road. The channel edge abutting the lane is at the same level and gradient as the edge of the lane but the other edge is constructed at a slope of not less than 0.3 per cent. Drop inlets would have to be placed at very short intervals, for example, less than 50 metres. An average kerb height is 300 millimetres, with about half of it being buried in the base course layer so that the clear height is typically 150 to 200 millimetres. After a distance of only 50 metres, the depth of the channel at the kerb face will have doubled to be at the bottom of the kerb.

Where kerbs are not provided and the water is simply allowed to flow over the side of the fill, channel grading is also applied to side drains. This is at the high points of crest curves and the low points of sag curves because the gradient at these points is effectively flat, that is, 0 per cent. For example, a 120 km/h design speed and a change of gradient from +5 per cent to –5 per cent results in a gradient across the crest of the vertical curve flatter than 0.5 per cent for a distance of 110 metres or 55 metres on either side of the highest point on the curve. Starting from this point, with a side drain depth of 0 millimetres and a gradient of 0.5 per cent, the side drain would have a depth of 300 millimetres less the height difference generated by the vertical curve after a distance of 60 metres Typically, the base course and sub-base layers of the pavement have a thickness of about 150 millimetres each so that, at this distance, the bottom of the side drain would approximate the bottom of the sub-base layer. As discussed in Chapter 23, side drains should desirably not be deeper than 150 millimetres with an absolute maximum depth of 300 millimetres. This is predicated on grounds of road user safety. Deep side drains can cause out-of-control vehicles to dig into the far side of the drain and somersault end-over-end with possibly fatal consequences to the occupants of the vehicle.

In the case of a sag curve, a design speed of 120 km/h and a change of gradient from –5 per cent to +5 per cent would cause the gradient across the bottom of the sag curve to be less than 0.5 per cent for a distance of approximately 70 metres or 35 metres on either side of the low point of the curve. Over this distance, channel grading should be provided. Starting with a depth of 150 millimetres, a channel grading of 0.5% suggests that, at the lowest point on the curve, the channel could be about 170 millimetres deep plus the depth generated by the vertical curve. This is well within the absolute limit suggested in the preceding text.

It is unlikely that the bottom of a sag curve would be in cut. And, for reasons of drainage, every effort should be made to avoid this. About the only place where a sag curve below ground level would be forgivable would be in the case of an underpass where the higher of the two gradelines is closer to ground level than the required clearance plus allowance for the deck thickness. In this case, the side drain would have to discharge via a drop inlet into an underground drain. The drain would have to be extended to where it could discharge either into an existing watercourse or simply at ground level in the road reserve.

7.5 CURVES

7.5.1 Introduction

Vertical curves can be either convex or concave, the crest and the sag respectively. Four types of curves are identified

- A positive gradient followed by a flatter positive gradient, or a negative gradient, a crest
- A positive gradient followed by a steeper positive gradient, a sag
- A negative gradient followed by a steeper negative gradient, a crest or
- A negative gradient followed by a flatter negative or a positive gradient, a sag

The standard notation is that a positive gradient results in an increase in height in the direction of increasing stake value and a negative gradient results in a decrease in height in the direction of increasing stake value. There is no difference in the form of calculation of gradeline levels between the four forms of vertical curve. They are highlighted only to point out the need for care in the correct selection of arithmetic signs of the gradient of the successive grades being connected by the selected vertical curve.

Vertical curves usually take the form of parabolas. Some countries, notably European, use circular curves both for horizontal and for vertical curves but generally, the differences in levels between the two forms of curvature are sufficiently slight to be ignored.

Curves may be symmetrical or asymmetrical. Symmetrical curves have equal lengths on either side of the vertical point of intersection (VPI). They are the most usual form of vertical curve, often referred to as simple curves. The lengths of the vertical curve on either side of the VPI do not, however, have to be equal. The shape of the topography, the land form, may, for example, require that a gentle curve be immediately followed by a sharper curve in the same direction, either convex or concave. In effect, a compound curve results. The two sides of the asymmetrical curve are calculated separately. It is necessary to ensure that the sight distance provided is adequate on both sections of the asymmetrical curve.

7.5.2 The properties of parabolic curves

The outstanding property of a parabola is that it provides a constant rate of change of gradient. In consequence, the extent of sight distance provided remains constant across the length of the curve. The basic formula of a parabola is

$$y = ax^2 + bx + c$$

where
- y = height of any point on the curve above the x-axis
- x = distance from the beginning of the vertical curve (BVC)
- a = rate of change of gradient
- b = gradient at BVC
- c = height of centreline at the BVC

The gradient is provided by the first differential

$$\frac{dy}{dx} = 2ax + b$$

and the rate of change of gradient is

$$\frac{d_2y}{dx^2} = 2a$$

This constant, which is the determinant of the shape of the parabola, is reformulated as K, the distance required to effect a 1 per cent change of gradient. It follows, with $K = 1/2a$, that the length of a parabola, L, is expressed as

$$L = AK$$

Calculation of gradeline levels on a parabola can be by one of two methods:

- The offset method
- The gradient method

The relationship for the parabola can be rewritten as

$$y = \frac{1}{2K}x^2 + \frac{G}{100}x + H_{\text{BVC}}$$

The latter two terms in the relationship give the elevation of the grade when extended beyond the beginning of the vertical curve (BVC) by a distance, x, and the first term is the offset from the grade level, with K being negative for a crest curve and positive for a sag curve, that is, the offset is below or above the grade level respectively. Using this method,

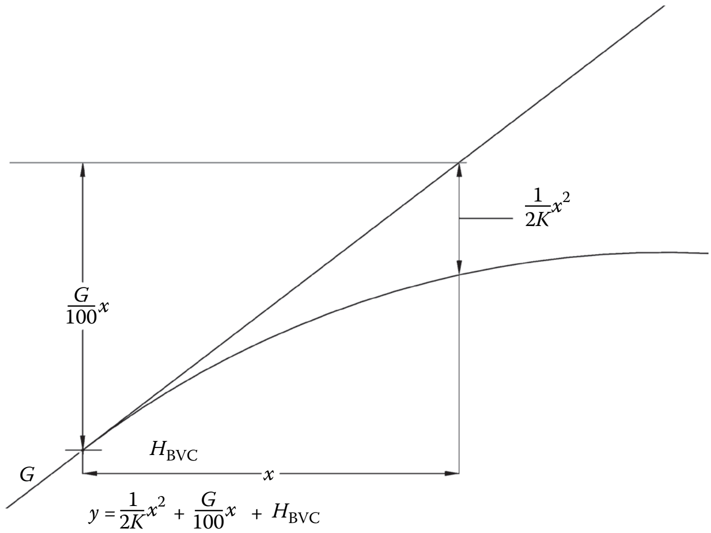


Figure 7.6 The offset method for calculation of gradeline levels on vertical curves.

the gradeline level can be calculated at any point, x, on the vertical curve. As illustrated in Figure 7.6, it is a very convenient form of calculation and is the recommended option.

To illustrate the offset method: The vertical point of intersection (VPI) of a crest curve is located at stake value (SV) 26 + 540 at an elevation of 1526.53 metres. The approach grade has a gradient of $G_1 = 4.03$ per cent and the departure gradient is $G_2 = -3.52$ per cent. The designer has elected to adopt a vertical curve with a K-value of 90. The height of the gradeline is to be calculated at 20-metre intervals along the curve.

Calculation. The curve length is

$$L = AK$$

$$= (4.03 + 3.52) \cdot 90$$

$$= 679.5 \text{ metres}$$

For convenience the curve length is changed to 680 metres so that the revised value of K used in the calculation becomes

$$K = L/A$$

$$= 680/(4.03 + 3.52)$$

$$= 90.06623$$

The location and elevations of the beginning (BVC) and end (EVC) of the vertical curve are

$$\text{Location}_{BVC} = SV_{VPI} - L/2$$

$$= (26 + 540) - 340$$

$$= 26 + 200$$

$$\text{Elevation}_{\text{BVC}} = H_{\text{VPI}} - G_1/100 \cdot L/2$$

$$= 1526.53 - 4.03/100 \times 340$$

$$= 1522.828$$

$$\text{Location}_{\text{EVC}} = SV_{\text{VPI}} + L/2$$

$$= (26 + 540) + 340$$

$$= 26 + 880$$

$$\text{Elevation}_{\text{EVC}} = H_{\text{VPI}} - G_2/100 \cdot L/2$$

$$= 1526.53 - 3.52/100 \times 340$$

$$= 1514.562$$

The calculation of the gradeline levels is illustrated in Table 7.2.

Using a spreadsheet, it is not necessary to calculate the elevation of the extension of the grade separately. It is offered purely for illustrative purposes. A single calculation run will suffice.

The Gradient method is quite cumbersome as it requires two runs through the vertical curve to establish the gradeline levels. For convenience, calculation is limited to that of elevations at set intervals. Its only advantage is that it is self-checking. This is because

1. The first run calculates the gradient between successive points along the curve. Starting with the gradient at the BVC, gradients between the various points are calculated. The gradient at the end of the curve (EVC) must equal the gradient between the Vertical Point of Intersection (VPI) and the EVC. The first gradient used is the gradient of the approach grade plus or minus the extent of the change of gradient for a distance which is half that of the distance between successive stake values. In the case of a staking distance of 20 metres, the first gradient calculated is midway, that is, at 10 metres, between the BVC and the first Stake Value thereafter. This is the average gradient

Table 7.2 Gradeline levels according to the offset method

Stake value	*Grade level = g/100x +* H_{BVC}	*Offset*	*Grade line level*
26,200 (BVC)	1512.828	0	1512.828
26,220	1513.634	0.022	1513.612
26,240	1514.44	0.089	1514.351
26,260	1515.246	0.200	1515.046
26,280	1516.052	0.355	1515.697
26,300	1516.858	0.555	1516.303
26,780	1536.202	18.675	1517.527
26,800	1537.008	19.985	1517.023
26,820	1537.814	21.340	1516.474
26,840	1538.62	22.739	1515.881
26,860	1539.426	24.182	1515.244
26,880 (EVC)	1540.232	25.670	1514.562

between the two points on the vertical curve. The average gradient between this and the following stake value is then calculated by adding or subtracting the change of gradient for a distance from 20 metres from the preceding gradient. The gradient at the EVC is calculated by adding or subtracting the change of gradient for a distance of 10 metres from the preceding calculated gradient.

2. The second run commences with the height, H_{BVC}, and calculates the heights of the various points along the curve. The height at the end of the vertical curve, H_{EVC}, must equal the height calculated from the height of the VPI, the second gradient of the curve and the distance from the VPI to the EVC.

7.5.3 Relationship of curve length to sight distance

Two conditions arise being where the curve is either longer or shorter than the required sight distance. In the first case, the K value is given by the relationship

$$K = \frac{S^2}{200(h_1^{0.5} + h_2^{0.5})^2}$$

In the second case, the relationship is

$$K = \frac{2S}{A} - \frac{200(h_1^{0.5} + h_2^{0.5})^2}{A^2}$$

where

S = stopping sight distance for selected design speed and object height (m)
h_1 = driver eye height (m)
h_2 = object height (m)
A = Algebraic difference in gradient between the approaching and departing grades

The first of the two relationships offered above is more convenient to calculate than the second and using it also in the case of the sight distance being longer than the curve does not result in significant errors (AASHTO, 2011a). Stopping sight distance and object height are discussed in Chapter 5 and driver eye height in Chapter 4.

7.5.4 Minimum values of *K* for crest curves

Minimum values of K for crest curves for various object heights are shown in Table 7.3. These values, based on Stopping Sight Distance, are generally adequate in terms of comfort and appearance (SANRAL, 2002).

7.5.5 Minimum values of *K* for sag curves

The K-values of sag curves could be based on any one of four criteria:

- Sight distance
- Passenger comfort
- Drainage
- Vertical obstructions

Table 7.3 Minimum *K* values for crest curves

Design speed (km/h)	*Stopping sight distance (m)*	K-*value for object height (m) equal to*		
		0.00	*0.15*	*0.60*
40	50	12	6	4
50	70	25	12	8
60	90	40	20	12
70	110	60	30	18
80	140	90	50	30
90	170	140	70	45
100	200	190	100	60
110	230	250	130	80
120	270	350	180	110
130	310	460	240	150

Source: Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited (SANRAL), Pretoria, 2002.

During the hours of daylight or on well-lit streets at night, sag curves do not present any sight distance problems. However, when the vehicle headlights are the only source of illumination, the line of sight from the driver's eye height has to be replaced by a line originating from the headlight. Generally, the mounting height of headlights on passenger cars is about 600 millimetres above the road surface. A line parallel to the central axis of the vehicle and commencing at this height defines the centre of the beam of light. The light beam diverges, however, and it has been widely accepted that the angle of divergence is of the order of 1°. The value of *K* providing the required sight distance, *S* (Austroads, 2009a) is

$$K = \frac{S^2}{200(0.6 + S\tan 1°)}$$

Minimum values of *K* for sag curves calculated on this basis are given in Table 7.4.

Table 7.4 Minimum values of *K* for sag curves

Design speed (km/h)	K *value*
40	8
50	13
60	19
70	24
80	32
90	40
100	49
110	57
120	68
130	80

Source: Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited (SANRAL), Pretoria, 2002.

Low values of K result in discomfort because the g-forces and centrifugal forces are working in concert. This is similar to the forces being inflicted on passengers on a roller coaster. It is generally accepted that values of vertical acceleration should not exceed 0.3 m/s^2. This requirement is met when the K-value is approximately half of that required in terms of headlight sight distance.

Drainage in the case of sag curves has been discussed in Section 7.4.2. From that discussion it follows that the lower the K-value of the sag curve, the shorter the distance is over which channel grading has to be applied. This K-value is thus a maximum whereas the K values for comfort and headlight sight distance are minima.

As discussed in Chapter 5, overhead obstructions such as a bridge deck crossing over the road can result in a loss of sight distance. In other cases, the additional eye height afforded the truck driver is a benefit but, in the case of sight distance under a structure, it constitutes a worst case. The driver eye height is thus taken as being 2.0 metres and the object height is usually selected as 0.60 metre. The sight distance under the structure should at least be stopping sight distance or greater.

7.5.6 Minimum length of vertical curves

It is customary to provide curves at all changes of gradient. However, if the change in gradient is slight, high values of K would have to be provided to ensure a satisfactory appearance of the vertical curve. It is thus acceptable to effect the change in gradient without employing a vertical curve. Table 7.5 illustrates grade changes where vertical curves can be dispensed with and also the minimum length of vertical curves required to ensure a satisfactory appearance (Queensland Department of Transport and Main Roads, 2002).

7.5.7 Curve combinations

As in the case of horizontal curvature, vertical curves can be remote from each other so that there is no aesthetic interaction between them or close where this interaction may occur. Combinations of curvature include the

- Reverse curve
- Broken-back curve
- Compound curve
- Hidden dip
- Roller coaster

A reverse curve comprises a crest curve followed by a sag curve or vice versa. If a sag curve abuts a crest curve, it will appear as though the road drops away sharply in front

Table 7.5 Appearance criteria

Design speed (km/h)	*Maximum gradient change without vertical curve (%)*	*Minimum length of vertical curve for satisfactory appearance (m)*
40	1.0	30
60	0.8	50
80	0.6	80
100	0.4	100
120	0.2	150

of the vehicle. Where a crest curve abuts a sag curve, the appearance is of a lump that has to be climbed over. The visual appearance of a reverse curve is improved by a short length of grade as a transition between the two curves. This will create the impression of a smoothly curving vertical alignment. A length of grade in metres that is equal to about half the design speed in kilometres per hour is all that is required. A design speed of 100 km/h would thus suggest that the grade between successive curves should be about 50 metres long.

The broken-back curve arises when a sag curve is followed by another sag curve. Successive crest curves do not usually result in a broken-back curve because the following curve is usually out of sight for a driver on the first curve. Generally, a distance of 500 metres or more between the sag curves would result in the curves not interacting with each other, that is, they would not appear to be broken-backed. Some authorities (Austroads) are prepared to accept a distance in metres of as low as $0.4 \times V$ (where V is in km/h) as being sufficient to avoid the appearance of a broken-back curve. Personal experience suggests that this is too close a spacing to avoid the appearance of a broken-back.

The compound curve arises from a combination of two curves, both either crest or sag, with different K-values. The change in K-value occurs above or below the VPI and the two curves are calculated separately.

The hidden dip is an unfortunate combination of the vertical and the horizontal alignment. It typically comprises a crest curve followed by a sag, a grade, a sag and a crest in that order. The grade on the far side of the hidden dip is often at about the same elevation and gradient as the nearside grade, creating the illusion of continuity of the alignment. In consequence, the driver may be unaware of an approaching vehicle and attempt an overtaking manoeuvre with potentially disastrous consequences.

The roller coaster arises from a succession of short crests and sags on a horizontal tangent. This form of alignment typically arises in rolling topography, with the road being located approximately at right angles to the contours as illustrated in Figure 7.7.

Figure 7.7 Typical roller coaster alignment.

7.6 GRADELINE DEVELOPMENT

The longitudinal section contains virtually all the information necessary to build the road. At the top of the sheet, the profile is presented. This includes the ground line, typically under the centreline of the road, and the gradeline. Because of its application in the development of the mass-haul curve, the gradeline represents the top of the earthworks, hence, as previously stated, often referred to as the subgrade. The data defining the vertical curves are normally shown on the profile. These data are

- The heights of the VPIs
- The length of curves
- The entry and exit gradients
- The *K*-value of the curves

The longitudinal section is drawn to a distorted scale, usually of 10*V*:1*H*. As gradients are very low, typically 6 per cent or less, it is very difficult to create a mental picture of the gradeline without this distortion. While drawing the gradeline is usually by the use of software, it is useful to have some idea of the mechanics of drawing it manually.

Plastic templates of different radii, known as ship or railway curves, are used to construct the vertical curves. These templates are, in fact, circular but, because of the distortion, are a reasonable approximation of the parabola.

As a rule of thumb, the radius of the railway curve in centimetres is equal to the *K*-value of the vertical curve when the distortion is 1:10. This applies to the middle third of the curve. The outside thirds can be approximated by using a longer radius railway curve between the plotted position of the BVC or EVC and the already drawn section of the curve. The outside curves can have anything up to double the radius of the inner curve.

It sometimes happens that the curve being drawn is so long that it cannot be drawn with a single template. A useful device then is to draw a line between the midpoints of the tangents connecting the BVC and EVC to the *PI*. The vertical curve falls on the midpoint of this line, which is also tangential to the curve. The process can be repeated by a series of halving of the tangent lengths and joining these midpoints. The curve will be tangential to the joining line at the midpoints of these lines.

An extract of the survey plan is typically shown below the longitudinal section as a strip plan. Data defining the horizontal alignment are shown on this survey plan. If space permits, the mass-haul diagram is shown below the survey plan.

At the bottom of the sheet, a table is provided containing the following data:

- The ground levels
- The gradeline levels
- The bearing of tangents and data defining the horizontal curvature
- The superelevation and location of the superelevation development
- The location and sizes of culverts
- The location of guardrails, barrier lines and so forth

These data are plotted against a baseline of increasing stake value.

1. The first step in the development of the gradeline is to add all the known information to the drawing, which consists of
   - Ground line and the associated levels added to the table
   - Survey plan

- Horizontal alignment data
- The location and height of the control points with the height of the control points also being listed in the table

2. The next step is selection of the grading point. This is the point on the cross-section, through which the gradeline passes. In the case of a two-lane road, the grading point usually is at the theoretical intersection point of the two slopes comprising the camber of the road. It is also the point around which the superelevation is developed.

   On dual-carriageway roads, the grading point can either be on the centreline of the cross-section or there may be two grading points with these located at the inner edge of the inside shoulder. This would support the development of split grading between the two carriageways. These grading points could be used in the case also of stage construction of a dual carriageway. In the case of there being two grading points, the ground line is that lying under the centreline of the road.
3. Prepare a trial gradeline, taking into account the vertical controls including culverts, and include coordination of horizontal and vertical alignments as far as practicable.
4. Calculate earthworks quantities and develop a mass-haul diagram.
5. Adjust the vertical alignment so that
   - All mandatory controls are met
   - Discretionary controls are met as far as possible
   - Other controlling criteria are satisfied with special consideration given to the location of intersections and points of access to ensure that minimum sight distances and critical crossfall controls are met
   - Earthworks are minimized

Where minimum standards cannot be achieved and compromises have to be made, the designer requires a broad understanding of basic theory and the assumptions made in the development of the standards. All design exceptions should be documented in terms of the process of design exceptions, variances and waivers discussed in Chapter 4.

Chapter 8

Cross-section design

CONTENTS

8.1 INTRODUCTION

AASHTO (2011a) defines the cross-section as 'A vertical section of the ground and roadway at right angles to the centreline of the roadway, including all elements of a highway or street from right-of-way line to right-of-way line'. Although many designers are happy to spend considerable periods of time on the horizontal and vertical alignments, the cross-section seems to be in the nature of an afterthought. Other than to accept whatever cross-section the road agency deems to be appropriate for the project, it seems to command little attention or thought. And yet this Cinderella is used to define the very nature of the road being designed. The two-lane two-way road, the 2 + 1, the six-lane dual-carriageway are nothing more or less than names describing the cross-sections of these roads.

Another feature of the cross-section is that, while the horizontal and vertical alignments require considerable calculation, the design of the cross-section is essentially a process of selection and engineering judgment. A belief common to all property developers and, in fact, shared also with many town planners, is that the cross-section is whole and indivisible and spans the full width of the road reserve (in American nomenclature the right-of-way). In fact, it comprises many discrete components, each with its own functions, location within the road reserve and dimensions. The design of the cross-section is thus a process of selection of the components required followed by their sizing.

The components which appear in a cross-section include

- The lanes
  - Basic
  - Auxiliary
- Shoulders and sidewalks
- Medians
- Outer separators
- Boulevards
- Cut and fill slopes
- Verges
- Drainage elements
- Provision for utilities
- Roadside amenities, including
  - Rest areas
  - Trading areas
  - Traveller facilities

Not all of these components appear in every cross-section. The designer is required to determine precisely what function the road is intended to serve and then to select the components appropriate to that function. The next step is concerned with locating these components in the road reserve and, finally, sizing the selected components to properly give effect to their individual functions.

These components are discussed in the following sections of this chapter.

8.2 BASIC LANES

The traffic lane is that part of the cross-section that is set aside for the one-way movement of a single stream of vehicles.

The nomenclature adopted in reference to the various lanes in the cross-section of dual carriageways can be confusing. Reference is made in various texts and manuals to the inner and the outer lane, where the inner lane may or may not be the one closest to the outer shoulder. United Kingdom references (Highways Agency et al., 2002) are often to the nearside and the far side lanes. These refer to the side from which one normally mounts a horse, to wit from the left. In the United Kingdom, to a person standing on the left side outer shoulder, the nearside lane is thus the one closest to him or her. In this text, reference throughout will be to refer to the lanes by number, with lane 1 being the one on the extreme right-hand side when looking in the direction of increasing stake value.

The number and width of traffic lanes have a significant influence on the safety, capacity and comfort of driving. The width of traffic lanes may also impact on the operating speed of the road. The number and width of traffic lanes usually depend on (Barton, 2009)

- Traffic volumes
- Traffic composition in terms of number of trucks and including the presence or otherwise of cyclists
- Available road reserve width
- Side friction generated by abutting accesses and adjacent lanes including for parking

Driver expectations would also guide the selection of lane width. Long distance travel is normally undertaken at high speeds for which drivers would expect to be provided with wide lanes. On the other hand, on urban streets providing access to single residential dwellings, low speeds would prevail and narrow lanes would be acceptable.

Previously, where imperial units of measurement were applied, lanes typically were 9, 10, 11 or 12 feet wide. When the United States started moving towards metric units, these were converted to 2.7, 3.0, 3.3 and 3.6 metres respectively. For the 12-foot-wide lane, the United Kingdom has adopted a width of 3.65 metres although, for six-lane cross-sections, lane 2 and, in some cases, lane 3 is 3.7 metres wide and, for eight-lane cross-sections, lanes 2 and 3 are 3.7 metres wide. Southern Africa has adopted 3.7 metres as the equivalent of the previous 12-foot lane width. Australia and New Zealand have adopted a 3.5 metre width as being the standard for all roads. Lane widths adopted by other countries are listed in Table 8.1.

As shown in Table 8.1, lane widths are typically predicated on the function of the road. As discussed in Chapter 3, urban roads may be

- Freeways
- Urban arterials
- Collectors
- Local access streets serving
  - Residential areas
  - Commercial areas
  - Industrial areas

The lanes on freeways and urban arterials are the widest, typically in the range of 3.5 to 3.7 metres wide. Collectors, which would be the major roads in local areas, would often serve as bus routes and their lanes should thus be of the order of 3.3 to 3.5 metres wide (or

Table 8.1 International lane widths

	Lane widths (m) by roadway classification		
Country	*Freeway*	*Arterial*	*Minor or local*
Brazil	3.75	3.75	3.0
Canada		3.0–3.7 Rural	3.0–3.3
China	3.75	3.5–3.75	3.5
Czech Republic	3.5–3.75	3.0–3.5	3.0
Denmark	3.5	3.0	3.0–3.25
France	3.5	3.5	3.5
Germany	3.5–3.75	3.25–3.5	2.75–3.25
Greece	3.5–3.75	3.25–3.75	2.75–3.25
Hungary	3.75	3.5	3.0–3.5
Indonesia	3.5–3.75	3.25–3.5	2.75–3.0
Israel	3.75	3.6	3.0–3.3
Japan	3.5–3.75	3.25–3.5	3.0–3.25
Netherlands	3.5	2.75–3.25	3.10–3.25
Poland	3.5–3.75	3.0–3.5	2.5–3.0
Portugal	3.75	3.75	3.0
South Africa	3.7	3.0–3.7	2.25–3.0
Spain	3.5–3.75	3.0–3.5	3.0–3.25
Switzerland	3.75–4.0	3.45–3.75	3.15–3.65
United Kingdom	3.65	3.65	3.0–3.65
USA	3.6	3.3–3.6	2.7–3.6
Venezuela	3.6	3.6	3.0–3.3
Yugoslavia	3.5–3.475	3.0–3.25	2.75–3.0

Source: Hall LE et al., *Overview of cross-section design elements.* Transportation Research Board, International Symposium on Highway Geometric Design Practices, Boston, 1995.

wider if space and cost permit). The local streets in commercial and industrial areas would have to be able to accommodate heavy goods vehicles with lanes also of a width of 3.3 to 3.5 metres. Lanes on local residential streets need not be more than about 3.0 metres wide and, if serving relatively few houses, could be as little as 2.7 metres wide. In the extreme case, local residential streets could be only one lane wide but provided with passing bays in support of flow in both directions (Wolhuter, 2000).

Rural roads connecting towns usually have two-lane two-way cross-sections. Speeds on these roads would be high and lane widths should thus be similar to those of freeways. Secondary roads connecting local areas to the primary network would carry relatively little traffic and be short. A lane width similar to that of the urban collector would thus suffice. Farm roads usually connect one or two farms across a third to the secondary road network and are often little more than single-lane tracks.

A distinction is drawn between basic lanes and auxiliary lanes. Basic lanes are those that are continuous over extended lengths of a road. However, the number of basic lanes can change across the length of the road. For example, in moving from a rural area to the centre of an urban area, the volume of traffic to be accommodated will increase. A two-lane two-way road could thus, at a certain point, change to having four lanes and thereafter become a four-lane dual carriageway or even a six-lane dual carriageway.

8.3 AUXILIARY LANES

8.3.1 Introduction

Auxiliary lanes are those that are added to the cross-section to serve a short-term need and are dropped as soon as the need has been met. Auxiliary lanes include

- Climbing lanes
- Passing lanes
- Turning lanes
- Parking lanes
- Additional through lanes between intersections
- Additional through lanes between interchanges

These lanes are discussed further below.

8.3.2 Climbing lanes

As traffic volumes increase, so does the need for passing opportunities. Slow-moving vehicles cause deterioration in the traffic flow and hence a reduction in the level of service (LOS) provided. Acceleration to an acceptable passing speed on steep gradients takes longer than on a level grade, exacerbating the difficulty of the manoeuvre. The provision of climbing lanes will increase the length of passing opportunities and thus support matching of the LOS on the upgrade to that of the flatter sections on either side of the upgrade. These lanes are also referred to as crawler lanes, truck lanes or passing lanes. The last-mentioned name is incorrect because, although a passing lane looks like a climbing lane, its function is totally different, as described in the following section.

As discussed in Chapter 7, a speed reduction of 15 km/h by slower moving vehicles is considered sufficient to warrant the provision of a climbing lane. On multi-lane roads, there are generally sufficient opportunities for passing but, if there is very heavy traffic on the road, a climbing lane may be warranted if the speed reduction is of the order of 25 km/h (Transit New Zealand, 2002).

Steep downgrades can also have a detrimental effect on the capacity and safety of a road with high traffic volumes and numerous heavy vehicles. Heavy vehicles descending steep downgrades at a crawl speed have an effect on LOS similar to that on the equivalent upgrade. However, acceleration of the passing vehicle is much easier on the downgrade than on the upgrade so that passing is less of a problem. In very hilly or mountainous areas, the horizontal alignment may include short radius curves at frequent intervals. The lack of sight distance would reduce passing opportunities, thus further warranting the use of climbing lanes, although these could have more of the function of passing lanes, as discussed in the following section.

A slow-moving vehicle should be fully on the climbing lane by the time its speed has dropped by 15 km/h. This can be determined by reference to Figure 7.2, which shows the distance required on an upgrade to cause a reduction of speed of 15 km/h starting from an entry speed of 100 km/h. It should remain on the climbing lane until its speed has increased again to 85 km/h.

Entrance to and departure from climbing lanes is eased by the provision of tapers. The lane addition taper is generally referred to as a passive taper, because the driver has the choice of changing lanes or not. The lane drop taper is referred to as an active taper because there is no real choice in using it. The alternative is to end up on the shoulder or on the median island.

The active taper should have a taper rate of about 1:50. Passive tapers are typically in the range of about 2 to 5 degrees, and a taper rate of 1:20 is approximately in the middle of this range. The more generous active taper allows for the complexity of the merging manoeuvre where the driver of the vehicle exiting from the climbing lane has to locate a gap in the through traffic and then match both the location and the speed of his or her vehicle to this gap. In addition to forward sight distance, the driver of the merging vehicle must thus also have sight distance to the rear.

Clear sight distance must be available also to the driver on the through lane. This driver must be able to see the road markings defining the taper so that the intended path of the merging vehicle will be clear. For this reason also, the climbing lane should extend beyond the end of the crest curve to ensure that the entire taper would be visible. Stopping sight distance must be replaced by decision sight distance, which is measured to the road surface.

In general, climbing lanes should be at least 800 metres long to allow for completion of the passing manoeuvre by more than just one vehicle.

The width of climbing lanes usually is the same as that of the through lanes, which on a rural road would typically be of the order of 3.6 to 3.7 metres wide. If space is limited, reducing the width of the climbing lane to 3.3 metres or perhaps even 3.0 metres would be acceptable because of the slow speeds of vehicles on the climbing lane. For the same reason, the shoulder width could also be reduced, perhaps to as little as 1.5 metres.

On mountain passes, the vertical alignment often comprises long distances with a fairly constant gradient typically with a side cut cross-section. Providing a climbing lane over the full length of the upgrade would invariably prove to be uneconomic because, particularly in mountainous terrain, moving the centreline by as little as 1 metre to one side or the other to accommodate the climbing lane could cause the fill slope either to extend almost as far as the valley below the pass or result in a cut slope almost to the top of the mountain. The cost of the earthworks would become prohibitively expensive.

As a subset of climbing lanes, recourse may be had to partial climbing lanes (PCLs), also referred to in the literature as passing bays, slow vehicle bays or turnouts (New Zealand Transit, 2002). Because of the low speeds of traffic likely to use PCLs, they seldom need to be wider than about 3.3 metres. They should not be longer than about 250 to 300 metres. This is to ensure that drivers wishing to overtake a slow vehicle should be able to see the entire length of the PCL and not confuse it with a full-length climbing lane. It would also support a decision whether the passing manoeuvre should be undertaken or aborted if already initiated.

The provision of PCLs is appropriate on winding two-lane two-way rural roads in mountainous, coastal and scenic areas where normal traffic flows are low, typically with average daily traffic (ADT) less than 2000 vehicles per day and there is a high proportion of slow-moving vehicles in the traffic stream. Furthermore, proposals to provide PCLs must include careful consideration of the need to minimise queue lengths, minimum sight distance requirements and the relative costs/benefits of providing other types of passing opportunities.

A single PCL would not, however, be all that useful because, as soon as the stopped vehicle moved out to join the main traffic stream, a queue would develop once more and impatient drivers would start to undertake passing manoeuvres even where sight distance is inadequate. Numerous PCLs should be provided and it is necessary to identify all the places where it is possible to build them economically and then to select those that are sufficiently spaced to disperse queues.

The first PCL, that is, at the bottom of the grade, should have the geometry of the full climbing lane in the sense that the lane should commence at the point at which the speed of the slow-moving vehicle has dropped by 15 km/h and have a passive taper upstream of this point of about 1:20. The passive tapers of the other partial climbing lanes would be negotiated only at a crawl speed and a taper rate of 1:10 to 1:12 would be adequate. Also, being traversed at a crawl speed and starting from a dead stop, the active tapers for the lane drops on the first and the subsequent PCLs could have a similar taper rate.

Where PCLs are located in the vicinity of crest curves, they should be located such that they end in advance of the crests, or well beyond the end of the vertical curve on the down sides of hills. It is essential that they do not end just beyond blind horizontal curves or just over the crests of hills, that is, where drivers do not have an adequate view of a slow-moving vehicle reentering the traffic lane immediately ahead of them. Similarly, drivers of vehicles merging from PCLs should have a clear view of vehicles approaching from behind.

8.3.3 Passing lanes

As discussed previously, the provision of a climbing lane is intended to match the level of service on an upgrade with that of the flatter grades on either side of it. The passing lane, on the other hand, is intended to improve the operational efficiency of the road as a whole by increasing its capacity and hence the LOS experienced at various flow levels and also reducing delays resulting from inadequate passing opportunities (Potts and Harwood, 2004). A typical passing lane is illustrated in Figure 8.1.

Perhaps the extreme example of this is the four-lane road, which could perhaps be described as a two lane road with continuous passing lanes on either side. It achieves this improvement in LOS by dispersing platoons of vehicles trailing slower moving vehicles.

Passing lanes also increase the safety of the road by providing passing opportunities without vehicles having to enter the opposing lane. Various configurations of passing lanes can be considered, ranging from the isolated passing lane to continuously alternating passing lanes as shown in Figure 8.2. The latter is typified by the 2 + 1 cross-section where the two lane sections continuously alternate between the two directions of flow, typically at 1.5 to 2.5 kilometre intervals. The only difference between a two-lane road with passing lanes and the 2 + 1 cross-section is the lane that is added and dropped. For the two-lane road with passing lanes and, for driving on the right, it is the rightmost lane that is added and dropped.

Figure 8.1 A typical passing lane. (From Potts IB and Harwood DW. *Benefits and design/location criteria for passing lanes*. Missouri Department of Transportation, Kansas City, 2004.)

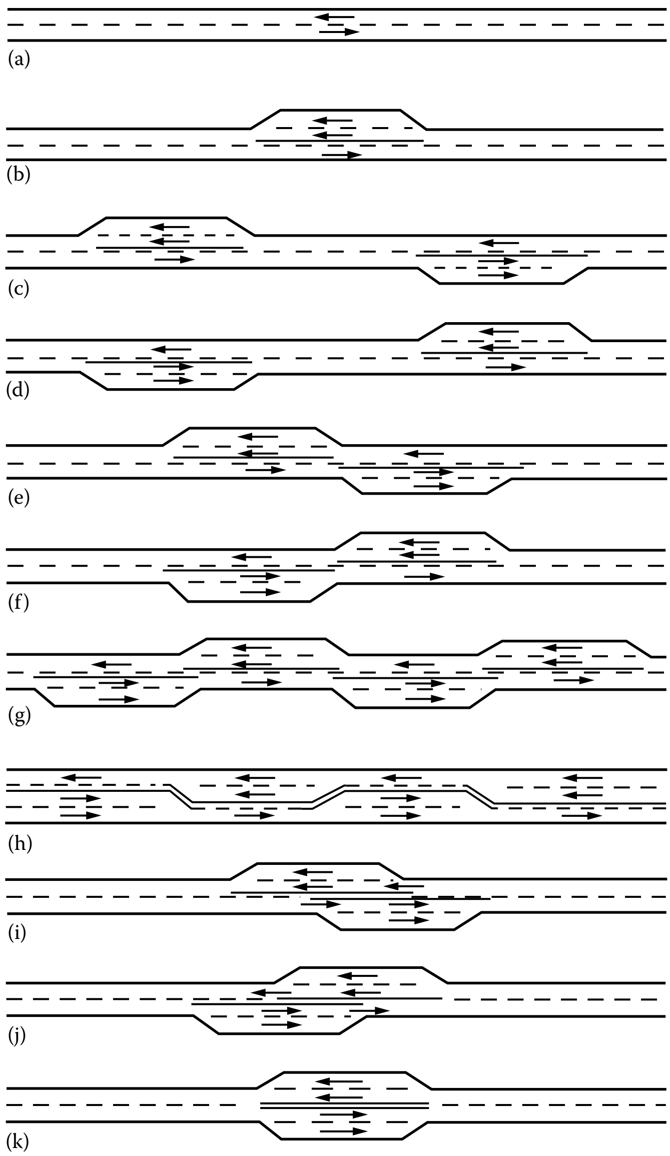


Figure 8.2 Alternative layouts of passing lanes. (a) Conventional two-lane highway, (b) isolated passing lane, (c, d) separated passing lanes, (e, f) adjoining passing lanes, (g, h) alternating passing lanes, (i, j) overlapping passing lanes and (k) side-by-side passing lanes. (From Potts IB and Harwood DW. *Benefits and design/location criteria for passing lanes*. Missouri Department of Transportation, Kansas City, 2004.)

In the case of the 2 + 1 cross-section, it is the centre lane (lane 2 in the two-lane section) that is added or dropped.

At medium and high volumes, roads with continuously alternating passing lanes can provide an improvement of up to two levels of service. With less frequent passing lanes, an improvement typically of one LOS is provided. The influence of the passing lane extends for a significant distance downstream, typically 5 to 13 kilometres downstream, dependant on the traffic volumes present (AASHTO, 2011a).

Table 8.2 Recommended traffic volumes at which passing lanes should be considered

Passing opportunities in the preceding 5 km		*Present year volumes (ADT)* *Slow vehicles (including light trucks and cars towing trailers, caravans or boats) (%)*		
Description	*Passing opportunities (%)*	*5*	*10*	*20*
Excellent	70–100	5670	5000	4330
Good	30–70	4330	3670	3330
Moderate	10–30	3030	28,500	2470
Occasional	5–10	2270	2000	1730
Restricted	0–5	1530	1330	1130

Table 8.2 gives the traffic volumes in the current year at which passing lanes would normally be considered. These are based on relatively short passing lanes at a spacing of 10 to 15 kilometres or more. As a general rule, where passing opportunities are less than excellent, they could be improved toward that rating by increasing the percentage length of road providing passing sight distance.

The passing manoeuvre typically occurs over a distance of 180 to 200 metres so that a passing lane of this length would allow only one vehicle to break out of the platoon. As in the case of the climbing lane, a minimum length of about 800 metres is thus to be preferred.

With signposting and sight of the upcoming passing lane, drivers are poised to make the most of the opportunity to pass the platoon leader. It has frequently been observed that the passing manoeuvre is initiated while the platoon leader is still on the taper. Approximately three to four vehicles successfully pass the platoon leader in the first 500 metres of the passing lane because they tend to drive at short headways, making the most of the passing opportunity provided by the lane. The length of passing lane provided should be sufficient to disperse the average number of vehicles found in the platoon. Too long a passing bay on the other hand could prove to be uneconomic.

Delay would be an obvious measure of LOS. This is the difference in travel time between the desired and the actual travel time. Unfortunately, desired times are a function of driver preferences and do not readily lend themselves to measurement in the field other than by assumptions regarding operating speed. Furthermore, actual travel times would be those achieved by all vehicles on the road over the full length of the road in question, which is significantly more difficult to measure. A surrogate measure is thus required. One of the criteria of LOS for two-lane two-way roads is Percentage Time Spent Following (Transportation Research Board, 2010). Unfortunately, this is also very difficult to measure and a surrogate surrogate criterion, percentage followers, is brought into play (Oregon Department of Transportation, 2010).

Passing lanes are normally used be vehicles travelling at or slightly below the operating speed. Their width should thus be the same as that of the through lanes. Given that passing lanes would normally be required only on roads with medium to high volumes of traffic, it follows that these lanes, and hence the passing lanes also, should have a width of 3.6 to 3.7 metres. The shoulder width should normally be of the order of 3 metres. This width of shoulder should be extended beyond the end of the taper at the end of the passing lane by about 100 to 150 metres to allow for the driver, who has experienced difficulty in completing the merging manoeuvre, to bring his or her vehicle to a standstill on the shoulder.

The lane addition, or passive, taper should have a length of

$$L = 0.4W \cdot S$$

where
L = length of taper (metres)
W = width of passing lane (metres)
S = 85th percentile speed for LOS B

In the case of the lane drop or active taper, the factor of 0.4 in the above equation falls away because of the greater complexity of the merging manoeuver (Potts and Harwood, 2004). Where a statutory speed limit of less than 70 km/h is in force, the speed, S, in the equation for the length of the active taper should equal the statutory speed limit.

Passing lanes are usually located at fairly regular intervals and the spacing between them should allow for a regular process of infilling as traffic volumes increase. For example, an initial spacing of 16 kilometres would allow for infilling to 8 kilometres and then 4 kilometres. At the next step down, it may be appropriate to consider upgrading the road to a 2 + 1 cross-section.

Numerous alternative layouts of passing lanes can be considered. The selection of the layout to apply depends on local circumstances starting with remote or isolated passing lanes and then separated passing lanes followed by various combinations of adjoining passing lanes as shown in Figure 8.2.

Passing lanes are normally built on relatively flat gradients where construction and maintenance costs would be low. The process of locating them thus commences with the location of climbing lanes. Thereafter, the gaps between climbing lanes are infilled on the basis of the anticipated or observed traffic flows. The second step involves examination of the passing opportunities that exist between the climbing lanes. There is little point in providing lanes where adequate passing opportunities already exist. It would also appear more logical to drivers to find a passing lane where otherwise it would not be possible to pass.

8.3.4 Turning lanes

Turning lanes are employed at intersections where turning vehicles could represent a major impediment to the smooth flow of traffic moving straight through the intersection. If there is a channelising island between the turning lane island and the through lanes, it is normally referred to as a turning roadway or slip lane.

The general layout of a turning lane is that it is preceded by a passive taper and deceleration lane and followed an acceleration lane and an active taper. The left turn lane should include a storage lane with sufficient length to store the number of vehicles awaiting a gap in the opposing vehicles prior to turning left. This matter is discussed in Chapter 21. In the presence of traffic flows of as little as 600 to 800 vehicles per hour per lane, one vehicle turning left could cause a significant back-up in the traffic stream if the opposing flow is of a similar magnitude.

Turning lanes are discussed in depth in Chapter 10.

8.3.5 Additional lanes between intersections

Between widely spaced intersections, traffic moves all the time, whereas, at signalised intersections, the traffic stream is halted periodically to allow for the movement of the crossing flows. The green time allowed by the traffic signals could be, say, 30 seconds. An evenly distributed flow of 600 vehicles per hour per lane would theoretically suggest 5 vehicles per lane passing through the intersection in a 30-second interval. However, time must also be allowed for the vehicles waiting for green to start from rest, resulting in fewer than

5 vehicles per lane passing through the intersection. In short, there is an imbalance between the LOS of the lanes at intersections and that between intersections. If this imbalance results in massive back-ups upstream of an intersection, it can be corrected by providing additional through lanes at the intersection.

If intersections are closely spaced and the downstream intersection serves high volumes of turning traffic, the upstream sections of its turning lanes may have to be extended towards the preceding intersection to the extent of becoming an auxiliary lane between intersections.

An intersection involving a two-way two-lane road could, theoretically, have auxiliary lanes for turns to the right and to the left and, in addition, an auxiliary lane for through traffic. A single lane in one direction could thus end up being widened to four lanes! Analysis employing the procedures defined in the *Highway Capacity Manual* is necessary to ensure that excessive auxiliary lanes are not provided.

8.3.6 Additional lanes between interchanges

The operational efficiency of freeways is largely dictated by the extent of weaving present. Weaving is defined in the *Highway Capacity Manual* (Transportation Research Board, 2010) as 'the crossing of two or more traffic streams travelling in the same direction along a significant length of highway, without the aid of traffic control devices' (Chapter 12, p. 1).

Weaving at interchanges arises from vehicles entering the freeway crossing the paths of vehicles exiting from the freeway. A particular example of this is the cloverleaf interchange. In this case, an auxiliary lane, the collector–distributor road, is provided within the interchange but separated from the mainline lanes.

Weaving between interchanges, however, has a greater impact on the LOS prevailing on the freeway. This is because this weaving may involve more than just one lane of the freeway, whereas at interchanges weaving is limited to lane 1 and the adjacent ramps. Vehicles travelling in lanes 2 and higher upstream of an interchange and wishing to exit at this interchange may have to weave across one or more lanes prior to exiting from the freeway. Joining flows from an on-ramp generally wish to escape the turbulence of lane 1 as soon as possible and weave across one or more lanes to enter a smoother flow regime.

Weaving is discussed in depth in Chapter 12, and can be one of three types if the merge/diverge manoeuvre is included. Weaves are classified as either ramp weaves or major weaves and also to one- and two-sided weaves. A one-sided ramp weaving segment is between a single on-ramp lane followed by a single lane off-ramp and connected by means of an auxiliary lane. Every weaving vehicle must make one lane change and the lane changing turbulence occurs between the auxiliary lane and the outermost through lane of the freeway. This is the type A weave of *Highway Capacity Manual* (Transportation Research Board, 2000). A major weave has one or other of the ramps having two lanes. This is the type B weave of *Highway Capacity Manual* (Transportation Research Board, 2000). In the case of a two-lane off ramp, the on-ramp to freeway movement requires one lane change whereas the freeway to off-ramp movement can be made without a lane change. In the case of a two-lane on-ramp, it is the freeway to ramp movement that has to make the lane change.

High volumes of merging or diverging traffic may require two lanes on the ramps to accommodate the ramp traffic. It will thus be necessary to determine whether the two lanes are continued on the freeway by the addition of a new basic lane or, alternatively, are dropped some distance downstream of the on-ramp. On the case of a two-lane on-ramp, the auxiliary lane would be added at a distance upstream of the off-ramp to allow for a smooth flow of diverging traffic onto the off-ramp.

Table 8.3 Camber or cross-slope for various pavement types

Surface type	*Camber or cross-slope (%)*
Gravel	4.0
Chip and spray surface dressing	2.5–3.0
Asphalt	2.5–3.0
Concrete	2.0–3.0

8.3.7 Camber and crossfall on the travelled way

Camber refers to the slope provided to the road surface from a high central point in both directions towards the shoulders whereas a crossfall is a slope in one direction only extending from shoulder to shoulder. The purpose behind providing these lateral slopes is to drain storm water off the road, thus seeking to ensure that vehicles do not go out of control as a result of hydroplaning. The steeper the slope the quicker the water will drain off the road.

However, there is an upper limit. If a two-lane road has a 4 per cent camber, vehicles seeking to overtake others will experience an instantaneous 8 per cent change in orientation as they move over to the opposing lane. This is at or just beyond the stable limit of many trucks – particularly those with high loads – and there is a distinct possibility that they could overturn.

Equally, there is a lower limit to the extent of camber or cross-slope provided. A camber of 1 per cent would not be sufficient to drain the road surface effectively. Furthermore, low values of camber could result in areas where, because of inaccuracy in the construction process, water would pond. This could result in hydroplaning and subsequent loss of control of the vehicle. It would also result in spray hitting the windscreens of vehicles travelling in the opposite direction and temporarily blinding the driver.

Between these two extremes, the camber or crossfall selected is dictated by the surface material of the pavement as shown in Table 8.3.

8.4 SHOULDERS

The shoulder is that portion of the cross-section contiguous with the travelled way. It serves many purposes, and these tend to dictate the width of shoulder provided. A distinction is drawn between the graded shoulder and the usable shoulder. The graded shoulder extends from the edge of the travelled way to the shoulder breakpoint, also referred to in the literature as the hinge point, as shown in Figure 8.3, whereas the usable shoulder extends only as far as the shoulder rounding where the fill slope is steeper than 1*V*:4*H*.

Where safety barriers have been provided, the usable shoulder extends to 0.3 metre short of the barrier. The graded width of the shoulder must, however, make provision of proper support of the barrier posts as they would otherwise be too easily dislodged and not provide the support required to prevent a vehicle from going over the edge of the fill. The total width of the shoulder to the shoulder break point should thus be the sum of the widths of usable shoulder, the safety barrier (typically of the order of 400 to 500 mm) and the width required for support of the barrier posts. The last mentioned width would usually have to be at least 600 mm.

On minor roads serving low traffic volumes, the principal function of the shoulder is to provide lateral support to the pavement layers of the travelled way. To provide this function, the shoulder need thus have a width of only 0.6 metre.

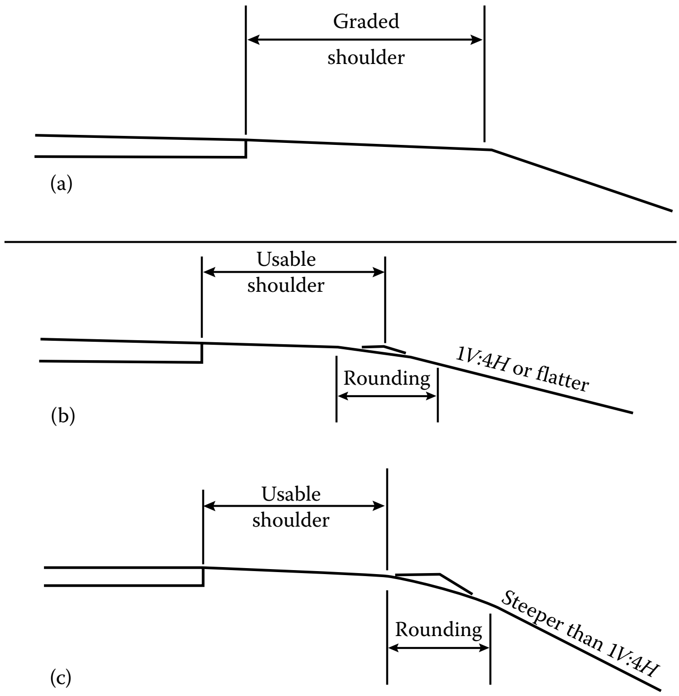


Figure 8.3 Graded and usable shoulder. (a) Gray shoulder, (b) usable shoulder or flat side shoulder and (c) usable shoulder or steep side slope. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on geometric design of highways and streets*. Washington, DC, 2011a.)

Other functions of the shoulder (AASHTO, 2011a) include

- Provision of space for emergencies stops because of, for example, mechanical difficulties
- Allowance for occasional short stops such as for looking at road maps
- Space for evasive manoeuvres to avoid potential crashes
- Improvement of sight distance by moving lateral obstructions to sight distance further from the travelled way
- Provision for stopping by maintenance and other service vehicles including snow removal and storage
- Discharge of storm water further away from the travelled way, thus minimizing the likelihood of seepage of water into the design layers of the road
- Provision of space for safety barriers
- Use by pedestrians and cyclists and as bus stops

In addition, some countries allow temporary occupation of the shoulder by slow moving vehicles enabling overtaking by other vehicles. South Africa has a charming display of courtesy whereby a vehicle that has moved onto the shoulder to allow another to overtake is thanked by a quick flash of the tail lights. The overtaken vehicle then flashes its lights in response, presumably saying, 'Thank you for saying thank you'! This passing lane function supports a higher capacity of the road and uniform speeds. As a safety measure, it is usually only permitted during the hours of daylight to ensure that drivers can clearly see other users of the shoulder.

In terms of the human factors approach to design, the sense of openness that shoulders create also contributes to driving ease and hence reduced driver stress.

As shown in Chapter 5, the design width of passenger cars is 1.8 to 2.0 metres and for semitrailers 2.5 to 2.6 metres. A vehicle stopped on the shoulder should be clear of the travelled way by at least 0.3 metre and preferably 0.6 metre. Internationally, a shoulder width 3.0 metres seems to be standard for high-speed, high-volume roads.

Where there is relatively little traffic or where the terrain may cause a full width shoulder to be unaffordable, a usable shoulder width in the range of 1.8 to 2.4 metres may be considered. These partial shoulders are to be preferred to not having shoulders at all. Although drivers would prefer to pull off the road completely, this may not always be possible so that the stopped vehicle would encroach on the adjacent lane. In the worst case, being of a 2.5 metre wide semitrailer stopping on a 1.8 metre usable shoulder, the encroachment would be 0.7 metre. The balance of the lane would still be available for passing traffic, which may, in turn, have to encroach slightly on the opposing or adjacent lane, effectively sharing the total available space with the other vehicles on the road.

It is useful to remember that geometric design determines the total width of the travelled way and road markings define the width of the individual lanes. The encroachment of 0.7 metre on the travelled way thus means that, on a 7-metre travelled way, each direction of movement would have 3.15 metres available to it.

Where a safety barrier is provided, a clear distance of at least 0.6 metre should be provided between the stopped vehicle and the barrier. With a lesser clear distance, passengers may have difficulty in getting out of the vehicle. This would require a usable shoulder width thus of about 3.6 metres. This width would also be appropriate on high-speed roads with a high percentage of trucks in the traffic stream. However, a possibility exists that drivers may perceive the shoulder to be a travelled lane and use it as such. The colour of the edge marking must differ from that of the lane markings to draw a clear distinction between the lanes and the shoulder.

Shoulders have to be continuous so that drivers have the freedom to leave the travelled way at any point along the road. Without this continuity, some drivers may have to leave the road where no shoulders are available. This creates a potentially dangerous situation.

On high-speeds roads, shoulders are usually surfaced. The knowledge of an available escape route in the case of an unexpected problem engenders a feeling of safety in addition to providing an opportunity for passing.

If shoulders are not surfaced, the provision of a surfaced strip abutting the travelled way may be considered (Irish National Roads Authority, 2007). This may be as little as 1.5 metres wide. The surfaced strip provides for

- Integrity and stability of the pavement layers
- Partial provision for stopped vehicles
- Space for vehicles to slow down while still encroaching on the travelled way, thus facilitating driving onto the unsurfaced shoulder at very low to crawl speeds
- Limited space for evasive manoeuvres
- Space for edge markings
- Reduction of encroachment by vegetation onto the travelled way
- Removal of storm water runoff to a point further away from the edge of the travelled way

Where the intention is to provide gravelled shoulders, it may be found that insufficient material is available for maintenance purposes or that suitable sources of gravel are remote from the road, with resulting cost consequences. Consideration may then have to be given to surfacing of the shoulders.

The crossfall on the shoulders has, if not identical to it, to match the crossfall on the superelevation of a horizontal curve fairly closely. The difference between them should be not more than about 4 per cent. A superelevation of 10 per cent thus suggests that the gravel shoulder should have a crossfall of at least 6 per cent. Water flowing off the high shoulder, in addition to that falling on the travelled way itself, would be flowing swiftly and probably scour the lower shoulder. As a maintenance matter, it may be advisable to surface the low shoulder across the length of the horizontal curve, including the superelevation development on either side of it. With curves following each other in quick succession, a patchwork appearance would be created. Under these circumstances, it may be desirable from an aesthetic point of view to provide continuous paved shoulders.

8.5 MEDIANS

8.5.1 Introduction

As Collins and Hart (1935) stated, 'The function of the central island is to separate the up traffic from the down traffic'. More recently, the median has come to include the inner shoulders on either side of the central island. It is now realised that the function of the median is significantly broader than originally perceived by Collins and Hart. These functions are listed in the following subsection.

Further subsections discuss

- The use of the median as two-way left-turning lanes
- The width of medians
- The profile of medians
- Median end treatments and openings
- Safety barriers on medians

8.5.2 The function of median islands

In addition to providing a separation between opposing traffic flows, medians also (AASHTO, 2011a)

- Provide a drainage channel, which is particularly important in the case of horizontal curves where the outer carriageway would drain towards the median
- Provide storage space for left-turning vehicles, including the auxiliary lane preceding the left turn
- Provide a recovery area for out-of-control vehicles and stopping space in case of emergencies
- Minimise headlight glare through the planting of vegetation on the central island, which vegetation could also serve as impact attenuation in the case of out-of-control vehicles leaving their carriageway
- Make provision for the anticipated provision of additional lanes at some time in the future
- Support split grading, where one carriageway is at a level different from that of the other
- Eliminate left turns, such as at accesses to filling stations or parking areas at shopping centres, between intersections in the presence of high flow rates of through traffic, thus reducing access at these points to right-in right-out operation
- Reduce air turbulence between opposing traffic

In urban areas, medians also

- Provide refuge space for pedestrians crossing the road
- Provide space for turning lanes including refuge space for turning vehicles

- Allow for centrally located street lighting, reducing the number of light poles in the road reserve by half and hence improving the safety of the road
- Control the location of intersection traffic conflicts
- Create open green space available for landscaping

8.5.3 Two-way left-turn lanes

Particularly in commercial areas and, to a lesser extent possibly in industrial areas, there could be a need for left turns between intersections in the road network into off-street parking areas or business premises. To accommodate this, the median could be converted into a two-way left-turn lane (TWLT) (AASHTO, 2011a). In these lanes, opposing traffic flows are channelled into a common lane. Although this may appear to be undesirable from a road safety point of view or perhaps merely inconvenient, it has been found to be effective in practice.

The operation of the lane is that, if there is a vehicle travelling in the opposing direction in the lane, the vehicle wishing to turn to the left either has to travel beyond the opposing vehicle before entering the TWLT lane or wait until the opposing vehicle has passed the point at which it wishes to turn.

The TWLT lane is in direct contrast to the use of the median specifically to eliminate these left turns. The designer must therefore be very clear as to which option in terms of context sensitivity is appropriate to the circumstances surrounding the design.

8.5.4 Median profile

In profile, medians may be raised, flush with the travelled way or depressed.

Raised medians are normally used in urban areas. The height of the median is generally dictated by the height of the kerb that defines it. The raised median is intended to serve as protection for pedestrians. This is principally through its role as a demarcation of the median. They cannot deflect a vehicle that has inadvertently mounted the kerb but motorists tend to stay well clear of kerbing.

Without kerbing, the median would either be slightly rounded, convex or concave, or flush with the travelled lanes on either side of it. The flush median offers no protection against vehicles veering out-of-control into the opposing lane. This intrusion generally results in a head-on crash which, because of the relative speeds of the vehicles involved, is invariably fatal. A further weakness of the flush median is that, on curves, storm water will flow across the outside carriageway and the median adding it to that falling on the inside carriageway. A considerably depth of water, sufficient to cause hydroplaning, on lane 1 of the inside carriageway could result. As a general rule, flush medians are not recommended practice.

A four-lane undivided cross-section normally separates the opposing traffic flows by only no-passing markings, which could perhaps be considered to be a median although with an effective width of only about 500 millimetres. It is generally acknowledged that the safety record of this cross-section is poor.

Depressed medians are usually used in rural areas. They are useful in support of drainage of the roadway as they ensure that water does not have to travel a long distance before clearing the travelled way. On curves, water from the outer carriageway would drain into the median where it could enter a longitudinal drain to a point where a drop inlet could remove it from the median.

Depressed medians should, for preference, have a slope from the inner shoulder to the low point of the median of 1:10 although this could be steepened to not more than 1:6 in constrained circumstances.

8.5.5 Median widths

The width of the median is dependent on the available space to accommodate it and the intended purpose of the median. For example, if the median is intended to accommodate split grading, its width would be dictated by the height difference between the two carriageways and a crossfall of about $1V{:}6H$.

The minimum width of the median could arise in the case of a zero width of the central island. With inner shoulders that are 1.5 metres wide, the median thus would be 3.0 metres wide. However, in the case of a 2 + 1 cross-section, there is often no median at all except where a barrier – typically a cable barrier – is provided. A median width of 0.5 to 1.0 metre would be sufficient to accommodate the barrier.

The maximum practical width of the median should be sufficient that an out-of-control vehicle would not enter the opposing carriageway. As a general rule, most vehicles leaving the road do not cover a lateral distance of more than 9 metres. There is thus little benefit in providing a shoulder wider than this. However, greater shoulder widths could be usefully employed in landscaping and the planting of shade trees, hence creating a park-like environment. Any trees planted in the median should not have a bole with a diameter thicker than about 150 millimetres, as this could become a dangerous obstacle for a vehicle leaving the carriageway out of control. If the natural ground slope is so steep that split grading has to be employed, the median may have to be wider than 9 metres to accommodate the height difference between the two carriageways.

Some authorities have opted for a median width of 18 to 20 metres on the basis of two vehicles simultaneously leaving the road from opposite directions and crashing into one another. It is suggested that the probability of such an occurrence is vanishingly small. In practice, unnecessarily wide medians effectively sterilise large tracts of land that could have been put to better use outside the road reserve, for example, for farming purposes.

Of course, if space permits and land acquisition is cheap, the median could be so wide that the two carriageways are, in effect, two separate roads with their own horizontal and vertical alignments.

There is a fundamental difference in the approach to selection of the median width between the urban and rural environments. At urban intersections, wide medians impose a long travel distance on turning vehicles. This is inimical to the efficient operation of these intersections, especially in the case of signalised intersections where the number of vehicles that could clear the intersection in one phase would be reduced. Narrow medians are thus to be preferred. In the rural environment the 9 metre wide median as discussed previously is the preferred option. Speeds are high and the consequences of a vehicle leaving the carriageway out of control would almost certainly be fatal.

The narrow versus wide dichotomy creates an interesting problem at the interface between the rural and the urban environments. In the one case, narrow is good and, in the other, narrow is bad and wide is good. A suggested approach to resolution of this problem could be to take cognisance of the increased traffic volumes in the urban area and to narrow the median down by adding lanes to the cross-section. A 9-metre median could be replaced by two 3.6-metre lanes and a 1.8 metre wide median comprising two 0.9 metre wide inner shoulders. As an intermediate step, one additional 3.6-metre lane could be provided in the inbound direction towards the urban area, accepting that, in the outbound direction, the lane drop to two lanes could have happened upstream at a point when the traffic volume has declined. The median in the transitional area could thus have the 9-metre rural median replaced by a 3.6-metre lane, a 4.0-metre central island and a 1.4-metre inner shoulder.

If the reserve width is restricted, a wide median should not be provided at the expense of the verges. Narrowing the verge may create the kind of obstacle that the median is designed to avoid. A narrow verge may result in an out-of-control vehicle passing the road reserve boundary and crashing into something or, worse yet, someone who has a perfect right to be there.

8.5.6 Median barriers

Median barriers are intended to prevent an out-of-control vehicle crossing into the path of opposing traffic. They would thus prevent cross-median crashes but the total crash frequency may increase because of the reduction in the space available for recovery manoeuvres. Furthermore, with an increase in the height difference between the two carriageways, an out-of-control vehicle may easily roll down the slope and into the path of an oncoming vehicle. The possibility of cross-median crashes could thus increase with increase in this height differential.

Barriers are discussed in depth in Chapter 14. As described in that chapter, there are three basic forms of barrier:

- Flexible, for example, cable, barriers
- Semirigid, for example, steel W-beam, barriers
- Rigid, for example, concrete New Jersey profile, barriers

In selecting the type of barrier to be deployed, the dynamics of lateral deflection have to be considered.

Cable barriers are relatively inexpensive to install and maintain and, because of their fairly low rate of lateral deflection, are the safest of all barriers in terms of injury to the occupants of the vehicle striking the barrier. The low rate of deceleration carries with it the penalty of a deflection of 2.5 to 4 metres. They are thus not appropriate to narrow medians. In spite of this, they are often used on 2 + 1 cross-sections.

Semirigid barriers require provision for a lateral deflection of 1.0 to 1.5 metres so that they can be used in the case of relatively narrow medians. In the final analysis, the width of the median is predicated on its function. Various median functions and the associated minimum median widths are illustrated in Table 8.4.

Concrete barriers allow for no deflection at all and rely on the shape of the profile to redirect vehicles in the original direction of travel. Vehicles with a high centre of gravity can, however, capsize over the barrier and spill their loads onto the adjacent carriageway. The possibility of their striking objects within 3 metres of the barrier face should not be disregarded. Relatively light structures such as sign supports or light poles mounted on the median may not be able to withstand the impact and result in secondary crashes.

Table 8.4 Suggested minimum median widths

Median function	*Desirable min (m)*	*Absolute min (m)*
Separate traffic flows with concrete barrier	3.7	1.9
Shelter a small sign	1.2	1.0
Shelter signal pedestals or lighting poles	2.0	1.4
Shelter pedestrians and traffic signals	2.5	2.0
Shelter turning vehicles and traffic signals	6.0	5.0
Shelter crossing vehicles	7.0	6.0
Rural median on high-speed road	9.0	6.0 plus barrier
Provision for road widening with concrete barriers	11.7	9.3

Source: Botswana Department of Transport. *Road design manual*. Gaborone, 2012.

Note: The absolute minimum rural median comprises two 1.5-metre inner shoulders and 1 metre of central island on either side of a semirigid barrier.

A further consideration in the selection of median barriers is their effect on sight distance on horizontal curves as discussed in Chapter 7.

8.5.7 Median end treatments and openings

It is necessary to provide median openings at all intersections except those where the left turn is prohibited. The end treatment of these openings can be

- Semicircular
- Bullet nosed
- Flattened bullet nosed

The end treatments are illustrated in Figure 8.4.

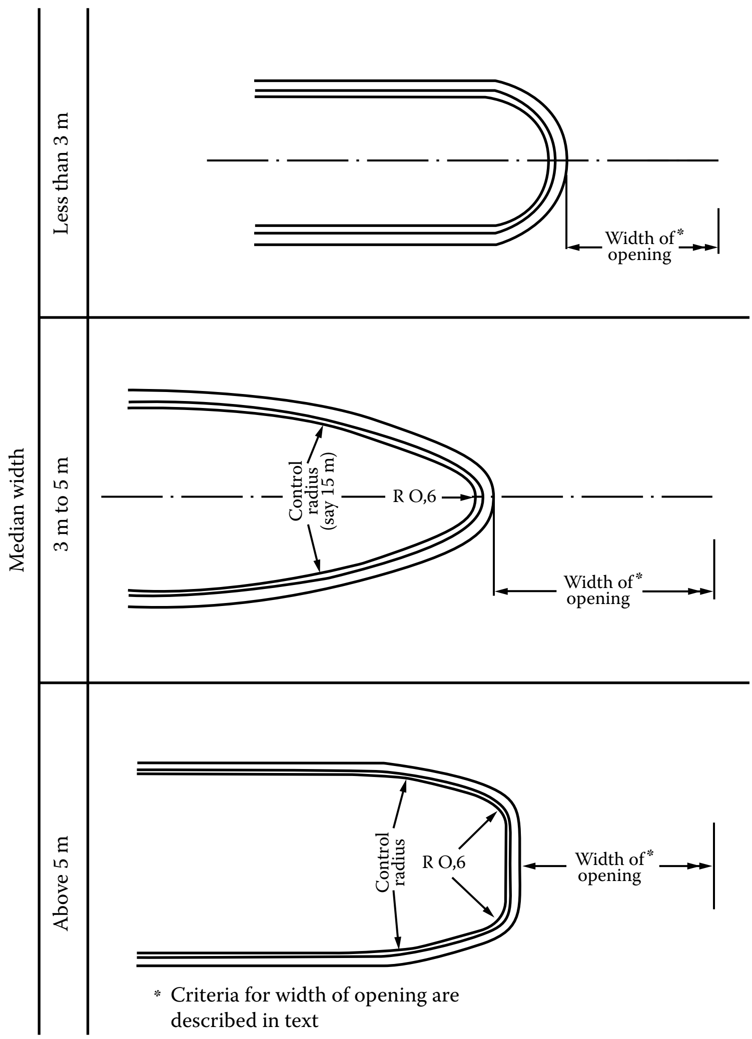


Figure 8.4 Median end treatments. (From Wolhuter KM. *TRH 17: Geometric design of rural roads*. Department of Transport, Pretoria, 1988.)

The bullet nosed and flattened bullet nosed end treatments provide guidance to the turning vehicle by being based on the turning templates illustrated in Chapter 5. These end treatments thus guide the turning vehicle towards the centreline of the crossing road. They are formed by two arcs of the control radius with a short intervening radius, typically 0.6 to 1.0 metre in extent. The control radius is determined by the selection of design:

- Passenger car 12 metres
- Single unit truck 15 metres
- Semitrailer 23 metres

The semicircular end treatment does not guide vehicles into the opposing stream of traffic on the cross-road as one may expect because the radius of this end treatment is typically of the order of 1.5 metres and too small to provide any form of guidance.

The absolute minimum width of the median opening is dictated by the width of the crossing road and should be equal to the roadway width, including the shoulders. In the case of the crossing road being kerbed, the absolute minimum width is the width of the travelled way of the crossing road plus a clearance of 300 millimetres on either side of the road to allow for the constricting effect of the kerbing.

8.6 OUTER SEPARATORS

The functional classification of roads as discussed in Chapter 3 implies that, desirably, a road should have one function and one function only. Unfortunately, it is possible that the function of a road can change across its length. A rural road entering a town or village would have a mobility function, whereas, as it passes through the central business district (CBD), it would acquire an accessibility function. Vehicles passing through the town and out the other side would, however, prefer that the mobility function be retained.

Where a hand tool such as a shifting spanner can perform a variety of functions, it normally does not do any of them particularly well. Similarly, the mixed functionality of the road in the CBD results in neither function being well served. Slow-moving traffic, where the drivers are seeking parking, for example, will inevitably impact unfavourably on traffic wishing to pass through town without hindrance. The faster moving vehicles would complicate the left turns of local traffic into crossing streets. Each function will be an impediment to the other.

The outer separator serves to split the two functions and is a longitudinal island between the faster moving through traffic and the slower local traffic. In effect, the through traffic is served by an arterial road and the local traffic by a frontage road. The outer separator includes the shoulders of the roads on either side of it. The outer separator provides space for

- Parking
- Use by pedestrians and cyclists
- Accommodation of height differences between the arterial road and the service road
- Underground services
- Street furniture such as road signs and traffic signals
- Landscaping

Parking is normally provided only on the service roadside of the outer separator. The island part of the outer separator could be as narrow as two kerbs back to back or as wide as required by the functions bulleted above.

The outer separator would have to be widened where the service road crosses a road that crosses or enters the arterial. This is particularly the case where the parking is provided away from the outer separator, that is, on the other side of the service road. As a minimum, the extra width allows for a vehicle on the arterial road turning through 180° onto the service road. Desirably, the entrance to the crossing road from the service road should be sufficiently removed from the arterial to include storage space on the crossing road in advance of its intersection with the arterial.

8.7 BOULEVARDS

The word boulevard derives from the Dutch *bolwerk*, with the English equivalent being bulwark. The original French meaning was a reference to the flat top of a rampart. The boulevards of Paris replaced the original city walls, and therefore encircle the city centre, in contrast to the avenues which are radial from the centre. A different application of the word is that it applies to the outer separator, which, in this case, is usually wider than the normal outer separator and liberally landscaped.

In terms of context-sensitive design, the boulevard seeks to replace an otherwise sterile environment with a lush parklike setting with shade trees, sidewalk cafes, memorials and public art where people could socialise and stroll as illustrated in Figure 8.5.

The sidewalks are thus wider than those provided on other streets and usually include cycle lanes. Typically, cyclists are catered for on separate cycle paths. These functions are combined with the provision of space for the unimpeded movement of traffic. Public transport is a major feature of boulevards and bus stops are provided at short intervals. The principal difference between the boulevard and the arterial/outer separator/frontage road cross-section is that, although they may end up looking similar, the former is designed as a landscaped unit whereas the latter is the sum of individual components.

Figure 8.5 Champs Elysees, Paris.

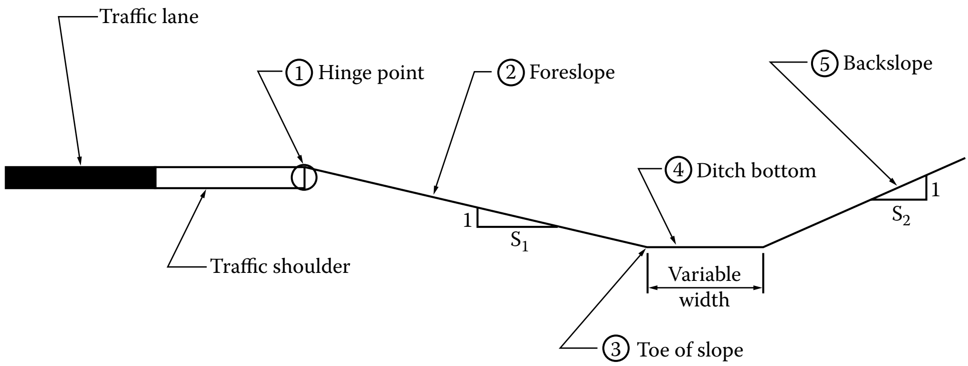


Figure 8.6 Components of the road prism side slopes. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on geometric design of highways and streets.* Washington, DC, 2011a.)

8.8 SIDE SLOPES

The roadside comprises the verge, discussed in the following section, and the section of the road prism beyond the shoulder breakpoint or hinge point. This section of the roadside has three components: the foreslope, the backslope and the side drain as shown in Figure 8.6.

The main functions of the side slopes are to ensure stability of the earthworks against failure such as slips and also to reduce the probability of injury or loss of life when a vehicle leaves the road. These two requirements both suggest that slopes should be as flat as economically possible within the constraints of the space available within the road reserve.

8.8.1 Slope stability

The maximum flow speeds that should be considered are listed in Table 8.5. These slopes are dictated by the angle of repose of the construction material involved, that is, the steepest slope at which the material is still stable. A decision of the slope to be employed should be informed by a geotechnical inspection of the material and the values offered in Table 8.5 are to be considered only as guidelines. As can be seen in the table, the fineness of the material, from rock to fine sand, and its cohesiveness dictate what these slopes could be.

Table 8.5 Maximum flow speeds in text for various materials

Type of material	*Max. permissible velocity (m/s)*
Fine sand	0.6
Loam	0.9
Clay	1.2
Gravel	1.5
Soft shale	1.8
Hard shale	2.4
Hard rock	4.5

Source: Burrell RC et al., *Geometric design guidelines*. South African National Road Agency Limited (SANRAL), Pretoria, 2002.

Where the natural cross-slope is steep, the cross-section would in all probability have a side-fill side-cut configuration. The side-fill, being placed on a steep ground slope, would constitute a super load potentially creating an unstable situation. It would thus be prudent to create a series of horizontal benches in the *in situ* material prior to construction of the fill.

As shown in Figure 8.5, a properly designed side drain should be provided at some distance away from the foot of a fill slope to avoid erosion caused by storm water flowing next to the toe at speeds sufficiently high to cause scour. Erosion could result in the bottom of the fill acquiring a steeper and unstable slope followed by the inevitable slip. The slip area could move higher and higher up the fill slope and even invade the shoulder area as well. Side drains should be trapezoidal in cross-section for ease of maintenance and removal of silt.

8.8.2 Safety

Safety considerations require that, if a vehicle leaves the road, the driver should be able to regain control and possibly also return to road. To achieve this, the cut or fill batter should not be steeper than 1*V*:4*H*. Even at this batter, recovery would not be easy. It is, after all, the equivalent of a negative superelevation of 25 per cent. At a batter of 1*V*:3*H* there is a strong possibility that the vehicle would capsize during its progress down the slope.

The side drain adjacent to the toe of the fill should not be deeper than 150 to 300 millimetres. This restriction on the depth of drain is to ensure that a vehicle can actually traverse the drain without digging into the far side of the drain and somersaulting. The preferred depth is 150 millimetres and the greater depth is a compromise between the safety of the vehicle traversing the drain and the integrity of the design layers of the pavement.

The shoulder break point should be well-rounded, as a sharp change of slope from the shoulder to the fill slope could cause a vehicle to become airborne and hence impossible to control. The consequent increase in speed may also result in the vehicle moving beyond the end of the clear zone before being brought to a standstill.

8.8.3 Aesthetics

Most land forms are gently rounded. Uniform values of cut and fill slopes would result in the earthworks, looking as though they had been cut out of cheese and be totally unnatural. The visual contrast between the side slopes and the natural slopes would be glaring and ugly. Generous rounding at the shoulder break point, without any narrowing of the usable shoulder, would be a start towards a more aesthetic cross-section. A further improvement could be by adopting a flat slope at the start of the cut or fill and then rotating it to a steeper slope as the height of cut or fill increased.

Flat slopes and generous rounding at the break point result in what is sometimes referred to as a streamlined cross-section (AASHTO, 2011a). Winds flow smoothly across the road without the creation of eddies and this is very useful in the natural removal of desert sand or snow from the travelled way.

8.9 VERGES

8.9.1 Introduction

The verge is the area of the cross-section between the road prism and the reserve boundary. Its width thus varies constantly as the height of the earthworks rises and falls relative to the natural ground level.

This area is referred to in the United Kingdom, southern Africa and Australia/New Zealand as the verge. In the United States, it has many names, some fanciful and some downright misleading, such as

- Berm
- Boulevard
- Curb strip
- Easement
- Grass bay or grass plot
- Neutral ground
- Parking, parkway or parking strip
- Street allowance
- Swale

The functions of the verge are to provide

- A buffer zone between the road and its adjacent land uses
- Space for a recovery area
- Space for utilities both above- and underground such as telephone and power lines, water reticulation, gas lines and cable TV
- Space for sidewalks and bicycle paths
- Space road signs and supports for overhead signs
- Space for street lighting where the light poles are not located in the median or on the shoulder
- Provision for landscaping
- Space for bus stops and lay-bys
- Space for aboveground storm water drainage and underground storm water reticulation
- Sight distance in horizontal curves and for intersections
- A hard surface for vehicles that need to stop on the road and want to pull off the shoulder as well
- Work space for maintenance vehicles
- Space for retained indigenous growth
- Berms to help to reduce the negative environmental impacts of the road such as noise and headlight glare
- Optional space for future expansions to the travelled way
- Accommodate driveway approaches for adjacent properties

To accommodate all of these functions, the combined width of the verges on either side of the road should be not less than 10 metres, usually 5 metres on either side but, if need be, the verges could be asymmetrical.

8.9.2 Drainage elements

Storm water drainage comprises longitudinal surface drainage whereby water from the road surface is captured in drains or kerbs and channels and transported to a point where it can be discharged either into an underground system or into a water course.

In rural areas, the components of this drainage include

- Catchwater drains, which collect water uphill of the top of cuts and transport it parallel to the top of the cut slope, or down a chute constructed in the cut face, to a side drain or water course and are intended to protect cut slopes from erosion

- Side drains, parallel to the road centreline and located in the verge at a distance away from the toe of fill or cut
- Mitre drains, which remove water from the side drain and discharge it outside the road reserve, normally in rural areas
- Edge drains, normally used on high fills and located near the shoulder break point because simply discharging the storm water over the edge of the shoulder may result in erosion
- Chutes, which convey storm water down the fill slope, usually into a stilling basin to reduce the risks of erosion caused by an otherwise highly energised stream of water

These components are illustrated in Figure 8.7.

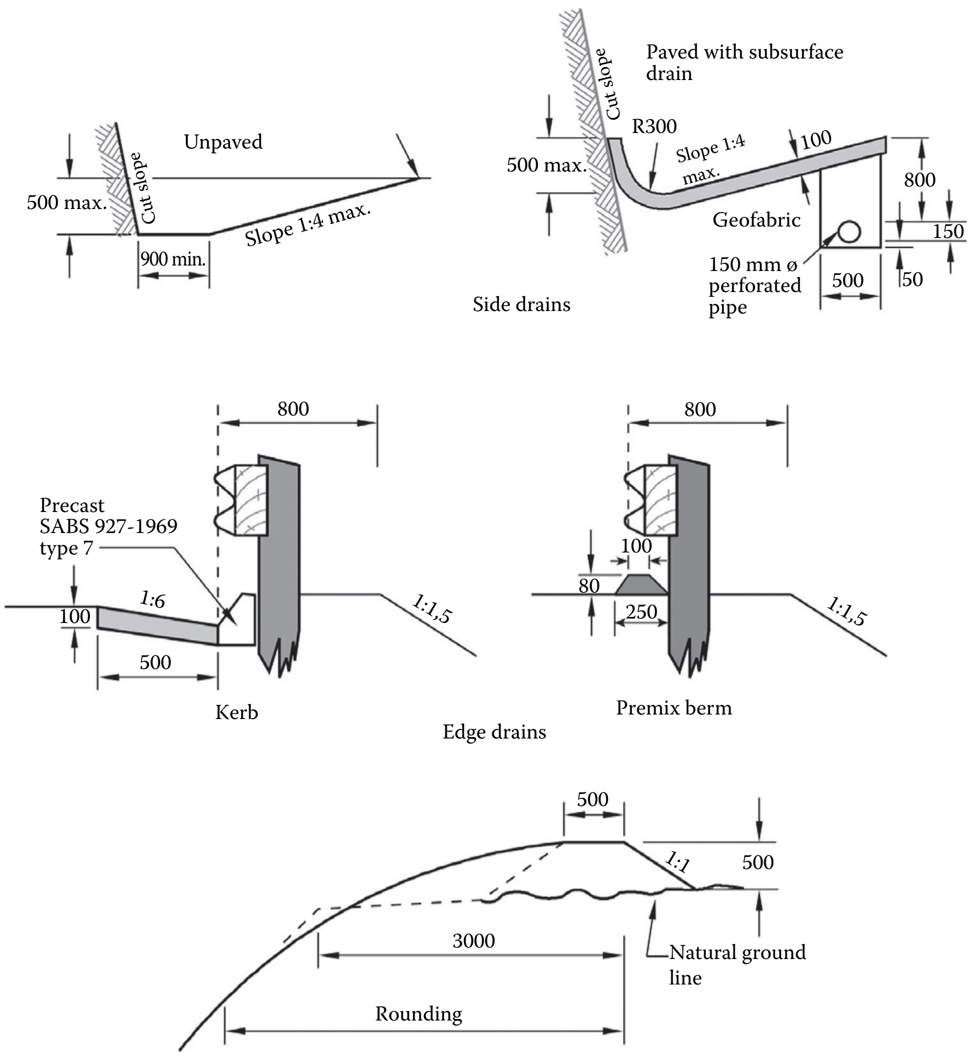


Figure 8.7 Elements of the rural drainage system. (From Burrell RC, Mitchell MF and Wolhuter KM. *Geometric design guidelines*. South African National Road Agency Limited [SANRAL], Pretoria, 2002.)

In urban areas, the side drains are replaced by kerbs and channels that convey storm water via drop inlets into the underground drainage system. Kerbs are divided into three categories:

- Barrier kerbs
- Semimountable kerbs
- Mountable kerbs

These are illustrated in Figure 8.8.

Kerbs and channels should not be flatter than 0.5 per cent because this is approaching the limit of self-cleansing flow speeds. If stringent quality control is brought to bear on the gradient of the channels, a gradient of 0.3 per cent could perhaps be considered. These gradients obviously apply also to the vertical alignment of the road.

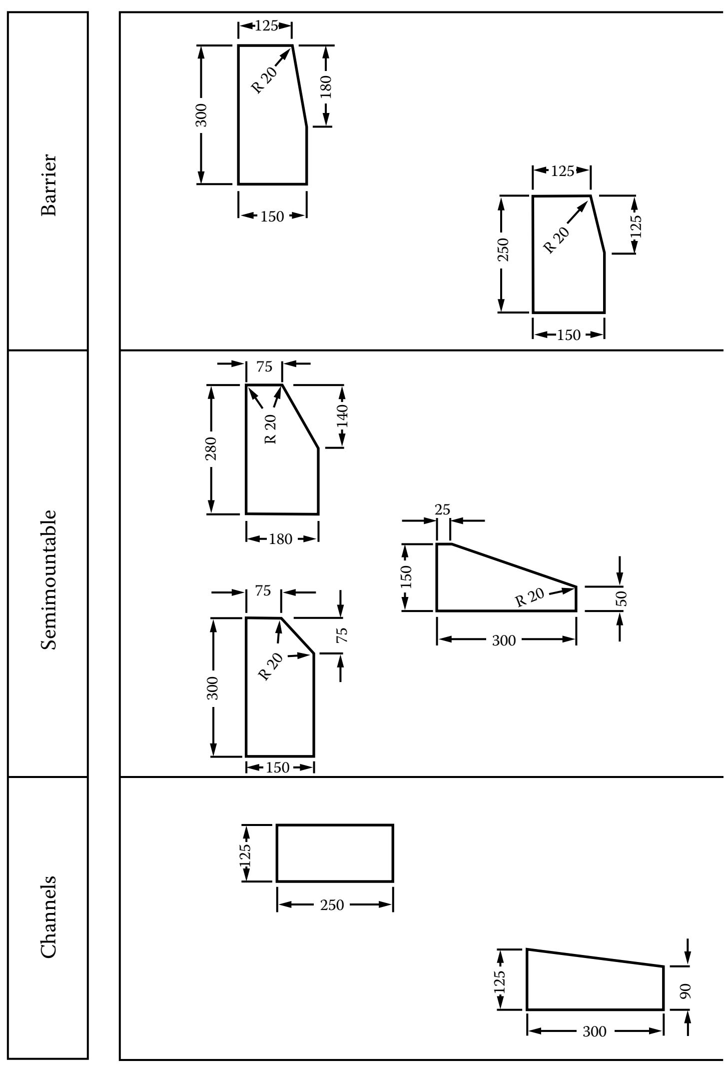


Figure 8.8 Typical kerb profiles.

Channel grading could, however, be considered as an alternative. In this case, the gradeline of the road would be at a gradient less than the 0.5 or 0.3 per cent but the channel would be at the minimum acceptable gradient. The edge of the channel abutting the travelled way has to have the same level at the outer edge of the travelled way but the channel edge abutting the kerb face could become steadily deeper. Barrier kerbs are usually 150 to 200 millimetres high and have a total height of about 300 millimetres, with a portion of their height being buried. The outer edge of the channel shouldn't be deeper than the bottom of the kerb face. This does place a limit on the possible length of the channel between drop inlets.

The sizing of the components of the rural and urban drainage systems are addressed in detail in Chapter 23.

8.9.3 Provision for utilities

Urban residents are provided with numerous utilities such as potable water, power, sewerage, telephone services and gas. These utilities require, in addition to their original installation, ongoing maintenance and this makes it necessary for service providers to be able to gain easy access to their service lines. If each utility had to be located within an exclusive reverse, the cost of provision of the utility would be very high. It is logical, therefore, that these utilities be accommodated within the road reserve (Sampson et al., 2011).

It seems almost to be the norm that, after a road has been constructed and opened to traffic, it is necessary to dig a trench across it to provide some or other utility. Proper planning and attention to detail in the design process in respect of the road as well as the utilities concerned is necessary to avoid interruption of the traffic flow and damage to the road surface and pavement structure. Furthermore, other issues regarding the sharing of the road reserve with other service providers need to be addressed. These include the fact that

- Services such as poles for the support of overhead power and telephone lines are obstacles that create a potential road safety hazard.
- Spillage from blocked or broken sewers can, apart from the unpleasant stench, be a health hazard to residents in the vicinity of the break.
- Burst water pipes invariably cause damage to the pavement layers of the road.
- The width of the road reserve may have to be increased to accommodate the utilities.
- The relocation of services, when it is necessary to upgrade or relocate the road, can be expensive.

The complexity of providing utilities within the road reserve tends to be legal rather than engineering. This arises from the fact that the road authority has a duty of care and obligation to the road users to ensure that roads are safe and can be used without any undue risk. Failure to comply with this obligation would result in the road authority be held liable for damages caused to road users.

Addressing the issues listed above require a clear contractual agreement between the road authority and the service provider concerned regarding responsibility for the physical repairs and the financial issues surrounding repair to the road and the other utilities that may be damaged either as a result of the failure of a utility or arising from the process of its reinstatement. It is necessary to register a servitude in respect of the utility concerned to define the responsibilities and rights of each party to the agreement. A servitude is defined as the right of use of land that is registered in favour of the service provider who is not the owner of the land.

The registration of the servitude is normally accompanied by a wayleave. The wayleave clearly spells out the terms and conditions subject to which the wayleave is granted and that the interests of the road users and general public are protected. The service provider is usually required to accept full responsibility for compliance with all safety standards and requirements during the initial installation of the utility and its subsequent maintenance. The road authority should also be indemnified against any claim that may arise as a result of the installation, maintenance and operation of the utility, including failure or other malfunction.

Most local authorities prefer that utilities should not be located under the shoulders or travelled way of the road. Access to the utilities through a manhole is hazardous not only for the workers engaged in maintenance works but also for road users. It is not unknown for manhole covers to be stolen so that, even when there is not work in progress requiring the removal of the manhole cover, the hazard remains. Every time the road surface is upgraded, for example, by applying a premix overlay, it is necessary to raise the manhole cover to be flush to the road at the new level and this is a distinct nuisance in maintenance operations. It follows that utilities should, as far as possible, be located in the verges. Crossings of the

Table 8.6 Typical accommodation of utilities in road reserves

Utility	*Type*	*Typical accommodation*
Storm water		Generally underground in urban areas. Surface in rural areas. Within road reserve
Water courses, irrigation pipes		Not in road reserve
Water supply	Potable bulk supply	Not normally in road reserve
	Municipal distribution	Underground, normally on high side of road reserve. Trench 0.7 m wide and 1.5 m deep
	Fire hydrants	Maximum 1.5 m from reserve boundary on low side of road
Sewerage	Bulk	Not normally in road reserve
	Local distribution	Normally on low side of reserve and adjacent to reserve boundary. Trench 1.0 m wide and a minimum of 1.5 m, for preference 3.0 m deep
Electricity	High voltage (>22 kV)	Generally overhead. Pylons too large to be accommodated inside reserve. If crossing the reserve, 7.5 m minimum vertical clearance
	Low voltage urban (<22 kV)	Either overhead or underground in reserve
	Low voltage rural (<22 kV)	Generally overhead immediately outside reserve and on opposite side from telephone lines
	Local distribution (125/220 V)	Generally underground near the high side boundary fence but possibly also above ground in same location
Telecommunication	Rural	Overhead within reserve, near reserve boundary. When crossing the road, vertical clearance 6.5 m. When if underground, within a sleeve at least 1.0 m deep. On opposite side of reserve from power lines
	Urban	Underground, near low side boundary
Gas or steam		Underground, remote from other utilities
Oil or fuel		Not in reserve except when crossing and then in a culvert

Source: Sampson JD et al., *Manual on services in road reserves*. South African Committee of Transport Officials (COTO), Pretoria, 2011.

road by utilities should be either in sleeves, pipes or culverts so that maintenance work can be carried out without excavating through the road surface.

Servitudes cannot be provided on all roads. For example, freeways normally carry large volumes of traffic at high speeds so that installation and maintenance of utilities in their reserves could be dangerous unless provision can be made for alternative access to the servitudes. Where road reserves are narrow, there may simply not be sufficient space to accommodate a verge wide enough to make provision for utilities in the reserve. If the road has to be widened, it is possible that utilities in the reserve would have to be relocated and the wayleave should clearly spell out that such relocation would be by and for the account of the service provider.

Many road authorities define precisely where in the verge utilities may be located. These definitions are to ensure that one utility is not contaminated or placed at risk by another. For example, potable water reticulation is often on the opposite side of the road from sewerage lines to ensure that leakage from the sewers does not contaminate drinking water. If this is not possible, sewer trenches are normally deeper than those for potable water reticulation. If underground, the width and depth of the trench in which the utility is to be installed or the road is crossed is usually specified. In the case of utilities located above ground, their location is usually right at the reserve boundary and where they cross the road the minimum clearance height is usually specified. The utility locations and clearances recommended for use in southern Africa are shown in Table 8.6.

Many local authorities prefer to employ the nomenclature of 'left side/right side' as opposed to 'high side/low side' in the location of utilities. This is useful in eliminating crossings of the road reserve in areas where the fall of the natural ground slope varies from side to side.

8.10 ROADSIDE AMENITIES

8.10.1 Introduction

Roadside amenities are provided for the benefit of road users. If it is not possible to provide these within the normal road reserve, the reserve width may be increased or, alternatively, they could be located outside the road reserve.

Discussed in this section are

- Rest areas
- Service stations and restaurants
- Trading areas

In addition to these, roadside amenities include pedestrian and cycle paths that do not form part of the road cross-section. These paths are discussed in Chapter 15. Bus stops could also be described as being roadside amenities as they do not form part of the normal cross-section. They are discussed in Chapter 16.

8.10.2 Rest areas

It is well known that drivers extend the length of time that they drive without a rest. It is equally well known that drivers are far more likely to be involved in a crash when they are tired than when they are rested. The general recommendation by safety practitioners is that drivers should not drive for longer than about 2 hours without a break. Drivers could be encouraged to take a rest by the provision of rest areas on roads that serve long-distance travel.

An attractive rest area would be a powerful inducement to drivers to stop and rest. It should therefore, as a minimum, have

- An interesting and beautiful view
- Numerous shade trees
- Rest area furniture

Rest area furniture includes

- Tables and chairs
- Barbecues
- Litter bins

Rest area furniture should be constructed of reinforced concrete and be well-buried to be vandal proof as far as possible.

Rest areas should be located by, in the first instance, identifying sites with attractive views. Further sites could then be located by infilling between the identified sites. These sites should typically be spaced alternately on either side of the road at about 15- to 20-kilometre intervals, that is, 30 to 40 kilometres apart for each direction of travel.

A typical layout for a rest area is shown in Figure 8.9.

Vehicles should be able to pull off the road completely onto a short length of road within the rest area. This inner road should be wide enough to allow for the movement of vehicles in both directions and a width of 6.6 to 7.0 metres would suffice for this purpose. The inner road should be surfaced to the same standard as the through road.

Drainage of the rest area is important and care should be taken to ensure that water does not drain from the rest area onto the adjacent through road. Care should also be taken that there is no ponding of water on rest area road or elsewhere in the rest area.

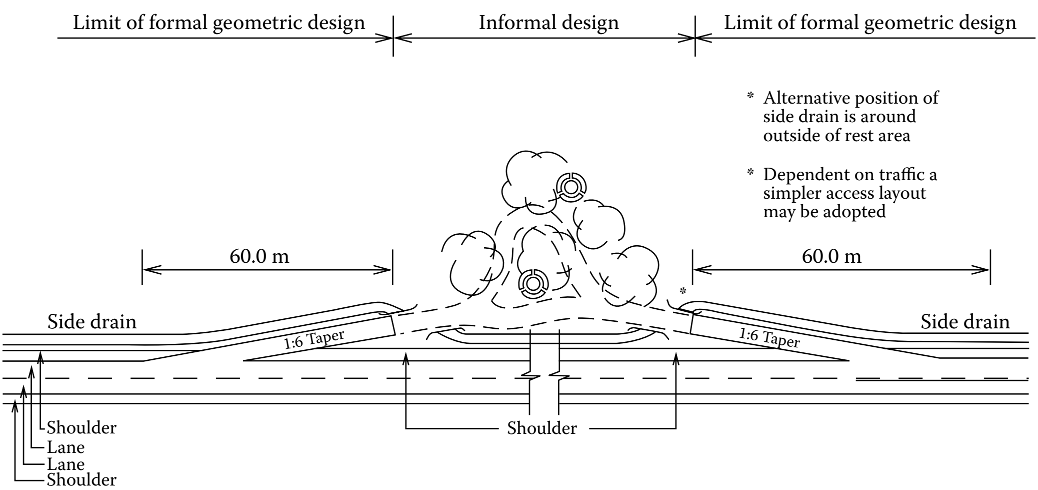


Figure 8.9 Typical rest area layout.

8.10.3 Service stations and restaurants

On a long journey, travellers will usually be unfamiliar with the layout of the towns and villages along their route and would thus not be particularly interested in exploring them to locate service stations, restaurants or toilets. These are normally provided by the private sector, usually by oil companies, on the outskirts of towns and adjacent to the road reserve.

The access to and from service stations should be by way of off- and on-ramps similar to those provided at interchanges. The reason for this generosity of design is that, after a long period of high-speed travel, drivers become desensitised to the speed at which they are travelling. For example, reducing speed to 60 km/h after some hours at 120 km/h generally feels as though getting out and walking is a logical alternative to the current 'crawl' speed!

In the case of service stations on freeways, the owner has two options:

- A duplication of the service station and restaurant on either side of the freeway reserve
- A duplication of the filling station and parking area and location of the restaurant on a bridge over the freeway

Although attractive to the owner because of cost savings and novelty value to customers, the latter is generally not a popular option as far as the road authority is concerned because the structure places a limitation on the abnormal loads that sometimes have to be transported along the freeway. This can be overcome by routing the abnormal load off the freeway, through the parking area and back onto the freeway via the off- and on-ramps. Should the road authority approve an application for a restaurant over the freeway, it would be prudent to ensure that the parking area is so designed that the abnormal load has a reasonably direct route from the off-ramp to the on-ramp.

These are discussed in Chapter 12 and referred to as service interchanges.

8.10.4 Trading areas

Particularly on routes serving tourist traffic, local inhabitants seek to make an income by selling object d'art to passers-by. In Africa, these normally take the form of small hand-carved statuettes of animals, wicker baskets and the like. The traders have no compunction in occupying the shoulder of the road, resulting in would-by buyers walking on the travelled way.

Policing of trading areas is an exercise in futility because, as soon as the police turn their backs, the hawkers emerge once more. In any event, chasing people away and confiscating their stock is inhumane, as this invariably is their only source of income. It is recommended that the presence of the hawkers be accepted and that the verge be designed to be safe both for them and for their clientele to shop in safety.

To this end, the shoulder should be 3.0 metres wide and, for preference, surfaced and extended to about 20 to 30 metres on either side of the 'shopping' area. A further 2.5 to 3.0 metres should be added onto the width of the shoulder to allow the traders to display their wares.

The trading area should not be located on the downstream side of a crest curve. This is to avoid the possibility of a driver being surprised into a panic stop by another vehicle pulling back onto the road after a stop on the shoulder. Similarly, location beyond a horizontal curve, where there is a possibility of vegetation obscuring the line of sight to the trading area, should also be avoided.

8.11 ROAD NARROWING

Road narrowing can take one of three forms:

- A reduction in the number of lanes comprising the travelled way
- Retaining the number of lanes but reducing the width of the individual lanes
- A change from a dual carriageway cross-section to a multilane undivided cross-section

There is, of course, a fourth possibility, which is a combination of a reduction in the number of lanes and a reduction in the width of the lanes. However, this should not be done simultaneously, as it would almost certainly provide a level of stress that the driver would no doubt prefer to be without.

Although a lane drop could simply be a case of ending an auxiliary lane, what is intended here is a reduction in the number of basic lanes comprising the travelled way. This typically applies to the road outbound from an urban area into an outlying or rural area where traffic volumes diminish with every passing possible destination. Ultimately, a point is reached where the number of lanes can be reduced while still maintaining an adequate LOS.

A lane drop in the outbound direction is normally accompanied by a lane addition in the inbound direction. Where the four-lane cross-section is a dual carriageway, the conversion to a two-lane cross-section requires a certain amount of thought.

The centrelines of the two cross-sections should be offset from one another. If they are not, outbound traffic, that is, the flow that is being confronted by the lane drop, has to merge and also follow a reverse curve alignment. In effect, the entire carriageway is dropped. The inbound traffic is also confronted by a reverse curve alignment but this is followed by a passive taper which is much more convenient to follow. The major problem, however, is that urban areas tend to expand into the surrounding rural area. A common centreline would thus result in the two-lane road having to be demolished and replaced by two new carriageways.

On the other hand, if the centreline of the two-lane section is lined up with the centreline of one of the carriageways of the dual-carriageway section, it follows that no abortive expenditure is incurred and the other carriageway is simply extended and the lane drop relocated further downstream. The question remaining is the selection of the carriageway to be dropped. If the two-lane cross-section lines up with the outbound carriageway, the possibility is created that vehicles inbound on the two-lane cross-section could continue in a straight line ending up on a reciprocal heading with outbound vehicles in the fast lane. The likelihood of a head-on crash is obvious. The two-lane cross-section should thus be aligned with the inbound carriageway of the dual carriageway as illustrated in Figure 8.10.

A further application of a reduction in the number of lanes is that aimed at traffic calming. This is in the case of reducing the volume of traffic passing through a residential area, that is, rat-running. Rat-running occurs when a commuter route becomes congested to an extent unacceptable to the commuters. They then seek an alternative route roughly parallel to the commuter route and through a residential area. This is obviously unacceptable to the local residents because of the danger caused to pedestrians, particularly children, by high traffic volumes in what should actually be a quiet residential street. Noise and air pollution are also unwelcome additions to the quality of life. The narrowing is from two lanes to one over a short distance, a choker in fact. Attempts by commuters to access the street result in a level of backup at the choker that is worse than what they would otherwise experience on the original commuter route. The resident's problem is thus resolved.

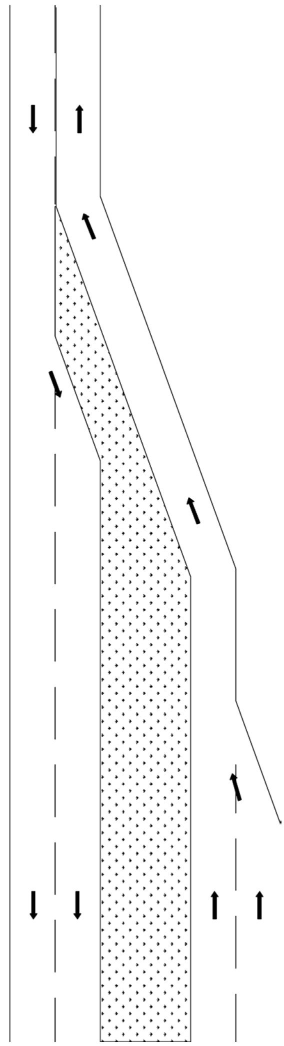


Figure 8.10 Transition from a dual carriageway to a single carriageway cross-section.

The problem of rat-running arises from the historical grid layout of residential townships. Traffic calming is a retrofit to reduce both the volume and the speed of through traffic on streets not designed to accommodate it. It is possible to design townships is such a way that there are no streets that offer short cuts to anywhere, or serve as alternatives to congested commuter routes. Although this is usually done in green fields (new) developments, the older areas of the town or city being closer to the city centre are obvious targets for rat-running and the typical grid structure allows it.

When roads are being upgraded by the addition of lanes, traffic has to be diverted onto deviations. In urban areas, it often happens that the deviation has to be located in the road reserve of the road being upgraded. As a consequence of the shortage of space, the lanes of the deviations are very narrow and typically of the order of 3 metres wide. Fortuitously, this slows the traffic considerably with a distinct safety benefit to the construction workers involved in the upgrade. Furthermore, under these circumstances, drivers are highly focussed on the task at hand and crash rates on deviations are typically low.

8.12 VERTICAL AND HORIZONTAL CLEARANCES

Vertical clearances are measured between the highest point on the road beneath a structure or other vertical obstruction and the bottom or soffit of the vertical obstruction at that point. In the case of a bridge, the clearance is often measured at the intersection of the centrelines of the road and the bridge deck. However, this may produce a misleading result because in the extreme case of the intersection of two roads, both of which are on steep gradients and minimum radius horizontal curves, the lowest clearance between them would be at a point remote from the intersection of the two centrelines. In cases where doubt may exist, it is thus prudent to calculate the clearance at a number of points in the intersection area bounded by the shoulder lines of the two roads. As a minimum, the matrix of the points at which the clearance should be checked is shown in Figure 7.3, repeated for convenience here as Figure 8.11.

The vertical clearance selected is the sum of the height of the highest design vehicle, the minimum acceptable gap between the vehicle and the obstruction above it and whatever allowance is deemed appropriate to accommodate future rehabilitation or resurfacing projects. In most countries the highest vehicle is about 4.1 metres high. Light-rail vehicles, generally referred to as trams, particularly those powered from overhead power lines, and the equivalent double-decker bus, also known as 'trackless trams', having a height of the order of 4.6 metres. The minimum gap between the top of the design vehicle and the soffit above it is generally in the range of 0.3 to 0.6 metres and the allowance for resurfacing is

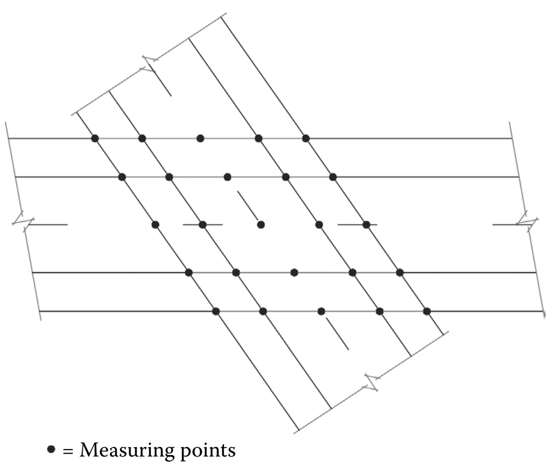


Figure 8.11 Location of check points for vertical clearance.

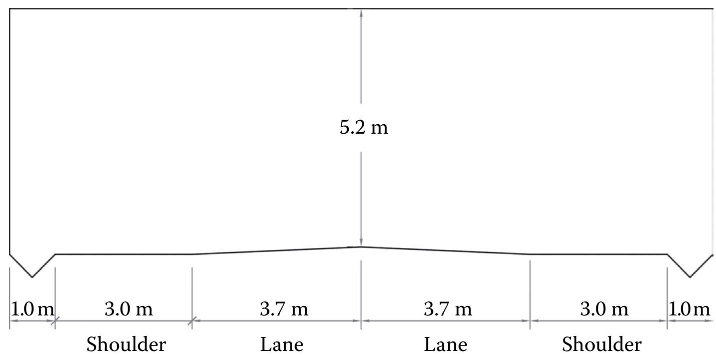


Figure 8.12 Clearance profile.

typically 0.3 metre. In consequence, the vertical clearance could fall in a range of 4.4 metres to 5.2 metres. Southern Africa has adopted the higher value and the United States a height of 4.9 metres.

Some countries have made provision for abnormal loads by declaring certain routes to be 'over-dimension load routes' (Transit New Zealand, 2003). These roads are characterised by heavier pavement designs and more generous geometric standards. For example, although the standard vertical clearance adopted by New Zealand is, like that of the United States, 4.9 metres, all new structures crossing over an over-dimension load route shall provide a minimum clearance of 6.0 metres over a carriageway that is at least 10.0 metres wide.

A typical clearance profile is shown in Figure 8.12.

8.13 TYPICAL CROSS-SECTIONS

The statement is made in Section 8.1 that cross-sections are a combination of different elements, each having their own function, dimensions and logical position within the road reserve. In spite of this, some road authorities are very inclined to lay down what the typical cross-section should be and then expect the designer to follow this prescription rigorously. This is convenient for property developers and their town planners because it immediately defines the space available for the creation of plots and hence the generation of income.

The main objection to the typical cross-section is the way in which the road authorities sometimes elevate it to the role of a law of the Medes and Persians.

However, the 'typical cross-section' should not be decried. It does have it uses, not least as a form of brief to the designer. For example, a road may have a two-lane two-way cross-section with the road authority anticipating a growth in traffic flow beyond that which the two-lane cross-section could accommodate. It is recommended that the designer and the road authority should then jointly decide on a cross-section that could accommodate most if not all of the features required of the new road. It is this cross-section that could serve as a point of departure in the design of the upgraded facility and, in fact, be the 'typical cross-section' adopted for design.

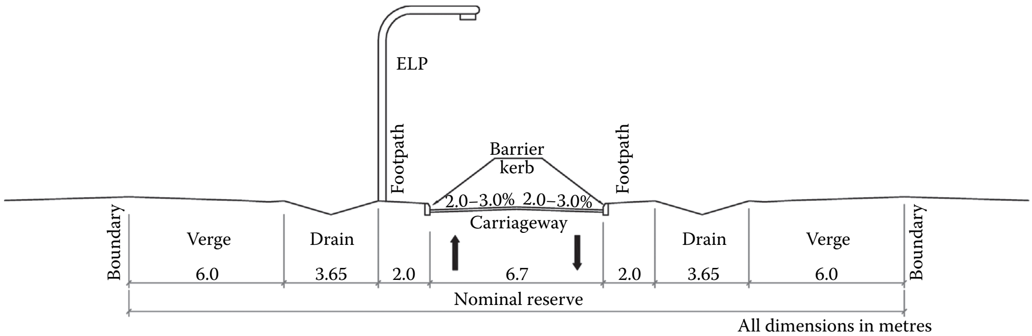


Figure 8.13 Typical two-lane cross-section.

A further application of the typical cross-section is that it is an aid to the determination of the required width of the road reserve. For example, a rural major road may comprise a two-lane two-way travelled way with surfaced shoulders and fill heights not exceeding 5 metres. For its own reasons, the road authority concerned may be of the opinion that the verges should be 5 metres wide. A typical cross-section addressing this specification could appear as shown in Figure 8.13.

Chapter 9

Aesthetics

CONTENTS

9.1 INTRODUCTION

A motorist may say that one road is 'nice' to drive on while another is not. It is most likely that a reason for this opinion would not be forthcoming. It is probably like a preference for roast potatoes. They are simply liked or not. The designer may thus be in the position of having to delve into the motorist's subconscious to establish the origin of the stated preference to establish a rationale whereby future designs could qualify for being 'nice'. However, it is reasonable to speculate that drivers' opinions are based principally on what they see and, to a lesser extent, feel. The latter refers to forces acting on the human body such as those resulting from braking, acceleration or horizontal curvature.

What drivers see can be divided into two categories: the microscopic and the macroscopic view. Both of these are aesthetic considerations and aesthetics is an important consideration in road safety in terms of the level of tension being experienced by the driver in execution of the driving task. The microscopic view is that which is visible in the immediate vicinity of the moving vehicle. It includes

- The condition of the road in terms of pot holes
- Oil stains and skid marks on the road surface
- Graffiti on and surface texture of retaining walls
- Weeds on an unsurfaced or gravelled shoulder
- Debris in the road reserve
- All those things that reflect on the quality of the maintenance of the road (Schutt et al., 2001)

It is a fact of life that drivers are more relaxed on a well-maintained road than on one that looks uncared for. And a relaxed driver is less likely to make hurried and potentially incorrect judgement calls in the driving situation than one who is tensed and totally focussed on the immediate environment to the detriment of attention on the longer range picture.

The geometric designer's role is focussed on the macroscopic view. This is the topic of this chapter, which addresses the harmony between the various elements of the road itself and its harmony with the environment in which it is placed. These are referred to as the internal and the external harmony of the design.

9.2 INTERNAL HARMONY

9.2.1 Introduction

This can perhaps be visualised as the appearance of an abstract ribbon in space. It relates to the interaction between the horizontal and vertical components of the road and the cross-section. The way in which successive horizontal and vertical curves follow or are superimposed on each other and the way in which the superelevation and its development bend the road edges in relation to the horizontal and vertical elements form a continuously flowing ribbon. This ribbon may be aesthetically pleasing or not depending on the skill brought to bear on its development.

9.2.2 Background

The Romans were the outstanding road builders of their time and achieved aesthetic road alignments characterised by long tangents and horizontal curves, and some of these alignments are still in use today. This was probably more by accident than design because the Romans needed to police an empire that extended from Hadrian's Wall in the north to the Sahara Desert in the south and from the Pillars of Hercules in the west to the Middle East. Their interest was thus purely military in nature. Aesthetics, if it arose at all, was purely a side issue.

Their design vehicles were the infantryman, by definition a pedestrian, the baggage wagon and the two-horse war chariot. It didn't seem to matter much if infantrymen tired on steep upgrades but, although not gradient sensitive, they were slow moving. Straight roads connecting camps located at a distance of a day's march, roughly 30 kilometres apart, were thus to be desired. It did matter that horses could tire, especially if ridden or driven at high speeds. This placed a limit on the maximum desirable gradient of the road. The Romans had discovered the need for the trade-off.

These ruler straight roads were, wherever possible, located on the high ground, which afforded a useful view of the surrounding territory. This was particularly important in the viewing of enemy movements and other potential threats. The legacy of the Roman approach to route location is that these roads now support the more peaceful view of a series of pleasing vistas.

The baggage wagon was also not particularly gradient sensitive because the horses pulling it could always be supplemented by the addition of slaves to achieve the required push. However, it needed long radius curvature because the pivoting front axle had yet to be invented. Although the slaves who had previously provided the motive force could be used to push the wagon from the side to assist in the change of direction, long and gradual changes of direction were necessary (Collins and Hart, 1935).

As has been suggested in the preceding text, the horizontal and, to a lesser extent, the vertical alignments are very long-lived features of a road. Apart from any aesthetic features these may possess, land use developments on either side of the road could make it impossible or at least costly to introduce later changes to the alignment. It is thus necessary for the designer to be extremely sure that the horizontal and vertical alignments are the best that can be achieved with regard not only to the whole-life economy of the road, but also in terms of the basic objectives of safety, convenience, accessibility minimum adverse effects on the environment (as discussed in Section 9.3.2) and internal aesthetics.

The internal harmony of the road has to be considered in terms of the horizontal and vertical alignments as separate entities and also in terms of the interaction between them. It is necessary to consider what can be seen by the driver because passengers in the vehicle are obviously free to consider the passing scene in addition to looking at the road ahead.

9.2.3 The driver's field of vision

The driver sees the road as if he or she were stationary and observing a 3-D movie. Peripheral vision sees nearby objects as a blur and objects in the foreground can be seen only briefly. It is only objects at a significant distance away that can be viewed at leisure. People with good eyesight can just make out an object that subtends an angle of 0° 1′. This corresponds to 3 centimetres at a distance of 100 metres. However, it is necessary to take account of people such as the elderly whose eyesight may well be less than acute. The ability of the head and eyes to swivel makes it difficult to define the driver's field of vision but the sensitive part of the retina at the centre of the visual field has an arc of 2.5°.

Five propositions applicable to the driver's field of vision were enunciated by Tunnard and Pushkarev in *Man-Made America: Chaos or Control?* (1964):

1. As speed increases the number of objects and incidents that must be reacted to increases proportionately. Observing irrelevant objects outside the necessary area of attention becomes more dangerous. Drivers would tend to focus on objects of immediate concern in the centre of the visual field. It follows that the road should aim the eye towards these objects of interest.
2. As speed increases the eyes are focussed further and further ahead. Drivers anticipate the distance at which they would have to respond to a situation demanding their attention. At a speed of 70 km/h, attention is focussed on a point 400 metres ahead and, at 120 km/h, the focal point is at a distance of 1000 metres ahead. Anything that has to be brought to the driver's attention thus needs to be large enough to be recognised at these distances and located close to the axis of vision.
3. With increasing speed and number of objects competing for the driver's attention, the driver concentrates to an increasing extent on the objects closest to the axis of vision. This is sometimes referred to as tunnel vision. The constant sounds of the tyres, the engine and the wind speed combined with the stroboscopic effect of the road markings disappearing under the bonnet of the vehicle when driving along a straight and uneventful road can cause the driver to fall into a light hypnotic trance. It then becomes possible for the driver to travel for a considerable distance while ostensibly asleep without incident. The only way to get out of the trance is either to wake up, in which case the driver might not even be aware of having nodded off, or to fall into a deep sleep. In this case, drifting off the road to the right or into the opposing lane to the left become inevitable.
4. As speed increases, features in the foreground tend to fade because the attention is focussed at a distance. At a speed of 100 km/h, objects closer than 50 to 100 metres would not be seen unless they are so demanding of attention that the driver's focus changes to see them.

5. Space perception becomes impaired with increasing speed. Small relative changes in distant objects cannot be detected. As speed increases, the time between discerning movement and passing the object decreases. The road should thus offer as many clues as possible to avoid the driver becoming desensitised to the speed actually being travelled.

These principles emphasise that, at the speeds encountered on freeways and main roads, drivers have to concentrate to survive. They must focus far enough ahead to anticipate changes in alignment and also the motion and speed of approaching vehicles.

At 120 km/h on a 30 metre wide freeway, the totality of the driver's view can be split at 30 per cent for the cross-section, 15 per cent for the roadside and 55 per cent for the sky. At lower speeds on a 15 metre wide two-lane two-way road, the cross-section takes about 15 per cent, 60 per cent for the roadside and approximately 25 per cent for the sky. It follows that a distinction must be drawn between designing the alignments of dual carriageways and two-lane two-way roads. In the case of the dual carriageway, the cross-section dominates. The concept of the abstract ribbon in space thus becomes very important in the creation of an aesthetic environment. The roadside predominates in the case of the two-lane road so that the opportunity to use the road as the basis for viewing the environment requires a design that seamlessly integrates the road with its environment.

9.2.4 The methodology of alignment design

The design of the horizontal alignment can be approached in two ways. Historically, and presumably dating back to Roman times, the tangents were located first and then linked by curves. The current approach is, in the first instance, to select the curves and then to connect them with tangents – the curvilinear approach. The belief is that this would tend to produce long curves with high values of radius and relatively short tangents whereas the tangential approach would lead to a series of long tangents clinked by short curves. This belief has nothing to recommend it and it doesn't really make any difference whether the starting point is the curve or the tangent. The end result should be the same depending on the designer's preferences.

The problem with locating the curves prior to location of the tangents is that there are simply too many variables that have to be selected more or less simultaneously. The radius should presumably have as a high a value as is practicable and the length of the curve should also be long. How far around the change of direction should the curve go? In short, what would the angle of deviation, also known as the deflection angle, be? This could dictate or be dictated by the bearing of the approach to the curve and also that of the departure from the curve. On the other hand, the bearing of the tangents could also be modified to a certain extent in support of the search for a curvilinear alignment.

If using the tangential approach, the location of the point of intersection of the two tangents is essentially automatic and dictated by the topography being traversed. The bearings of the two tangents provide the angle of deviation and selection of the highest possible radius follows. This thus tended to be a linear process with the shortcoming of the long tangent linked by proportionately short curves.

The ideal approach is to do a preliminary location of the point of intersection by determination of the location of the two tangents and then follow this by consideration of the radius of the horizontal curve. A very long radius may cause the alignment to move into deep cut or high fill in the vicinity of the centre of the curve. This could be rectified either by shortening the radius or, alternatively, shifting the point of intersection to allow for a greater external distance measured from the point of intersection to the centre of the curve. The point is that the approach to location of the centreline should be iterative rather than linear. This would achieve the best of both worlds in being a strongly directional alignment with long flowing curves. The curves and the abutting tangents could probably end up being of approximately equal length and this conclusion should be actively pursued.

9.2.5 The do's and don'ts of internal harmony

1. Aesthetically, the short curve linking long tangents could create the impression of a kink in the alignment, particularly if viewed from of some distance away whereas the longer curve produces a more flowing aesthetic appearance. It also shortens the total length of the road between the two tangent points, albeit only slightly.
2. The rolling gradeline arises from superimposing a short vertical curve on a long horizontal curve as shown in Figure 9.1.

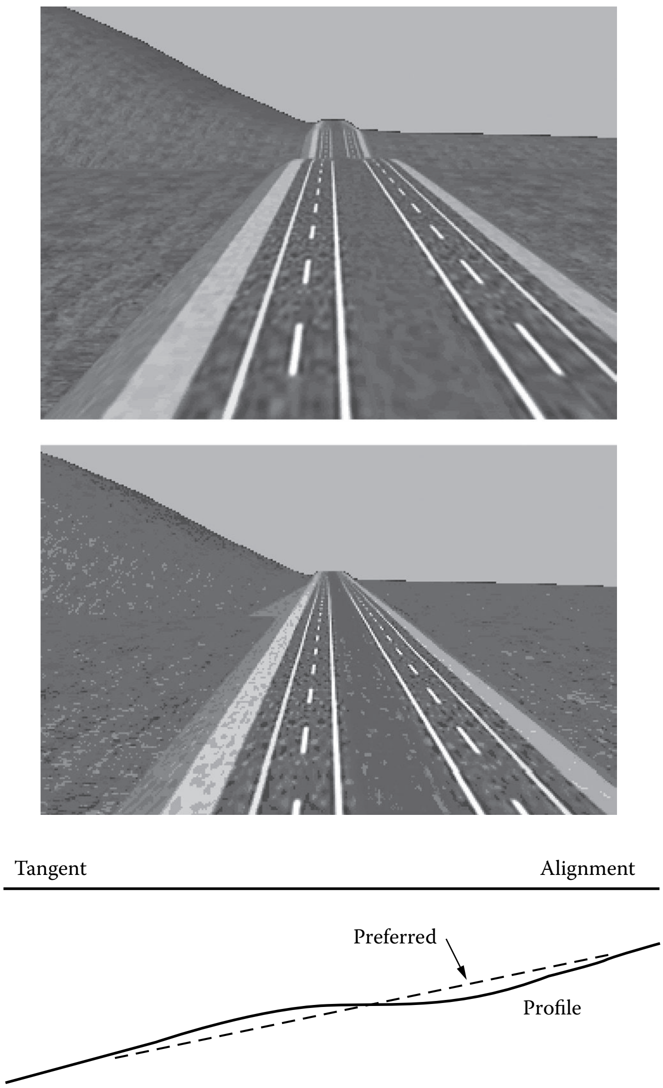


Figure 9.1 A short vertical curve on a long horizontal curve.

3. A series of short humps on a long horizontal curve, as illustrated in Figure 9.2, or tangent can arise from following the natural ground line too closely, creating the well-known roller-coaster effect.
4. It is important that there be coordination between the horizontal and vertical alignment. A local dip on a long tangent may minimise the earthworks but will be there for the life of the road and will always be unsightly, as illustrated in Figure 9.3.
5. Where the short dip is preceded by a short hump, particularly if located in advance of a long curve, the effect is even more unsightly, as illustrated in Figure 9.4.
6. The broken-back curve, arising from two horizontal curves in the same direction linked by a short tangent, is singularly ugly and also creates the impression that the designer misjudged the required extent of the change of direction. Sometimes, in restricted terrain such as on a mountain pass, elimination of the broken-back curve may be physically impossible or carry a penalty of extremely costly construction. However, where possible, the broken back should be replaced by a longer radius single curve, as illustrated in Figure 9.5.
7. A poorly located vertical curve, that is, where the vertical alignment is out of phase with the horizontal alignment, can, from a distance, create the impression of a break in the horizontal alignment. This is shown in Figure 9.6. Drivers may probably not notice this effect if their focus were towards the immediate foreground but it is undeniably unaesthetic and should be avoided if at all possible.

Figure 9.2 The roller coaster effect.

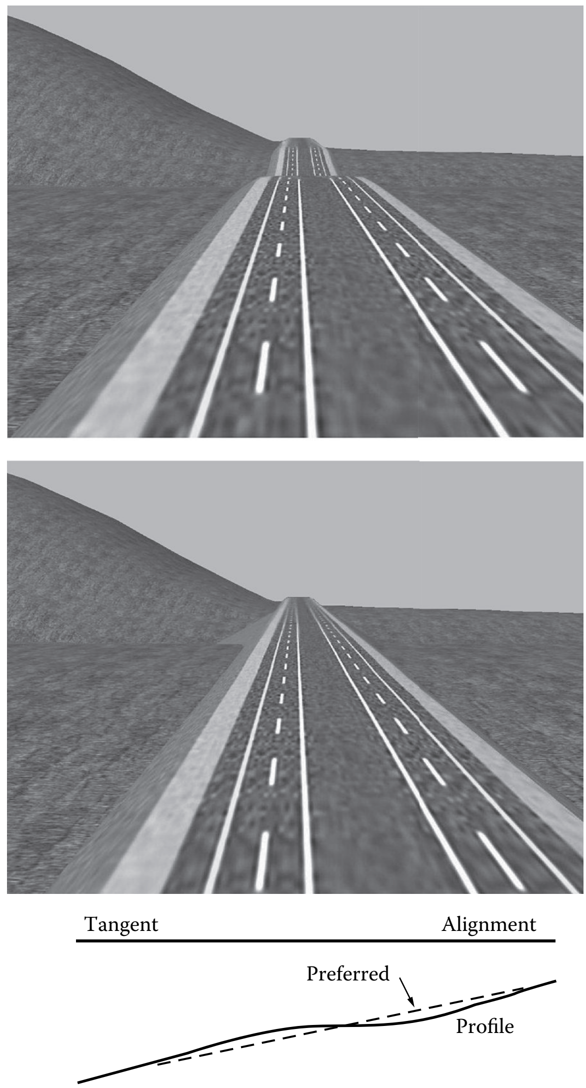


Figure 9.3 Local dip on long tangent.

8. At one time it was believed, principally for economic reasons, that river crossings had to be kept short. They were thus square across the flow of the river. This had the effect of producing a distorted alignment whereas a skew crossing, apart from being more directional, also made aesthetic sense. A comparison between the square and the skew crossing is shown in Figure 9.7.

In general, the most aesthetic alignment results when the horizontal and the vertical alignments are well coordinated. This is achieved by having the horizontal and vertical curves of similar lengths and superimposed upon each other. It is possible to leave out or add a vertical curve between to horizontal curves as long as the preceding and following horizontal and

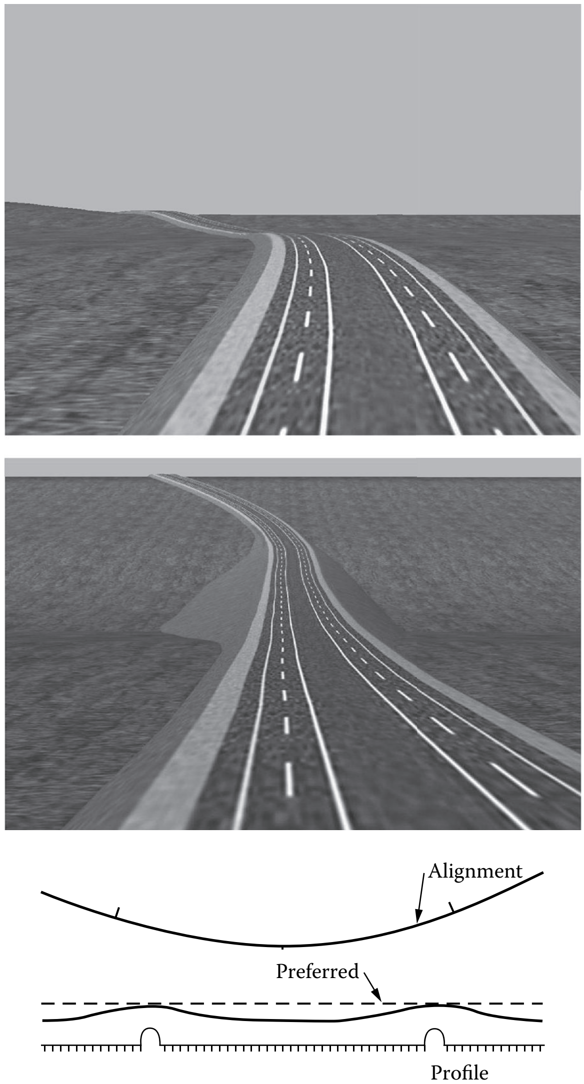


Figure 9.4 Short vertical curves in advance of a long horizontal curve.

vertical curves are superimposed. This superimposition is achieved by having the centres of the horizontal and vertical curves at a common stake value or chainage. The crest curve should be slightly led by the horizontal curve so that the driver can have advance notice of the impending change of direction, as shown in Figure 9.8. Examples of good alignment coordination are shown in Figures 9.8 and 9.9.

The curvilinear alignment illustrated in Figure 9.9 is more in keeping with the natural flow of the land form than would be an alignment predicated on a series of tangents. Although the tangential approach has its merits, it also has its aesthetics limitations. Using the tangential approach, it would be difficult to achieve the harmony between the alignment and the landscape shown in this figure.

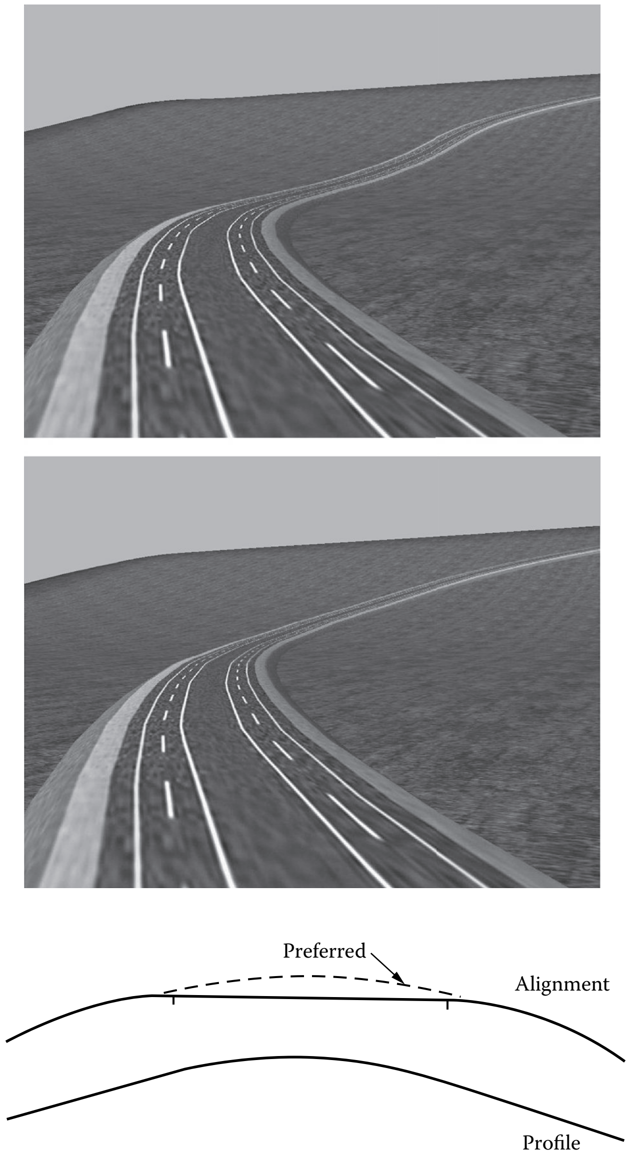


Figure 9.5 The broken-back curve.

As illustrated in the preceding photographs, it is important that the vertical alignment be in scale with the horizontal alignment. Apart from not being aesthetic as previously suggested, the local dip on a long grade is physically dangerous, as an approaching vehicle could be hidden in the dip. Thinking that the road ahead is clear, the driver may initiate an overtaking manoeuvre with potentially catastrophic consequences. Similarly, a short hump could also obscure an approaching vehicle from view. Both situations, in addition to being dangerous, are unsightly. The two in combination, that is, the short hump and dip preceding a long curve, is particularly unsightly.

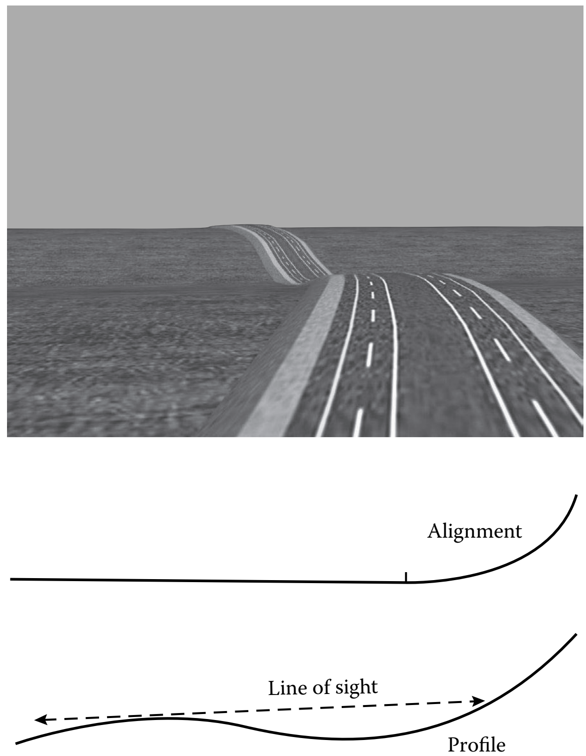


Figure 9.6 A break in the horizontal alignment.

9.3 EXTERNAL HARMONY

9.3.1 Introduction

Internal harmony can be achieved by abiding by the few simple suggestions provided earlier. Furthermore, its approach is focussed on the driver's view of the road ahead. External harmony, being the interaction between the road and its environment, is more difficult to achieve. The driver would only be able to treat the environment to the odd passing glance unless pulled off the road into a rest area, whereas the passengers' view could include a scrutiny of the environment. This would be the view from the road. The view of the road from outside the road reserve could attract vociferous and usually negative comments from local residents. The Shakespearian 'beauty lying in the eye of the beholder' suggests that numerous beholders, viewing the road from a variety of vantage points and with differing points of view regarding what constitutes a beautiful road, all have to be satisfied. A further complication is that they all are comparing the view that is with the view that was with, usually, a strong bias in favour of what was.

Creating an aesthetic road needs skills additional to those of the discipline of geometric design. These skills lie more in the fields of architecture and landscape engineering, which is also architectural in its approach but on a significantly larger scale than a focus on buildings and other structures. This further stresses the need for a multidisciplinary approach to geometric design. The document *Beautiful Roads – A Handbook of Road Architecture*

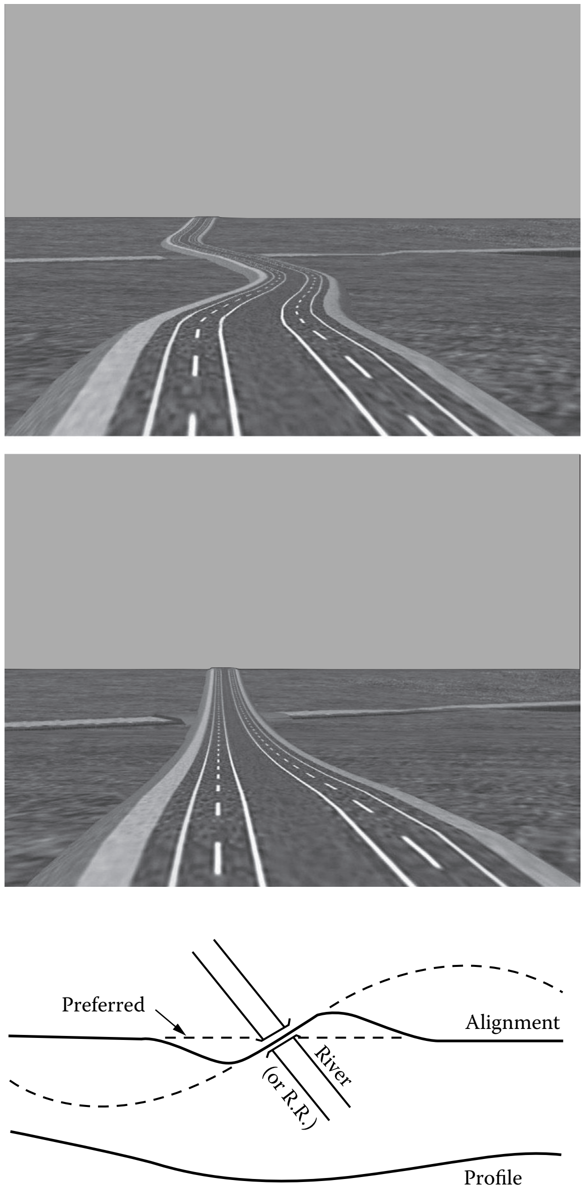


Figure 9.7 The skew versus the square river crossing.

published by the Danish Road Directorate (Egebjerg, 2002) richly captures the complexity of road architecture and is recommended reading for all who would venture into this field.

Concern for the environment includes attention to

- Air and water pollution
- Noise
- Environmental degradation including fragmentation of habits
- Very importantly, visual pollution

Figure 9.8 The horizontal curve leading the crest curve. (Photo courtesy of RC Burrell.)

Figure 9.9 Landscaped curvilinear alignment with stream in variable width median. (Photo courtesy of RC Burrell.)

It follows that visual pollution must be addressed as part of the Environmental Impact Assessment (EIA) and (in the United States) Environmental Impact Statement. There should thus be a full Visual Impact Assessment (VIA) study that can be referred to by the EIA.

9.3.2 The Visual Impact Assessment

There is a strong connection between a community's values and the environment. Visual quality is not an absolute that can be built into the environment. It is, in fact, necessary to recognise that visual quality is the perception that the community as a whole has of the environment in which it finds itself. It is thus necessary that the community be directly involved at an early stage in defining how it understands visual quality (Churchward et al., 2013). The community must also define visual impacts that are important to it. This definition is required to be restricted to the area that would be visible from the road and also to external views that would include the road. Analogous to the watershed of hydrology, this area is usually known as the 'viewshed' of the proposed project.

VIAs are normally undertaken by professionals, who frequently but not invariably are landscape architects. What the professionals consider important may not be important at all to the local community, assuming they can actually see it. An exclusively professional assessment thus runs the risk of not addressing the needs and quality judgments of the community.

Best practices for the conducting of a VIA have been identified (Churchward et al., 2013) and these include

- Conduct an inventory of the existing landscape and its physical attributes, which include
  - Landform relief
  - Vegetation, specifically woodland presence, area and configuration
  - Presence and configuration of bodies of water
  - Apparent natural appearance of land uses or character of the built environment
  - Length or extent of view
  - Visibly flowering plants
  - Apparent maintenance as shown by a neat landscape and tidy farming practices.
- Identify whose views will be affected by the proposed project; with viewers subdivided into
  - People on properties adjacent to the road, that is, 'neighbours' who have views *to* the road and are frequently subdivided by land use.
  - People on the road, that is, 'travellers' who have views *from* the road and are frequent subdivided by mode of travel (car, truck, bicycle, etc.) or reason for travel (commuting, freight haulage, recreation).
- Identify key views, which may be either representative or iconic and that can be used to analyse visual quality and visual impacts with the viewing public assisting in the selection of these views as part of a Visual Quality Plan. The application of key views is in generating 'before and after' images defining the impact of the project on the environment.
- Determine the visual impact of the project, which is a comparison between two future states – one with the proposed project and one without, effectively the 'no-build' visual quality for a selected future date. The starting point of this comparison is an assessment of the current visual quality. This visual quality is that placed on the existing landscape by people who currently have views of the existing landscape.

In summary, the alternatives to be compared and documented include

- The effect of the project on the affected environment
- The effect of the project on the affected population
- The impact on visual quality by the project

The prime objective of any EIA is to quantify the impact that a project will have on the environment and then to identify strategies aimed at avoiding, reducing or minimising negative impacts or, if none of these options are viable, to determine the extent of compensation payable and to whom. As the Visual Quality Assessment (VQA) is a subset of the EIA, it follows that it shares this objective and definition of ameliorating strategies.

Although the focus has been on seeking to avoid negative impacts of a project, the VQA could also identify the possibility of positive impacts. It may be possible to use these as a trade-off against negative impacts elsewhere along the project or perhaps even to cost them and add this sum to the cost of the project to the greater benefit of the community at large. Most authorities would, however, probably accept possible improvements but without any commitment to the cost implications involved in achieving them.

Chapter 10

Intersections

CONTENTS

10.1 INTRODUCTION

Road networks comprise links and nodes. The nodes are the points at which links (highways, roads or streets) cross. The three general types of crossings are at-grade, grade separations without ramps and grade separations with ramps – otherwise interchanges. They are referred to in most countries as intersections, differentiating between three-legged, four-legged and multi-legged intersections although, in the United Kingdom, reference is to three-, four- and multi-armed junctions. In this book, reference is to 'intersections' and 'legs'.

The distinction drawn between interrupted and uninterrupted flow should be understood. Interruption refers to stops resulting from traffic control measures such as signalisation of intersections or the various forms of priority control at intersections. The delay they cause is often referred to as 'geometric delay' as a differentiation between them and delays caused by congestion and poor levels of service (LOS). It follows that an urban arterial operating under green wave with no stoppages in traffic flow is said to be operating under interrupted flow conditions, whereas a freeway that, in the peak hour, becomes an extended parking lot is experiencing uninterrupted flow. 'Rush hour' is often a total misnomer.

In this chapter, aspects of design such as the general principles governing intersection design, the form of control applied to intersections, the channelisation of the various flows through intersections and two special cases of traffic control – roundabouts and signalisation – are discussed.

10.2 HUMAN ERROR

The three causes of crashes are human error, mechanical malfunction of the vehicle and poor condition of the road. The second and third of these are outside the scope or terms of reference of the geometric designer. As safety is one of the prime concerns of the designer, it behoves him or her to reduce the likelihood of human error as far as possible and to minimise the consequences of the human errors that still occur.

Human error can be that of various individuals being drivers or occupants other than the driver of vehicles, pedestrians or cyclists. It is the driver's responsibility to ensure that the actions of the other occupants of the vehicle do not cause distraction potentially interfering with the driving task. The safety of pedestrians can be ensured by providing them with

- Sidewalks that are removed as far as possible from the travelled way
- Clearly demarcated and signposted pedestrian crossings
- Education in road safety

Similarly, cyclists provided with dedicated lanes or paths that are removed from the travelled way will be safer than those who are required to share the lanes of the travelled way with other vehicles. Education in road safety and the rules of the road is obviously a prerequisite for the safe operation of a bicycle.

The area where the designer can play a major role is in minimising the circumstances in which driver error is likely to occur. Driver error is frequently cited as being the predominant proximate cause of crashes, and various percentages of cause are quoted in the literature ranging from 75 per cent to more than 90 per cent.

Errors can include simple inattention to the point where, on a long tangent, a driver can fall asleep through sheer boredom. This is actually a light hypnotic trance as opposed to normal sleep and it has been theorised that the stroboscopic effort of the road markings disappearing under the bonnet (hood) of the vehicle in combination with the constant engine and tyre sounds is, at least partially, to blame for this phenomenon. This theory explains why drivers can control their vehicles for long distances and even negotiate signalised intersections while apparently asleep.

Failure of reaction time resulting from tiredness or substance abuse, and so forth, and incorrect response to a situation are also potential sources of error. The last-mentioned could possibly be because of information overload or misreading the information read from the road. The crash black spot, where numerous drivers make the same mistake at the same place, is a well-known phenomenon. It is quite possible that this arises from a design error or an optical illusion.

10.3 CONFLICT RESOLUTION

Run-off-the-road (ROR) crashes are often single-vehicle crashes and can result from the driver losing control of the vehicle on a slippery surface or horizontal curve. Where more than one vehicle is involved, conflict between the travel paths of the various vehicles would often be the cause of crashes. The majority of conflicts occur at intersections and the design and operation of intersections are thus exercises in conflict resolution.

The conflict is between vehicles competing for right-of-way through the intersection area. It can also be between vehicles and pedestrians or between cyclists and pedestrians. The conflict can be resolved by (Florida Department of Transportation, 2007)

- Rules of the road where, at an uncontrolled intersection, right-of-way is granted to the first vehicle arriving at the intersection or, in the case of simultaneously arrival, to the vehicle on the left
- In the case of a multi-leg (e.g. four-way stop) stop, right-of-way is accorded similarly except that vehicles are required to come to a complete stop, which is not the case at uncontrolled intersections

- Rules of the road apply at all intersections, where left-turning vehicles have to yield right of way to vehicles moving straight through the intersection
- Fixed priority where vehicles on the minor legs of an intersection are controlled by a Yield or a Stop sign
- Alternating priority where priority is switched between opposing movements by traffic signals
- Weaving with the onus on the drivers to change lanes in a safe manner
- Roundabouts, where conflict resolution is, in principle, similar to that at uncontrolled intersections except that crossing movements are at flat angles and low speeds
- Grade separations where the spatial separation between conflicting traffic streams automatically resolves the problem of the temporal separation between them

10.4 FORMS OF TRAFFIC CONTROL AT INTERSECTIONS

One of the first steps in the process of intersection design is to select the most appropriate form of control, as this will have a significant role in the layout of the intersection. Apart from obvious concerns of safety, a logical basis for selection of the type of control is efficiency of operation of the intersection. Delay is a measure of efficiency, as time wasted has economic consequences. Vehicles standing with their engines idling also have economic consequences in terms of wear and tear (and possible overheating) on their engines and fuel consumption. It follows that delay should be minimised as far as possible.

In Figure 10.1, the extent of delay experienced at various forms of control and for a range of traffic flows on the major and minor legs of intersections are shown (PWV Consortium, 2002). These were modelled using computer simulations, specifically SIDRA and SIMTRA

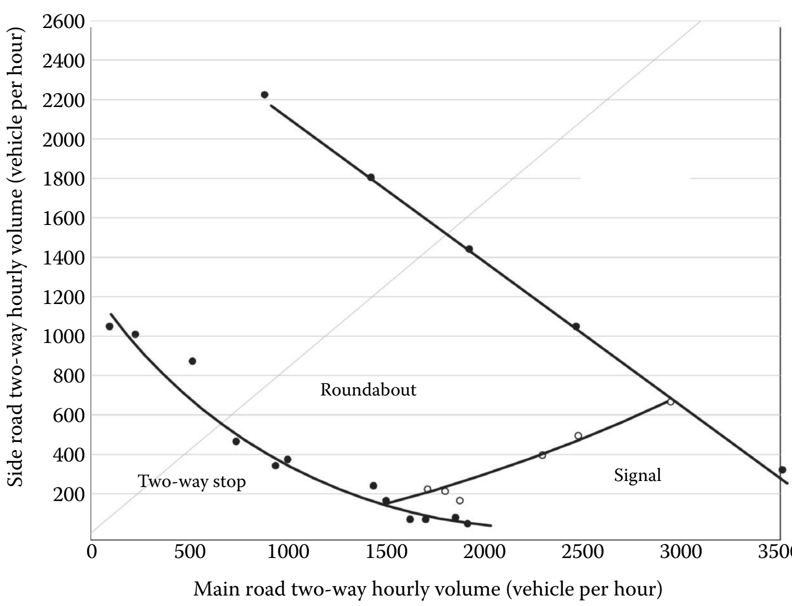


Figure 10.1 Volume range over which the indicated form of control minimises delay. (From Gauteng Department of Public Transport, Roads and Works. Road Design Manual: Volume 1: Geometrics, Johannesburg, 2007.)

developed by Professor van As of Pretoria University. In total, about 6500 different configurations, arrival patterns and volume combinations were tested.

It is to be noted that the arrival pattern has a significant influence only at signal-controlled intersections. The priority control cases, Stop, Yield and Roundabout, are insensitive to arrival patterns and were simulated under random arrival conditions only. As far as possible, 'typical' conditions were simulated, which are considered to be as follows:

- The major road was set to have a 60:40 split.
- The major road two-way volumes were increased at 200 vehicle per hour intervals from 200 to 2400 vehicles per hour.
- The side road was assumed to have a 67:33 split.
- Side road two-way volumes were increased at 75 vehicle per hour intervals from 75 to 1050 vehicles per hour.
- Main road turning movements were set at 15 per cent, 70 per cent, 15 per cent left, through and right respectively.
- Side road turning movements were set at 40 per cent, 20 per cent, 40 per cent left, through and right respectively.

10.5 DESIGN PRINCIPLES

10.5.1 Angle of skew

Elderly drivers often have reduced flexibility of their necks and sometimes also suffer a loss of peripheral vision. If the angle of skew between the major and the minor roads is too far off the ideal of a 90° cross, these drivers would, if on the minor road, have difficulty in seeing opposing vehicles. Drivers of trucks may also experience difficulties caused by the bodywork of their vehicle obscuring their line of sight to vehicles approaching from the right. It is generally accepted that the angle of skew should not be outside the range of 60° to 120° as an absolute maximum and between 75° and 105° as a desirable maximum.

Acceptable angles of skew are illustrated in Figure 10.2.

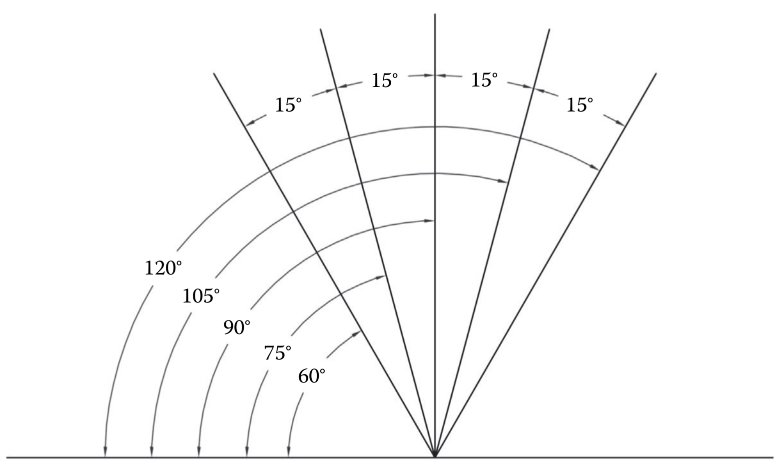


Figure 10.2 Acceptable angles of skew.

10.5.2 Gap acceptance

Stop and Yield control are the most commonly applied forms of priority control at intersections. Their successful operation relies on a driver on the minor leg(s) of an intersection being able to assess a gap between vehicles on the major road and to take a decision on whether to accept or reject it. If a vehicle on a minor leg arrives at the intersection partway through a gap between vehicles on the major road, reference is then not to a gap but to a lag. In short, a lag is the unexpired portion of a gap.

The general approach to gap acceptance is that, for design purposes, the critical length of gap is that which the 85th percentile driver would accept (van As and Joubert, 1993). The value of this gap does not lend itself to easy measurement. The problem is that the driver makes a selection from gaps *that are on offer*. If traffic flows are such that the minimum acceptable gap hardly ever appears it follows that conclusions drawn are likely to be skewed. Furthermore, drivers will often reject gaps only to accept a gap shorter than those already rejected. Gaps acceptable to commuters may be shorter than those the same drivers would accept over weekends.

The first work done regarding the size of the critical gap was by Harmelink, who was concerned primarily with the development of a warrant for left-turn lanes at priority controlled intersections (Harmelink, 1967). He found that the critical gap should be of the order of 5.0 seconds, which is near the values found in recent research. If a more conservative gap value for use in design is desired, then the critical gap value should be increased to 5.5 or 6.0 seconds. For trucks, a higher critical gap should be considered, generally of the order of 0.5 to 1.0 second increase in the value assumed for passenger cars (Fitzpatrick and Wolff, 2003). The values of gap acceptance now adopted in the calculation of intersection sight distance (AASHTO, 2011a) are more generous, at 7.5 seconds for a passenger car and 9 and 11 seconds for trucks and semitrailers respectively.

Prior to this human factors approach to the operation of a priority-controlled intersection, an elaborate model of intersection sight distance was in vogue. This drew a distinction between Yield and Stop control in the determination of the distance between the driver's position and the stop/yield line at the time when a decision whether to stop or to go had to be made. This was in terms of the approach speed for Yield control and for a dead stop for Stop control. Various assumptions were then required regarding

- The length of time it would take the driver to start the vehicle from rest
- Whether the driver intended turning to the left or to the right or to proceed across the intersection
- In the last-mentioned case, the distance that the vehicle had to travel to clear the intersection included not only the width of the major road but also consideration also of the length of the crossing vehicle
- The speeds of the turns to the left or the right
- The rate of acceleration of the vehicle having completed the turn
- The rate of deceleration applied to a vehicle on the major road to ensure that it assumes a station behind the turned vehicle corresponding to a headway of, say, 2 seconds at the speed then achieved by the leading vehicle

It is a matter of some surprise that fair uniformity of practice regarding intersection sight distance was ever achieved.

10.5.3 Conflict reduction

Every conflict point in an intersection is a point at which a crash may occur. It follows that reducing the number of conflict points would have an effect on the overall safety of the intersection. A three-legged intersection has 9 conflict points and a four-legged intersection has 24. A five-legged intersection has 120 conflict points. Some diagrams in the literature also include diverge conflicts, although it is difficult to imagine these as conflicts in the strict sense of the word. A roundabout has 4 conflict points excluding the diverge conflicts. It is thus reasonable to assume that this is the safest of all the forms of intersection, and this assumption is borne out by the statistics.

Apart from the sheer number of conflicts, the more conflict points there are in a confined space the closer together they will be. The rate at which drivers would have to take a decision would thus increase until the point is reached where they simply cannot cope anymore. The consequence of this is that they either simply stop in the intersection area or hope that other drivers will allow them a gap to proceed. The probability of crashes occurring could then become very high.

10.6 GENERAL CONTROLS FOR INTERSECTION DESIGN

10.6.1 Introduction

The most obvious requirements for a safe intersection design is that drivers should have adequate sight distance in the approach to the intersection and sufficient time while within the area of the intersection to observe what is going on around them and to decide on their own appropriate course of action.

The shape of the road surface is defined by the combination of the horizontal and vertical alignment and the cross-section. Where two roads intersect or cross, it follows that the resulting surface will be very complex, with the possibility of local low or flat spots occurring. These could create ponding on the road surface leading to hydroplaning and the loss of control of vehicles in the intersection. In any circumstance, hydroplaning is dangerous but, in intersections where other vehicles are also moving around generally in close proximity to one another, it is doubly dangerous. Furthermore, a vehicle moving through a pond could throw water up against the windscreen of another vehicle, with a consequent loss of visibility to the driver of that vehicle. Identifying problem areas of drainage is not easy but is essential for the safety of road users.

The physical location of the intersection relative to other intersections and the alignment, both horizontal and vertical, will have an impact on the safety of operation within the intersection.

The issues of sight distance, drainage and the location of intersections are discussed in this section.

10.6.2 Intersection sight distance

Intersection sight distance (ISD), previously called shoulder sight distance, is measured from a driver eye height of 1.05 metres to an object height of 1.15 metres as discussed in Chapter 5. The operation of intersections relies on gap acceptance and it is this that determines the length of the legs of the sight triangles.

10.6.2.1 Sight triangles

It is a sine qua non that stopping sight distance is provided at all points along the road. At intersections, a greater extent of sight distance is necessary because the driver should have an unobstructed view of the entire intersection and any potentially conflicting vehicles before entering it. Sight distance is thus not merely along the length of the road but to the side as well. In short, what is required is a sight triangle, defined as the area clear of any obstructions that might block the view of potentially conflicting vehicles. Two forms of sight triangle have to be considered: approach sight triangles and departure sight triangles (AASHTO, 2011a) as illustrated in Figure 10.3. The approach sight triangle refers to intersections subject to Yield control, specifically that required by a driver on a minor leg of an intersection to be able to see opposing vehicles while approaching the intersection. The departure sight triangle refers intersections subject to Stop control or where traffic on the major road forced a stop on vehicles at a Yield controlled intersection. The driver at the Stop line on the minor road must be able to identify a gap in the traffic on the major road before deciding whether to enter the intersection.

The extent of intersection sight distance required is dependent on design speed and the type of control provided. Control types are (AASHTO, 2011a)

- Case A – No control other than rule of the road
- Case B – Stop control on the minor legs of the intersection

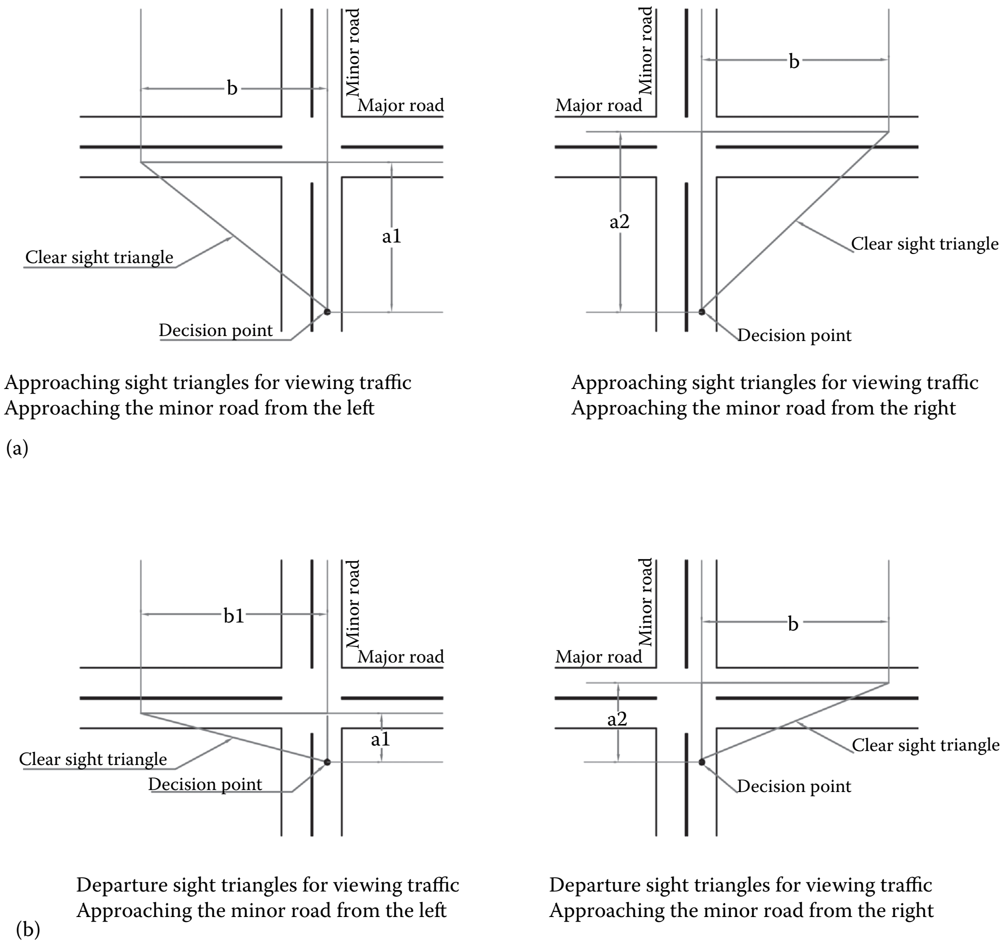


Figure 10.3 Sight triangles: (a) approaching sight triangles (uncontrolled or yield-controlled) and (b) departure sight triangles (stop-controlled). (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

- Case C – Yield control on the minor road
- Case D – Signal control
- Case E – All-way stop control
- Case F – Left turn from major road

Cases B and C draw a distinction between left turns, right turns and crossing manoeuvres from the minor road.

The distance from the major road along the minor road to the point at which the driver on the minor road should begin to brake is shown in Figure 10.2 as being a_1 to the right and a_2 to the left. Distance a_2 is equal to a_1 plus the width of the lanes departing on the major road to the right. Both distances are measured from the centre of the approach lane on the minor road to the centre of the lane of interest on the major road. The distance a_2 also includes the width of a median if present. If the median is sufficiently wide to provide refuge to the vehicle turning to the left from the minor road, intersection sight distance to the left as for Case B1 should be provided, with the dimension a_2 being 3.0 to 3.5 metres.

Distance a_1 is shown in Table 10.1.

In the case of an uncontrolled intersection, this distance applies also to the major road, as this form of control does not differentiate between major and minor roads in the strict sense of the word.

Distance b is shown in Table 10.2 in respect of Case B operation for turns to the left and the right and the crossing manoeuvre (i.e. subject to Stop control) and also turns to the left and right for Class C operation (i.e. Yield control). The length of the sight triangle leg along the major road in respect of the Class C crossing manoeuvre (from Yield control) is given in Table 10.3.

10.6.3 Drainage of intersections

For reasons of safety, it is essential that water should not pond on the road within the intersection area. Low points should therefore not occur anywhere other than at the edges of the intersecting roads. Where the edges are demarcated by kerbs, drop inlets routing storm

Table 10.1 Length of sight distance triangle leg Case A

Design speed (km/h)	*Length of leg (m)*
40	35
50	45
60	55
70	65
80	75
90	90
100	105
110	120
120	135
130	150

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.

Note: For approach gradients greater than 3%, multiply the sight distance values in this table by the appropriate adjustment factor from Table 9.4.

Table 10.2 Length of the sight distance triangle leg on the minor road Case B

Design speed (km/h)	*Case B1: Left turn from stop*	*Case B2: Right turn or cross from stop*	*Case C2: Left of right turn from yield*
40	85	75	90
50	105	95	115
60	130	110	135
70	150	130	160
80	170	145	180
90	190	165	205
100	210	185	225
110	230	200	245
120	255	220	270
130	275	235	290

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.

Table 10.3 Length of major road leg for Class C crossing from yield

Major road design speed (km/h)	*Major road leg lengths for minor road design speeds (km/h) of*					
	30–80	*90*	*100*	*110*	*12*	*130*
40	75	80	80	85	90	90
50	95	95	100	105	110	115
60	110	115	120	125	130	135
70	130	135	140	145	150	160
80	145	155	160	165	175	180
90	165	175	180	190	195	205
100	185	190	200	210	215	225
110	200	210	220	230	240	245
120	220	230	240	250	260	270
130	235	250	260	270	280	290

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.

water into subsurface drains have to be provided. In the absence of kerbing the shoulders should be warped to ensure that water that has reached the edge of the travelled way is provided with a flow path across the shoulder.

It is not possible to calculate where flat and low spots would occur and the solution is to be found by graphical grading. The first step in the process is that contour lines showing the surfaces of the intersection roads have to be drawn through the intersection area as though the opposing road were not there. These contours arise from the camber or crossfall of each road as well as their gradients. The contours then have to be edited to reflect a compromise between them, resulting in a smooth albeit warped continuous surface and not a series of spot heights across a jagged surface.

It must be borne in mind that, in terms of intensity of rainfall, an intersection has a small surface area so that the intensity could be very high. The likelihood of a storm being limited to the confines of the intersection area is very limited, but a localised shower could have a high intensity. The runoff from a surfaced road is 100 per cent. It follows that the proper drainage of an intersection is an important contribution to road safety.

10.6.4 Spacing of urban intersections

In urban areas, intersections on arterials should be spaced such that traffic flows along them are not unnecessarily constrained. The ideal is that flow should be at green wave conditions. These are the situations whereby traffic signals change to their green phase as approaching platoons of vehicles reach them. In the case of one-way streets, this is very easy to achieve, as the start points of the green phase of traffic signals at successive intersections need to be offset from those at the previous intersections only by the travel time between them. The distance between successive intersections is thus irrelevant. On two-way streets, optimisation of the signal settings is applied to the heavier of the two flows along the street. The lighter flows simply have to accept whatever phasing is made available to them. The morning peak is from outlying residential areas towards the central business district (CBD) and the afternoon peak is in the reverse direction. Traffic signal control has achieved a very high level of sophistication, but even the simplest electromechanical controller can accommodate more than one signal plan. There are usually three signal plans installed in the controllers of fixed time signals, for morning inbound peak traffic, evening outbound peak traffic and off-peak traffic, where the flows are of similar magnitude in both direction.

It is possible that green wave flow can, in the case of two-way roads, be achieved simultaneously in both directions. This requires that the spacing of successive intersections is such that the travel time between them at the design speed is equal to the sum of the durations of the yellow and red phases and the all-red phase. With a cycle length of 60 seconds, these phases typically would total about 30 seconds. The spacing between successive intersections should thus be of the order of 500 to 700 metres for design speeds ranging from 60 km/h to 80 km/h to achieve bidirectional green wave operation. Bidirectional green wave flow is a special case, seldom likely to be realised in practice. As stated previously, signals are set to favour the peak flow and the reverse flow simply has to accept whatever constraint this may have on its progression through the series of signals. It is pointed out that signalising flows moving at speeds higher than 80 km/h is not wise, as inattentive drivers may suddenly find themselves in the position of not being able to stop in time at a yellow or red signal.

Intersections on collectors can be at shorter intervals. The spacing of intersections on collectors is determined to large extent by the layout of the access roads serving individual properties. It is, however, suggested that collectors very often serve as bus routes and this function also has a bearing on the location of intersections. For bus services to be viable, they have to be attractive to would-be passengers. Attractiveness is related to many features such as cost, frequency and regularity of service, comfort and accessibility. Accessibility implies, inter alia, the distance that passenger have to walk to reach the nearest bus stop. Assuming that a walk of 10 to 15 minutes duration at a speed of 1.5 m/s would be acceptable to the average commuter, this would represent a maximum distance of 900 to 1350 metres between home and bus stop. In general, the spacing between successive bus stops should be about half of this distance to allow for walking along the cross-street. According to Vuchic (1981), spacings should be of the order of 400 to 600 metres. Ideally, bus stops are located at intersections to shorten walking distances as far as possible. From a town planning perspective this suggests that intersections on bus routes should, if possible, follow the spacing of the bus stops and thus also be located at 400- to 600-metre intervals.

10.7 FUNCTIONAL AREAS OF INTERSECTIONS

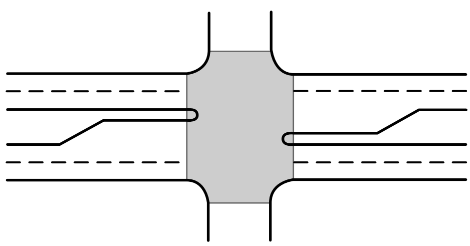
In the design of intersections, two forms of area are of interest. In the first instance, the physical area encompassed by the intersection is important, as this ultimately defines the reserve area that has to be acquired for its construction. In the second instance, the functional area extends up- and downstream of the intersection as influenced by the need of drivers to decelerate upstream of the intersection or accelerate downstream of it, whether or not auxiliary lanes and the need to change lanes into them are present.

The minimum spacing of intersections is determined by the fact that the functional distances of two successive intersections desirably should not overlap. Drivers are required to take numerous decisions and take several actions in the functional area of an intersection. Any overlap between the functional areas of successive intersections would probably generate a significant increase in driver's workload, resulting in a heightened risk of crashes. It would also be inimical to smooth traffic operation.

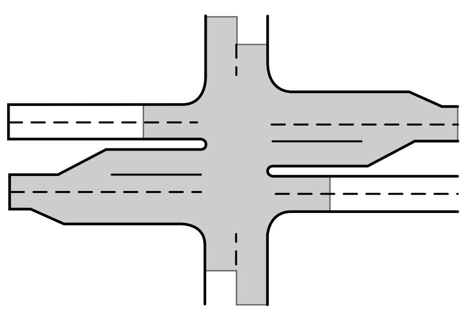
The upstream functional area of an intersection comprises three distinct phases. The first is the distance, d_1, travelled during reaction time when, in addition to a decision to stop or to proceed through the intersection without stopping, the driver may also be considering the need to change lanes in anticipation of a turning movement. The second phase is the distance, d_2, required to execute any required lane change manoeuvres and the third phase is the distance, d_3, required to decelerate to a stop. It is necessary to provide storage space, d_4, beyond the end of the third phase to allow for the maximum number of vehicles likely to arrive at the intersection at any one time. The upstream functional area is illustrated in Figure 10.4.

The derivation of the desirable length of the upstream functional area is based on:

- Distance d_1 being calculated as the distance traversed at the design speed during a reaction time of 2.5 seconds
- Distance d_2 is the length of the passive taper defining a lane change to the left or to the right typically with a taper rate of 1:20
- Distance d_3 is the distance traversed during deceleration from the design speed to a stop at a rate of 3 m/s^2



Defined by physical area



Defined by functional intersection area

Figure 10.4 The physical and functional areas of an intersection. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

Distance d_4 is not shown in the calculation and must be added during design to the length shown in Table 10.4. Distance d_4 is site specific and dependent on the number of vehicles that would probably need to be stored prior to their executing the turn. This is derived as discussed in Chapters 17 and 18.

The downstream functional area comprises two phases, the first being where vehicles accelerate from a dead stop to the operating or posted speed of the road and the second where they merge either with turning vehicles or at the end of auxiliary through lanes. In the case of the right turn, the lane drop is adjacent to the shoulder. A driver unable to complete the merging manoeuvre would thus be able to escape onto the shoulder. The left turn is more problematic, as the acceleration lane will precede a merge into the inside or fast lane.

Rates of acceleration vary according to the power of the vehicle in combination with the preferences of the driver. The maximum acceleration that a passenger car is capable of is, for a wide range of cars from a subcompact to a large saloon, of the order of 3.5 m/s^2 for speeds of 60 km/h or less, thereafter diminishing along a hyperbolic curve to zero at the maximum speed of which the vehicle is capable. Although this is an indication of what vehicles can do, the human factors approach suggests that it is what drivers wish to do that is more important. Driver preference factors are thus applied to the maximum vehicle performance. These factors have been measured to fall in a range of highly aggressive driving with a factor of 1.0 to 0.85 to a very nonaggressive value of 0.4. A factor of 1.0 represents the maximum that the vehicle is capable of. Most drivers' preferences seem to fall into the boundary range of 0.85, corresponding to an acceleration rate of about 3.0 m/s^2 (Snare, 2002) matching, in fact, the desired rate of deceleration. The distances required to achieve various design speeds from rest are shown in Table 10.5. And added to these is the desired length of active taper at a rate of 1:50 with an assumed lane width of 3.7 metres. The desirable length of the downstream functional area is shown in Table 10.5.

The stopping sight distance is sometimes considered to be an indication of the desirable length of the downstream functional area. The rationale behind this is not understood but this, too, is shown in Table 10.5. The 1:50 taper length should be added to the Stopping Sight Distance to establish the length of the downstream functional area, in which case it is longer than the functional area based on the more logical acceleration rate.

The acceleration rates of trucks are much lower than those appropriate to passenger cars being, typically, in the range of 0.3 to 0.35 m/s^2. Where trucks are present in significant numbers it may be necessary to increase the length of the downstream functional area to 100 or 180 metres respectively to enable them to reach speeds of 60 or 80 km/h (Gattis et al.,

Table 10.4 Desirable length of upstream functional area

Design speed (km/h)	*Reaction distance* d_1 *(m)*	*Passive taper length (m)*	*Length of deceleration lane (m)*	*Total length (m) of upstream functional area*
60	42	74	289	405
70	49	74	394	516
80	56	74	514	644
90	63	74	651	788
100	69	74	804	947
110	76	74	973	1123
120	83	74	1157	1315

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.

Table 10.5 Desirable length of downstream functional area

Design speed		*Stopping sight*	*Distance for*	*Taper length*	*Total length of*
km/h	*m/s*	*distance (m)*	*acceleration (m)*	*(m) at 1:50*	*functional area (m)*
60	16	90	46	185	231
70	19	110	63	185	248
80	22	140	82	185	267
90	25	170	104	185	289
100	27	200	129	185	314
110	31	230	156	185	341
120	33	270	185	185	370

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.

2008). The 185 metre long tapers have to be added to these distances to provide the length of the downstream functional area.

The sum of the downstream functional area of an intersection and the upstream functional area of the following intersection is the desirable minimum acceptable distance between them. For a design speed of 60 km/h the spacing should thus be of the order of 640 metres and 760 metres for 70 km/h to accommodate truck performance.

10.8 CONSTRAINTS ON THE LOCATION OF RURAL INTERSECTIONS

In rural areas, intersections are spaced at anything up to several kilometres apart so that the prime influences on their location are more those of horizontal and vertical alignment and topography than of operational issues.

Intersections should, ideally, be located on tangents. If located on curves, drivers entering the major road from a minor road on the inside of the curve could have the problem of the sightlines being behind them. If the intersection is skewed, it is possible that the curve may result in the line of sight being outside the desirable range illustrated in Figure 10.2. This is particularly the case where the radius of the curve is close to the minimum for the design speed. The problem can be minimised by relocating the minor road to be at right angles to a tangent to the curve at the point of intersection.

A more critical problem is that of drivers entering from the outside of the curve. They have the advantage of having their sightlines in front of them. The additional driver eye height created by the superelevation also significantly increases the available sight distance. But they have to turn against the superelevation onto the major road. The driver of a truck with a high load would face the upcoming manoeuvre onto a 10 per cent adverse superelevation with a serious degree of trepidation.

It is not always possible to avoid locating the intersection on a curve, and the adverse superelevation confronting drivers entering from the outside of the curve should be checked. For preference, this should not be more than 6 per cent (Illinois Bureau of Public Roads and Streets, 2009) to prevent slow-moving vehicles from sliding down the superelevation in cases of heavy rain, snow or icing conditions. This corresponds to a radius of 1000 metres for a design speed of 120 km/h and this minimum radius is recommended for all design speeds in areas where snow and icing are likely to occur, particularly in the case of intersections where the major road has a four-lane divided cross-section (Savolainen and Tarko, 2004).

Intersections should be located as far as possible on flat grades and a gradient of less than 3 per cent would be acceptable. Steeper gradients could create problems of sight distance as the intersection sight distance would have to be increased to allow for the longer stopping sight distance required on the major road. Steep gradients on the minor road would require a clear sight of the intersection ahead and signing to ensure that drivers start their deceleration timeously to come to a stop before entering the intersection area.

High fills, defined as fills where the fill batter is flatter than 1*V*:4*H*, should be avoided as sites for intersections. Crashes at intersections tend to involve more than one vehicle and one of the typical consequences of crashes is that one or both the vehicles could end up running off the road. The combination of the impact and rolling or sliding down the fill slope could possibly turn a serious crash into a fatal one. If this is an unavoidable location for an intersection, it should have safety barriers on all legs of the intersection. It would then be necessary to ensure that the barriers do not obstruct the sightlines of the intersection. Large quantities of fill may be required to ensure adequate sight lines past the barriers and signs.

Cuts deeper than about 1 metre or, alternatively, the driver eye height, should also be avoided as sites for intersections. If this is not possible, it will be necessary to 'daylight' the intersection, that is, to increase the scope of the excavation to ensure that the sight lines are not compromised. This may require large quantities of excavation and also, possibly, a change from what otherwise would be a Yield controlled intersection to Stop control.

10.9 CHANNELISATION

10.9.1 The application of channelisation

As traffic volumes increase, the complexity of the driving task increases exponentially. At four-legged intersections, conflict points follow each other in quick succession so that the time available for drivers to perceive and react to the changing circumstances becomes very limited. The pressure can, however, be at least partially relieved by identifying precisely where the conflict points are located and moving them as far apart as circumstances allow. Defining the travel paths for the various movements in the intersection area is achieved by providing islands that deliberately restrict the options available to drivers to select their own paths. This design process is referred to as channelisation.

In this section the following are discussed:

- The edges of turning roadways
- Channelising islands
- Auxiliary through and turning lanes
- Turning roadway widths
- Tapers

10.9.2 The edges of turning roadways

The edges of the turning roadways at an intersection should provide guidance to the driver regarding the appropriate path to follow. Three possible shapes of roadway edge can be considered. These are a simple radius, a simple radius with tapers on either end or a compound curve. The minimum radii for the various design vehicles listed in Chapter 5 are based on a turning speed of about 15 km/h and allow for an offset between the vehicle path and the kerb line. Offsets should be of the order of 600 mm. Radii of curvature are shown in Table 10.6 for a range of angles of skew from 60° to 120°.

Table 10.6 Edge of roadway dimensions

		Radius (m)		
Design vehicle	*Angle of skew (degrees)*	*Single curve*	*Single curve with taper*	*Compound curve*
Passenger car (P)	60	12	–	–
	75	11	8	30-8-30
	90	9	6	36-12-36
	105	–	6	30-6-30
	120	–	6	30-6-30
Single-unit truck (SU)	60	18	–	–
	75	17	14	36-14-36
	90	15	12	36-12-36
	105	–	10	30-11-30
	120		9	30-9-30
Semitrailer (WB19)	60	50	43	–
	75	–	43	134-23-134
	90	–	39	120-21-120
	105	–	35	160-15-160
	120	–	30	160-21-160
Semitrailer (WB22)	60	75	43	122-30-120
	75	–	39	128-23-128
	90	–	36	134-20-134
	105	–	33	152-15-152
	120	–	30	168-14-168

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.

The simple radius requires the least space and, with low volumes of turning traffic, is adequate. As volumes both of through and turning traffic increase, it is preferable that turning vehicles be removed from the through traffic flow. The provision of tapers allows for this but without going to the extent of providing deceleration and acceleration lanes. The best guidance is provided by a compound, usually three-centred, curve as it can fairly closely approximate to the actual vehicle path as the curve is negotiated.

The radii that should be offered to the various design vehicles are shown in Table 10.6. Acceptable taper rates and three centre curve radii are also shown (AASHTO, 2011a). The basis of selection of the design vehicle is discussed in Chapter 5. It should be noted that, as shown in Table 5.1, American vehicles are generally slightly larger than those of other countries. They have accordingly been selected as the basis of design of intersections, thus providing a larger level of safety for these other countries.

Turns at intersections take place at slow speeds so that, where a taper is provided, it need not be any flatter than about 1:10. The principal benefit of the taper is twofold. It effectively removes the turning vehicle from the through traffic stream and also reduces the angle through which the vehicle has to turn while executing the turn by nearly 12° in total. Usually, a triangular island is provided at intersections with tapers to offer guidance to drivers and also to reduce the dead area, that is, areas not normally traversed by vehicles in the intersection area.

10.9.3 Channelising islands

Channelising islands are provided to offer guidance to drivers regarding the path they should follow through the intersection area. As part of the guidance offered to drivers, it should be very difficult if not actually impossible to execute prohibited movements. Guidance is achieved by reducing the area of the intersection available to vehicles. This limits the propensity of vehicles to wander, making their paths difficult to anticipate by the drivers of other vehicles. The paths of conflicting movements are made to merge at flat angles or cross at angles as close to 90° as possible to ensure that opposing vehicles can easily be seen either in the rear-view mirror or by small movements of the head. Channelisation seeks to ensure that only one conflict occurs at any one time and furthermore that successive conflicts are removed as far as possible from each other. The priority that major movements should enjoy should be made clear to drivers by making these movements as directional as possible.

In addition, areas clear of those allocated to moving vehicles should be provided

- As a refuge for pedestrians
- For storage for vehicles waiting to turn
- For traffic control devices to ensure that they can easily be seen

As the majority of the crashes occurring on the road take place at intersections, channelising can reduce their consequences by forcing a reduction of the speed differentials at which they occur.

The temptation to provide numerous small islands should be resisted, as an archipelago tends to confuse rather than guide drivers through the intersection. A few large islands provide more positive guidance.

Some examples of the application of islands are shown in Figure 10.5. They can be seen to fall into three broad categories: median islands, triangular islands and splitter islands. Splitter islands, like median islands, are those located in the vicinity of the centreline of a road and generally serve to separate the two streams of traffic. Their main function is to provide a warning to drivers of the upcoming intersection. For this reason, the profile of the approach side of the island should be such that, from a distance preceding the intersection, it appears to block off the entire lane. When the vehicle is closer to the intersection, the redirection of the lane becomes clear.

The central figure in Figure 10.5 illustrates that the splitter island can also serve as an effective prohibition of the left turn at an intersection, for example, at a right-in/right-out access. The illustrated median islands include the deceleration components of left-turn lanes out of the major road. Left-turning vehicles often have their line of sight obscured by vehicles stored in the opposing left-turn lane. This can be rectified by a widening of the median island to include the provision of an offset deceleration lane as shown in Figure 10.6.

10.9.4 Auxiliary through and turning lanes

Auxiliary lanes are aimed at removing or, at least, minimising some localised impediment to smooth traffic operation. These impediments may refer to flow or to speed and are often the result of the interference caused by an opposing stream of traffic. A right turn may be impeded by through traffic moving from the left. Left-turning traffic has to yield right of way to traffic crossing its path from both directions as well as from the opposite direction. In consequence, one left-turning vehicle can cause a considerable backup of the vehicles behind it.

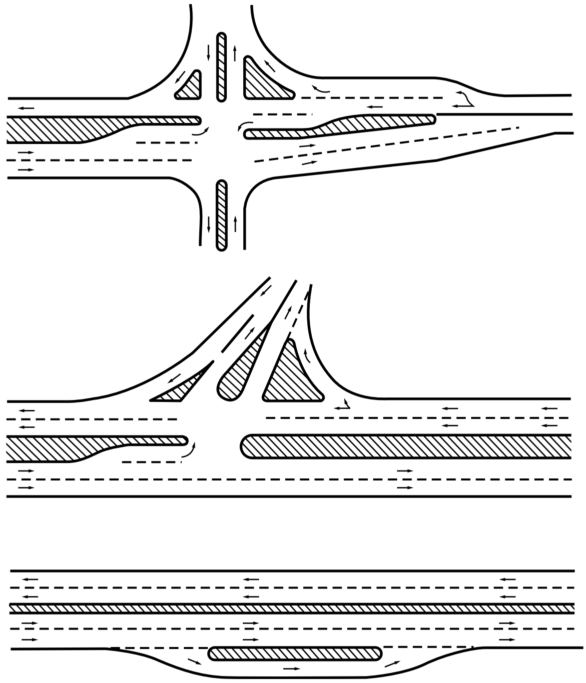


Figure 10.5 Typical layouts of channelised intersections. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.)

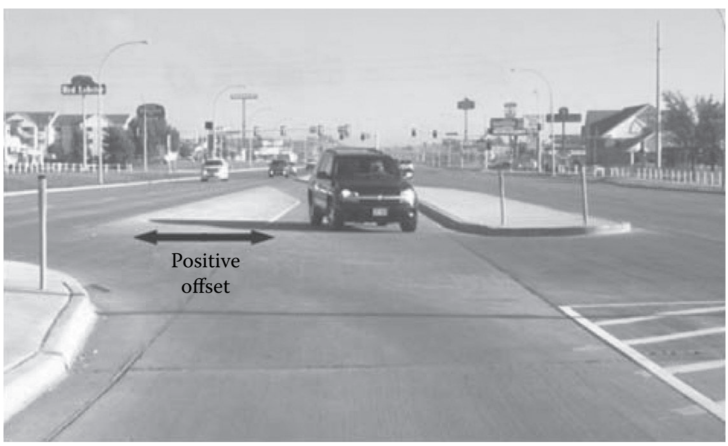


Figure 10.6 Offset left-turn lanes.

Auxiliary through lanes match the flow through the intersection to that up- and downstream of the intersection and are thus normally of application to signalised intersections as discussed further below. At priority controlled intersections they may, however, be applied to the minor road to increase the number of vehicles that can cross the major road during gaps in the flow along the major road.

Auxiliary lanes may comprise four components being a deceleration lane (including a taper) followed by a storage area for vehicles waiting to turn or to cross opposing streams of traffic, the turning roadway or lane and the acceleration lane (also including a taper). Each of these components needs to be considered separately because it does not automatically follow that all of them would be required at any specific location.

The length of the deceleration taper is a function of the speed of the traffic upstream of it and the preferred rate of deceleration from that speed to, in the worst case, a stop or, in the case of continuous movement through the turning lane, the radius and hence the speed of turning. This is based on the assumption that a vehicle will leave the through lane at the design speed of the road and commence its deceleration as soon as it starts to enter the taper, that is, as the outer front wheel crosses the edge of the through lane.

The length of the storage lane is a function of the number of vehicles that may have to come to a dead stop prior to moving through the turning roadway. This number would be estimated by application of probability theory as discussed in Chapters 17 and 18. The length of the storage lane is the product of the number of vehicles and their overall length plus an allowance for the gap between each pair of vehicles typically of the order of about 1.0 to 1.5 metres.

The width of the turning roadway or lane is determined by the selection of the design vehicle appropriate to the intersection being designed and its dimensions as discussed in the following section.

The acceleration lane in combination with its taper must be long enough to allow a vehicle that has completed its turn to accelerate to the speed of the through traffic and to merge with it. The taper rate is a function of the speed to which the turned vehicle must accelerate for successful merging, that is, the higher the speed the flatter the taper.

Seeing that urban design speeds are typically of the order of 60 km/h, a taper rate of between 1:10 and 1:15 for both the deceleration and the acceleration tapers would be adequate. As a first approximation, the designer could thus assume a taper rate of 1:12.5 for both and refine them as the design progresses. Refinement should typically be to flatten the taper rate of the acceleration taper towards 1:15 and sharpen the taper rate of the deceleration taper towards 1:10.

In rural areas, design speeds are more likely to be between 100 and 120 km/h, suggesting a taper rate of about 1:50 for the acceleration taper, which is, in fact similar to that adopted for interchange on-ramp tapers.

At interchanges, an off-ramp taper of 1:15 is typically considered acceptable. However, in that case deceleration continues long after the end of the taper. It is recommended that this taper rate be adopted in the case of intersections also and provision made for the continuation of deceleration in an auxiliary lane immediately adjacent to the through lanes.

10.9.5 Turning roadway widths

As soon as channelisation is applied to an intersection, reference is no longer to turning lanes but to turning roadways, also referred to as slipways. The facility of vehicles being allowed to encroach on adjacent lanes is no longer available and space must be provided between the islands and roadway edges to accommodate them. It is necessary to provide for off-tracking, which is the offset between the outside front wheel and the inside rear wheel of the vehicle. The lateral distance between the front overhang and the inside rear wheel is referred to as the wall-to-wall width of the turn or also as the width of the swept area covered by the vehicle.

Turning roadways are normally provided only for right turns. Having an auxiliary lane merging into the fast lane, as would be the case for a left turn, is problematic, as an opposing vehicle may be in the blind spot of the driver of the merging vehicle. Furthermore, merging into the fast lane is in conflict with driver expectations. Reference to a turning roadway implies the presence of a channelising island. For left turns, a turning lane is usually sufficient.

The width of turning roadways is dependent on the selection of design vehicle, the radius of curvature and the type of operation with the radius of curvature being dictated by the volume of traffic. As stated previously, American passenger cars are slightly larger than those of other countries. Using them as the design vehicle thus provides a modest margin for these other countries. The type of operation is categorised as (AASHTO, 2011a)

- Traffic condition A: Predominantly P vehicles, but some consideration for SU trucks
- Traffic condition B: Sufficient SU vehicles to govern design, but some consideration for semitrailer vehicles
- Traffic condition C: Sufficient buses and combination trucks to govern design

Traffic condition A would be sufficient for intersections on access roads in residential areas, whereas, on more important routes such as urban collectors and rural inter-town roads, design would normally be predicated on traffic condition B. In CBDs, it would be necessary to make provision for busses, and, in industrial areas, there would be sufficient semitrailers to dictate design.

Table 10.7 Derived turning roadway pavement widths

Radius on pavement inner edge of pavement (m)	*P*	*SU*	*BUS*	*A-BUS*	*WB-19*	*WB-20*	*Interlink*
Case I operation – no provision for passing a stalled vehicle							
15	4.0	5.5	6.6	6.7	8.8	13.5	–
25	3.9	5.0	5.7	5.7	6.8	8.5	9.6
30	3.8	4.9	5.4	5.5	6.3	7.8	8.6
50	3.7	4.6	5.0	5.0	5.5	6.6	6.8
75	3.7	4.5	4.8	4.8	5.1	5.7	6.0
100	3.7	4.4	4.7	4.7	5.0	5.3	5.6
Case II operation – provision for passing a stalled vehicle by another of the same type							
15	6.0	9.2	11.9	12.4	25.2	15.4	–
25	5.6	7.9	9.6	9.9	15.0	11.2	16.9
30	5.5	7.6	9.0	9.3	13.4	10.4	14.9
50	5.3	7.0	8.0	8.1	10.4	8.7	11.2
75	5.2	6.7	7.4	7.5	9.1	7.9	9.6
100	5.2	6.5	7.2	7.2	8.4	7.5	8.8
Case III – two-lane operation by same type vehicles in both lanes							
15	7.8	11.0	13.7	14.2	27.0	17.2	–
25	7.4	9.7	11.4	11.7	16.8	13.0	18.7
30	7.3	9.4	10.8	11.1	15.2	12.2	16.7
50	7.1	8.8	9.8	9.9	12.2	10.5	13.0
75	7.0	8.5	9.2	9.3	10.9	9.7	11.4
100	7.0	8.3	9.0	9.0	10.2	9.3	10.6

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.

Operational considerations are defined (AASHTO, 2011a) in one of three cases with these being described as

- Case I: One-lane, one-way operation with no provision for passing a stalled vehicle
- Case II: One-lane, one-way operation with provision to pass a stalled vehicle
- Case III: Two-lane operation

Case I is normally acceptable in the case of minor turning volumes on relatively short turning roadways. The chances of vehicles being stalled are relatively remote but the edges of the roadway could either be edged or have mountable kerbs enabling stalled vehicles to move onto the shoulder before they run out of speed.

Case II allows for higher turning volumes that are still within the capacity of one-lane operation and also make provision, albeit at lower speeds, for the situation where the stalled vehicle was unable to move onto the shoulder. Even so, the edge of the turning roadway should still either be edged or have mountable kerbs as for Case I. For the longer design vehicles, Case II widths may be impractically large and it may be necessary either to limit the design to Case I operation or, alternatively, omit the channelising island.

Case III would be used where the turning volume exceeds the capacity of a single lane. At an intersection, operation would obviously have to be one-way.

The widths of turning roadways for various design vehicles, traffic conditions and operational considerations are taken from AASHTO (2011a), with minor modifications, in Table 10.7.

10.10 ROUNDABOUT DESIGN

10.10.1 Introduction

A distinction is drawn between traffic circles and roundabouts, with traffic circles going as far back in time as the Bath Circus built in 1768. More recently, Columbus Circle in New York City was built in 1905 and the Place de l'Étoile around the Arc de Triomphe in 1907. Less famous but numerous traffic circles were built elsewhere including in the United Kingdom and the north-eastern states of the United States. A significant difference between the traffic circle and the modern roundabout is that, in the former, right of way was accorded to entering vehicles, that is, they took precedence over vehicles already in the circle. In consequence, traffic circles were very inclined to suffer gridlock. Many were subsequently replaced by conventional or signalised intersections.

In Britain, the Transport Research Laboratory developed the 'off-side priority rule' during the 1960s. This meant, in a country where vehicles drive on the left, that priority was accorded to vehicles approaching from the right. The roundabout thus emerged as a viable entity. In the United States the name traffic circle still enjoys usage but this is to differentiate between circular intersections controlled by stop signs or traffic signals and those controlled by priority rules. In the case of countries where driving is on the right, priority is given to the vehicles approaching from the left.

It has been suggested elsewhere that intersection design is one of the more complicated issues that can occupy the designer's mind. Of these, the roundabout is the most complex by a considerable margin. The conventional three- or four-legged intersection normally commences with selection of the form of control based on the flow that has to be accommodated and a decision regarding the need for channelisation. This is followed by ensuring an acceptable angle of skew. Once these have been finalised, the design vehicle determines the dimensions of the various elements of the intersection.

The roundabout, on the other hand, is beset by numerous interlocking variables. The intersection sight distance should, unlike the case of the conventional intersection, deliberately be kept low both to focus the driver's attention on the immediate environment of the roundabout and to force a low operating speed on the vehicles within it. The angles of skew of the various legs of the roundabout are not fixed and should be selected to ensure that the selected entry speed not exceeded. In the conventional intersection, the centrelines of the crossing roads are continuous whereas the centrelines of the crossing roads in a roundabout are invariably discontinuous and offset either to the right or to the left.

In spite of all the complexities that beset the designer, the fact of the matter is that the roundabout presents the driver with significant advantages. If one includes diverging manoeuvres as conflict points, four-legged turbo roundabouts, discussed in Section 10.10.3, have 10 conflict points and conventional four-legged roundabouts a total of 16 conflict points compared to the 32 of the conventional four-legged intersection. It is to be expected therefore that roundabouts would be significantly safer than conventional intersections. Because of low absolute speeds and speeds relative to other vehicles, crashes that do occur on roundabouts are usually less serious than those on conventional intersections.

In this section, the design elements that comprise roundabouts, their various forms and the design principles that govern their design are discussed.

10.10.2 Geometric elements

The principal characteristics of the roundabout are illustrated in Figure 10.7.

The elements that the roundabout requires to give effect to these characteristics are

- The central island, which defines the radius of the inner edge of the circulating roadway and provides space for landscaping which, apart from any aesthetic value it may possess, also provides a clear signal to drivers of the approach to the roundabout
- The apron, which, if needed to accommodate the turning roadway width required by long vehicles, is the mountable portion of the island immediately adjacent to the circulatory roadway
- The circulatory roadway, which comprises the lane(s) used by vehicles travelling in a counterclockwise direction for driving on the right-hand side of the road or clockwise for driving on the left
- The splitter islands, which are kerbed or painted areas on the approaches to the circulatory roadway with the function of deflecting and hence slowing entering vehicles, separating entering and exiting flows and making it possible for pedestrians to cross the road in two stages
- Landscaping strips, which separate pedestrian and cycle traffic from the circulatory roadway and provide an indication of the location of pedestrian crosswalks
- Pedestrian crosswalks, which are normally set back from the circulatory roadway by the length of one of two waiting passenger cars and may include cutting through the splitter island for the benefit of wheelchair users and cyclists
- Yield lines at the entrance to the circulatory roadway, which are the points at which vehicles are required to yield to vehicles on the circulatory roadway although it is understood that, in the United States, where they are still referred to as Yield lines they do not in fact have this function and merely serve to demarcate the edge of the turning roadway

The elements of a two-lane roundabout are illustrated in Figure 10.8.

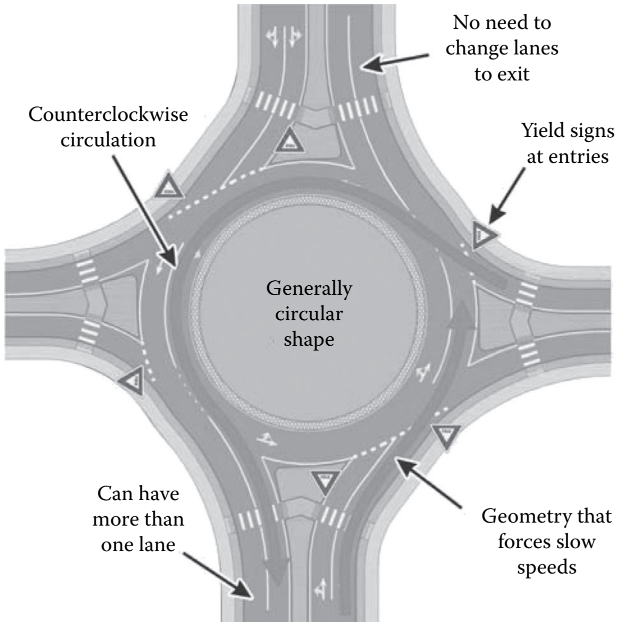


Figure 10.7 Characteristics of the roundabout. (From Rodegerdts L et al., *Roundabouts: An informational guide*. National Cooperative Highway Research Program [NCHRP] Report 672, Transportation Research Board, Washington, DC, 2010.)

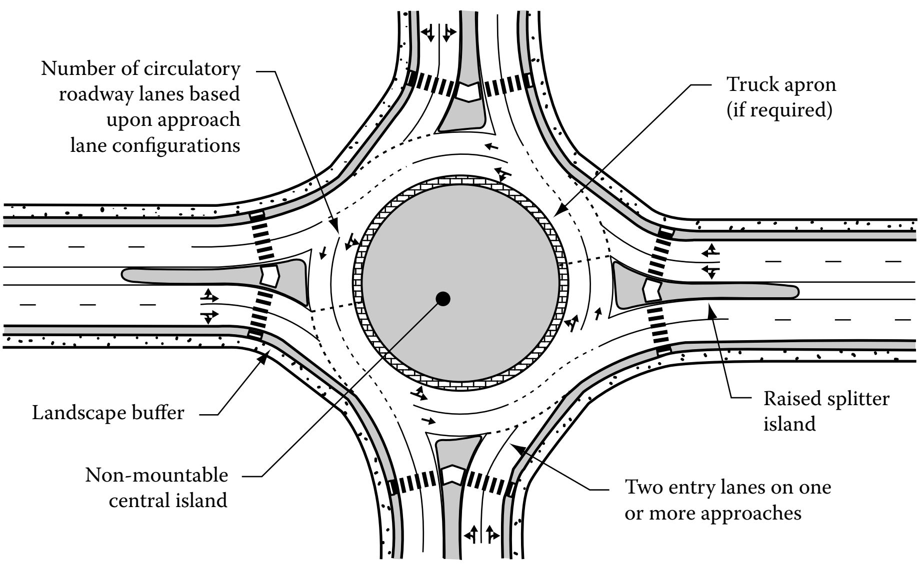


Figure 10.8 Elements of the two-lane roundabout. (From Rodegerdts L et al., *Roundabouts: An informational guide*. National Cooperative Highway Research Program [NCHRP] Report 672, Transportation Research Board, Washington, DC, 2010.)

10.10.3 Types of roundabouts

Roundabouts are classified as

- Mini roundabouts
- Single-lane roundabouts
- Multilane roundabouts
- Turbo roundabouts
- Tear drop roundabouts
- 'Magic' roundabouts

Mini roundabouts are a subset of single lane roundabouts and are normally used in residential areas because of limitations of space. In the extreme, the mini roundabout can be contained within the width of the road reserve, with the central island being just a painted circle with a diameter often as small as 1 metre on the road surface. In this case, the inscribed diameter, illustrated in Figure 10.9, may be as little as 6.5 metres, suggesting a circulating roadway with a width of 3.0 metres around a central island with a diameter of 1 metre. This design would be suitable only for passenger cars. It is not the intention that vehicles should attempt to drive around such a small centre island but simply to indicate that control is by priority to the near side for driving on the right or offside for driving on the left.

The single lane roundabout with an inscribed circle diameter of 15 metres, and a central island diameter of 10.5 metres would accommodate the SU design vehicle. The WB19 semitrailer would require an inscribed circle diameter of 30 metres. The off-tracking of the inner rear wheels would require the provision of an apron around the central island because, with this diameter of inscribed circle, the radius of their path would only be 2.5 metres.

Two-lane roundabouts would typically have an inscribed circle diameter range of 55 to 67 metres and three-lane roundabouts a diameter of 60 to 75 metres to accommodate the WB20 semitrailer.

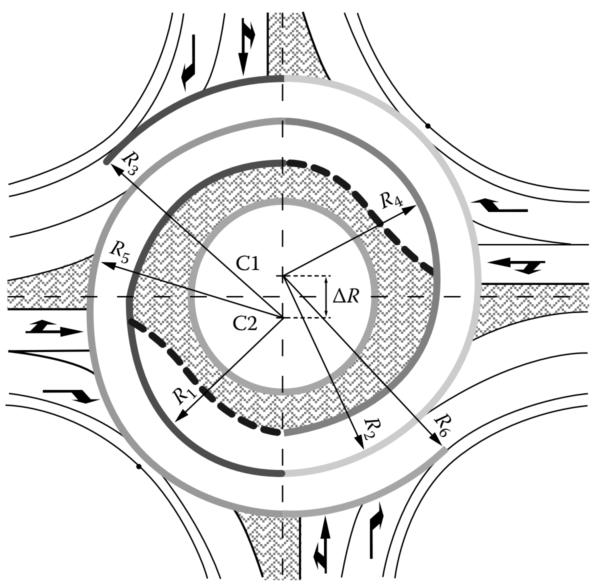


Figure 10.9 The geometric design of the turbo roundabout. (From Giuffrè O et al., *Turbo roundabout general design criteria and functional principles: Case studies from the real world.* 88th Annual Meeting, Transportation Research Board, Washington, DC, 2009.)

Table 10.8 Comparison of number of conflict points

	Number of conflict points		
Number of legs	*Unsignalised intersection*	*Two-lane conventional roundabout*	*Turbo roundabout*
3	9	16	7
4	32	22	10

The turbo roundabout was developed in the Netherlands (Giuffrè et al., 2009). The geometry of the turbo roundabout is illustrated in Figure 10.9. It has fewer conflict points than a conventional roundabout as shown in Table 10.8. The central island of the turbo roundabout is not normally circular.

For ease of setting out, the lane layout may be modified from true spirals to compound curves comprising different radii. To apply this methodology, the distance, ΔR, equals the selected lane width and the centres C1 and C2 of the arcs are located symmetrically ($\Delta R/2$) along the centreline of the major road around the intersection of the two roadways. The lane width is based on the smallest radii, R_1 and R_4, which are determined by the selection of the design vehicle and design speed of the roundabout. The remaining radii are calculated as

$$R_2 = R_1 + \Delta R \quad \text{and} \quad R_3 = R_2 + \Delta R$$

$$R_5 = R_4 + \Delta R \quad \text{and} \quad R_6 = R_5 + \Delta R$$

The teardrop island, illustrated in Figure 10.10, is usually associated with the crossing road ramp terminals at interchanges, specifically diamond interchanges. It has two advantages. The first is that vehicles on the crossing road of the interchange are not confronted by yield, stop or signalised traffic control measures and the second is that it is relatively easy to develop a design inhibiting wrong-way driving.

The so-called Magic Roundabout is a British invention that is extremely frightening to drivers encountering it for the first time. It is intended to address the problem of having more than four legs in the intersection and reportedly can accommodate high volumes of traffic. As shown in Figure 10.11, traffic flows in both directions around a large central island and a mini roundabout is placed at the intersection of each leg with the circulating roadway.

By all accounts, these roundabouts, of which five have now been built in the United Kingdom, have a very good safety records because of the low speeds at which they operate. Furthermore, they are claimed to have high capacities.

10.10.4 Design principles

The fundamental design principles relating to roundabouts take cognisance of human factor needs by providing

- Appropriate sight distance and visibility
- Adequate accommodation for the design vehicles
- Channelisation that is intuitive to drivers resulting in their naturally using the intended lanes
- Slow entry speeds and consistency of speed through the intersection area
- Sufficient lanes to achieve adequate capacity
- Clear road markings defining assignment of the various lanes and lane continuity
- For the needs of pedestrians, cyclists and other vulnerable road users

Figure 10.10 The teardrop central island. (From Rodegerdts L et al., *Roundabouts: An informational guide*. National Cooperative Highway Research Program [NCHRP] Report 672, Transportation Research Board, Washington, DC, 2010.)

Part of the complexity of roundabout design stems from the fact that adjusting one element of the roundabout may adversely affect another with a strong following ripple effect. The selection of the semitrailer as the design vehicle would result in a very wide circulating roadway and this may lead to unacceptably high entry speeds by smaller vehicles. A wide turning roadway may be perceived by the drivers of passenger cars as being a two-lane roadway. The difference in speeds between entering and circulating vehicles has a negative impact on safety and could be reduced by offsetting the entering roadway more to the left, for example. This would have the effect of increasing the angle of skew to the right and this would have an impact on the sight distance afforded the entering driver. Alternatively, the size of the central island could be increased, hence narrowing the circulating roadway, which negates the original intention of favouring the selected design vehicle. This could be addressed by the provision of an apron around the central island.

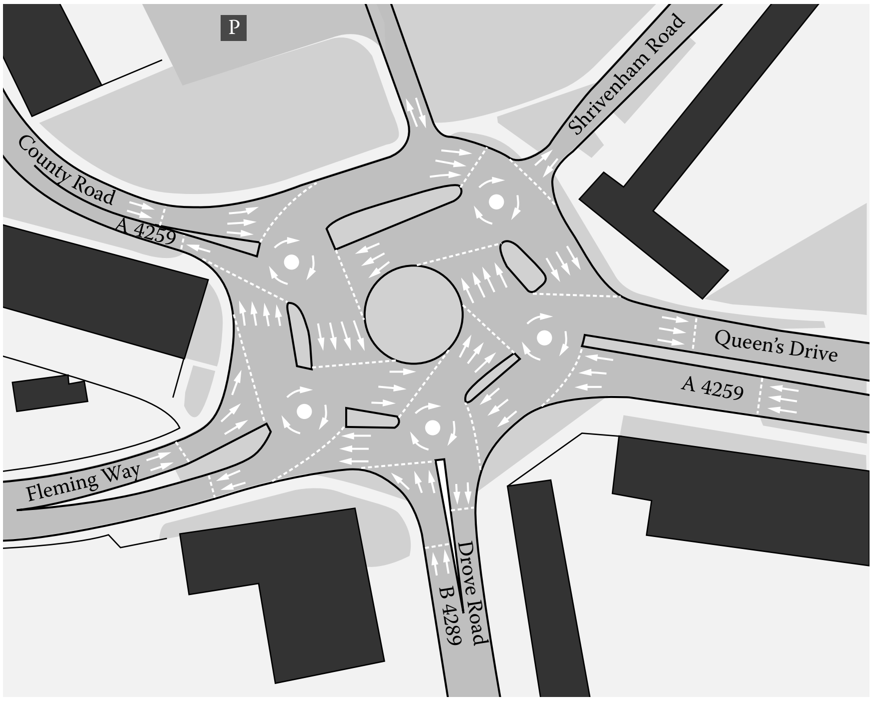


Figure 10.11 The 'Magic Roundabout', Swindon, UK.

To ease the navigational problems that drivers may experience in negotiating a roundabout, a form of lane balance should be provided. For example, if the major route involves a turn to the left, a two-lane entrance followed by single-lane exits at the intervening roads would guide drivers to the two-lane exit to the major route. A similar situation is illustrated in Figure 10.12.

The preceding text illustrates that design of a roundabout, far more so than for any other form of intersection, is a highly iterative process. Every design decision must be carefully scrutinised to ensure that it does not force a negative consequence on some other aspect of the design.

10.10.5 The application of roundabouts

As shown in Figure 10.1, roundabouts can handle a wide range of traffic volumes. Flows of 3000 vehicles per hour on the major road in combination with 700 vehicles per hour on the minor road are about the upper limit of what can be accommodated by a roundabout. Once these flows are exceeded, it would be necessary to consider signalisation of the at-grade intersection or perhaps even the inclusion of a grade separation.

A combination of a major road flow of 1600 vehicles per hour with a minor road flow of 200 vehicles per hour produces an interesting anomaly. According to the simulations on which Figure 10.1 is based, all three forms of control are equally effective in minimising delay at these flows. Selection of the appropriate form of control would thus be guided by minimisation of the cost of construction and maintenance whereby it is obvious that the two-way stop would be preferred to the roundabout which, in turn, would be preferred to signalisation.

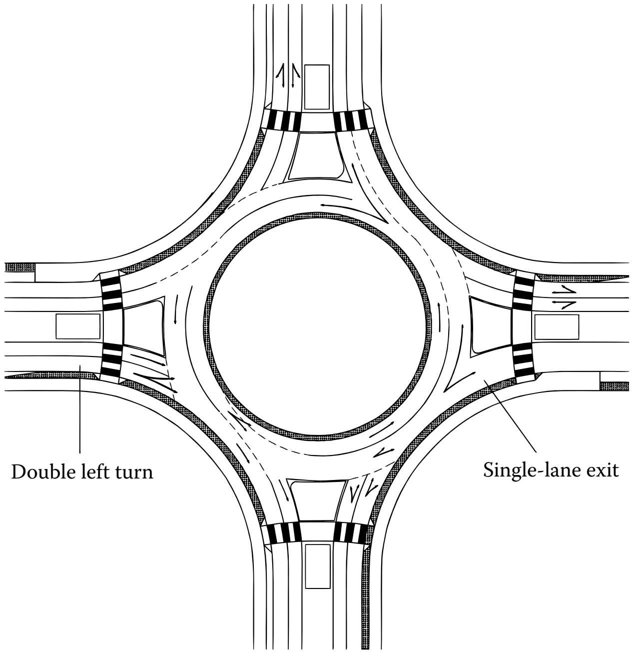


Figure 10.12 Lane balance at roundabouts. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.)

There are circumstances where roundabouts should not be used. As a general rule, drivers do not expect to see isolated roundabouts in rural areas. This is borne out by the tendency of local authorities in rural towns to install roundabouts on the outskirts of their towns 'to break the speed of the vehicles entering town' and invariably end up breaking the vehicles because the drivers did not react in time to the unexpected obstruction. In Section 10.7, sites in rural areas that would be inappropriate for conventional intersections are discussed. These apply equally to roundabouts.

In urban areas, where traffic flows are unbalanced with high volumes on one or more approaches, some vehicles will experience long delays. Roundabouts are similar to all-leg stops in that they cause delays to all vehicles whereas priority control results in delays only to the minor road traffic.

As discussed in Section 10.9.9, roundabouts present problems to pedestrians, particularly if visually impaired. Crossings are between vehicles and multi-lane roundabouts can thus be particularly hazardous. Where large numbers of pedestrians are anticipated, particularly in the presence of high traffic volumes, it may be difficult for the pedestrians to cross any leg of the intersection.

10.10.6 Sight distance at roundabouts

At a roundabout, as in the case of other intersections, sight triangles are required to ensure that adequate intersection sight distance (ISD) has been provided. Sight triangles may either be approaching or departing as shown in Figure 10.2. However, as illustrated in

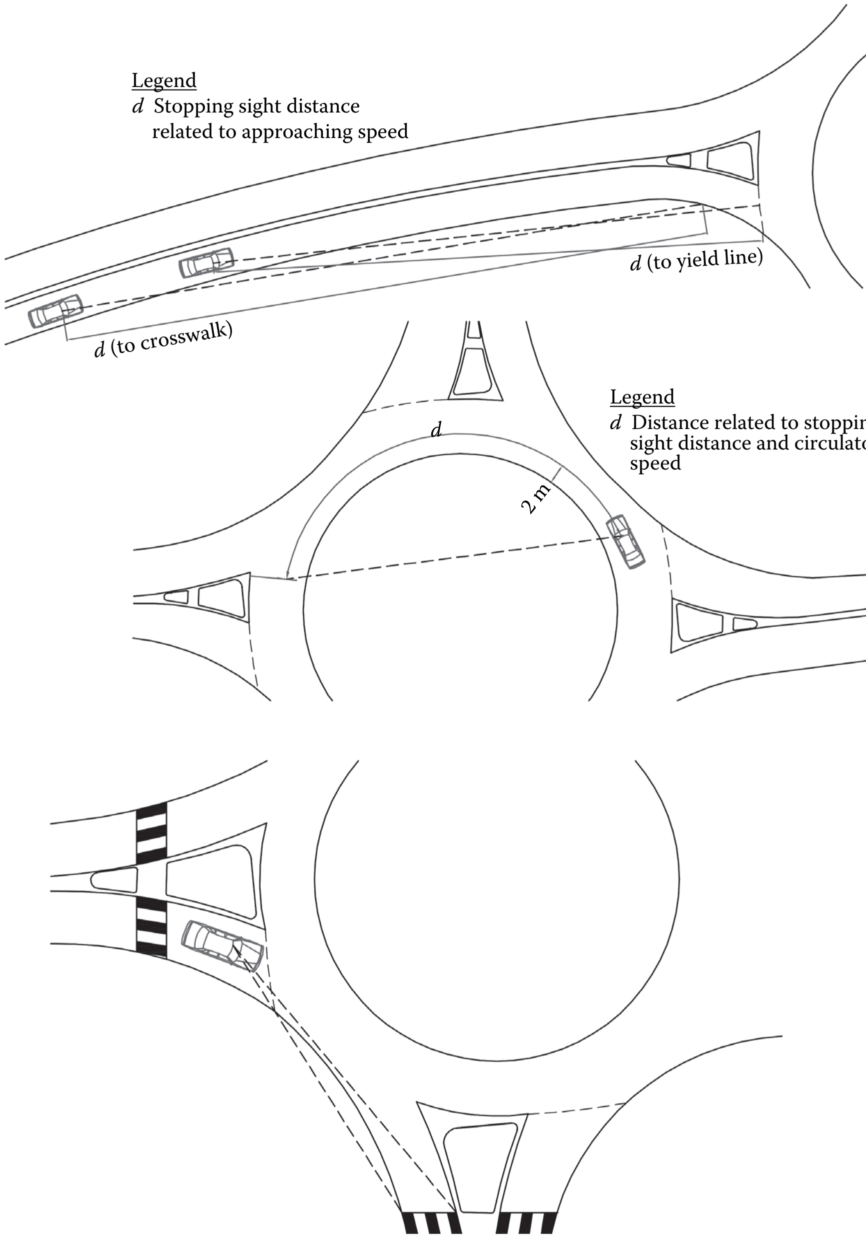


Figure 10.13 Sight distance at roundabouts.

Figure 10.13, resemblance to conventional sight triangles is slight. ISD is required at every point through the functional area of the roundabout but specifically

- In respect of vehicles entering the roundabout from the left (in the case of driving on the right)
- To vehicles already on the circulating roadway
- To pedestrians crossing the following leg of the roundabout

Sight distances as provided in Tables 10.1, 10.2 and 10.3 are appropriate to roundabouts. The major difference between roundabouts and other intersections is that, in the case of the other intersection forms, the ISD required is a minimum whereas, according to British experience, no more than the minimum should be provided for roundabouts. This has the effect of focussing the attention of drivers on the task at hand. Furthermore, higher values of ISD may lead to higher approach speeds, which could impact negatively on the safety of all road users. Higher approach speeds could also result in the need to brake sharply once on the circulatory roadway. Landscaping of the central island would be helpful in limiting the sight distance to vehicles on the opposing leg of the roundabout, simultaneously increasing the target value of the roundabout by improving the visibility of the central island.

It should also be noted that, with conventional intersections, sight distances are typically provided for design speed of 60 km/h or more. In the case of roundabouts, sight distances are for speeds of seldom more than 40 kilometres.

10.10.7 Alignment of approaches and entries

In the case of conventional intersections, the centrelines of the crossing roads are normally continuous through the intersection area. However, in the case of roundabouts, continuity of the centrelines is less important than the need to reduce vehicle speed at the approaches and through the roundabouts. The alignment of the approaches plays an important role in the extent of the control of speed achieved. To highlight the iterative nature of roundabout design, effective control of speed by changes in the alignment of the approaches may also result in the angle of skew to the upstream leg being outside the desirable limit of 75° and a compromise between these conflicting requirements would have to be achieved.

The alignment of the approaches and entries can be modified by moving them either to the left or to the right relative to the central island as illustrated in Figure 10.14.

Alignment to the left allows for better speed control by increasing the angle of deflection. The exit tends, however, towards being more tangential with a corresponding reduction of speed control on exit. Speed control on exit is important to ensure the safety of people on the downstream pedestrian crossing. With no offset of the alignment, the approaches are less subject to the introduction of curves and the layout has better speed control on the exits than does the offset to the left. Alignment to the right may be useful for meeting speed control objectives with large diameters of inscribed circle, but increases the amount of exit curvature to be negotiated. With offsetting to the right, there is always the risk that vehicles may enter roundabouts at too high a speed, resulting in loss-of-control and possible crashes. In general, an offset to the right is not all that successful but can be accepted if the speed control requirement can be met.

10.10.8 Inscribed circle diameters

The inscribed circle comprises the central island and the circulatory roadway. Its diameter is thus the sum of the central island diameter and the twice the width of the circulatory road. This diameter is determined in part by the radius of curvature appropriate to the design vehicle at the selected design speed and the swept area of the design vehicle at that radius. It is also a function of the level of speed control required. It must be large enough to provide adequate deflection to ensure safe travel speeds by vehicles smaller than the design vehicle. However, the entry and exit widths of the approaches, the entry and exit radii and the entry angles also play a role in providing deflection and hence travel speeds. Careful selection of these inputs may make it possible to employ a smaller diameter of the inscribed circle and hence a smaller footprint overall for the roundabout.

Typical ranges of the inscribed circle diameters are shown in Table 10.9.

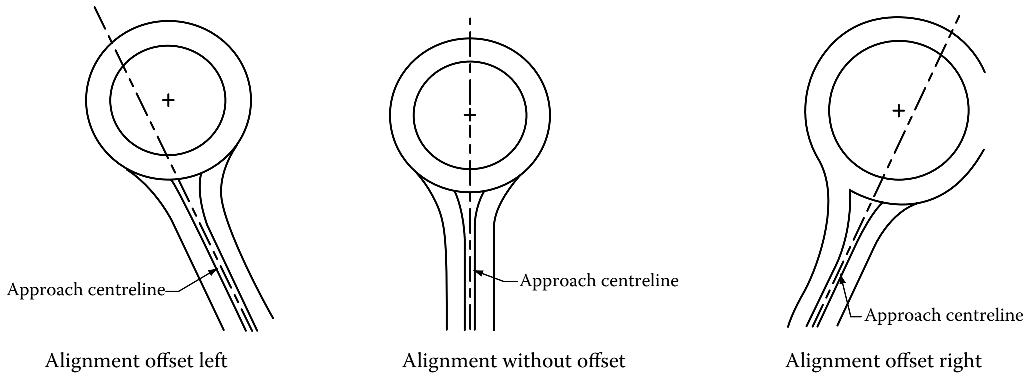


Figure 10.14 Alignment of approaches. (From Rodegerdts L et al., *Roundabouts: An informational guide*. National Cooperative Highway Research Program [NCHRP] Report 672, Transportation Research Board, Washington, DC, 2010.)

Table 10.9 Typical inscribed circle diameters

Roundabout configuration	*Typical design vehicle*	*Inscribed circle diameter range (m)*
Mini-roundabout	SU-9	14–27
Single-lane roundabout	B-12	27–46
	WB-15	32–46
	WB-20	40–45
Two-lane roundabout	WB-15	46–67
	WB-20	50–67
Three-lane roundabout	WB-15	61–76
	WB-20	67–91

Source: Rodegerdts L et al., *Roundabouts: An informational guide*. National Cooperative Highway Research Program (NCHRP) Report 672, Transportation Research Board, Washington, DC, 2010.

10.10.9 Nonmotorised traffic

Cyclists and pedestrians are not well-served by roundabouts. Cyclists typically ride close to the shoulder or kerb if they are sharing the roadway with other vehicles. On the circulating roadway, if wishing to make a left turn they therefore have to weave across the paths of

- Right-turning vehicles from the road on which they entered the roundabout
- Vehicles entering the roundabout from their left and travelling straight through the roundabout or turning to the left
- Left-turning vehicles entering the roundabout from straight ahead

While these manoeuvres are difficult enough on a single-lane circulatory roadway, they are significantly more difficult on multi-lane circulatory roadways. Even if there aren't cycle paths up- and downstream of the roundabout on all the approach legs, it is thus recommended that these be provided outside the landscaping strips of the roundabout as illustrated in Figure 10.15.

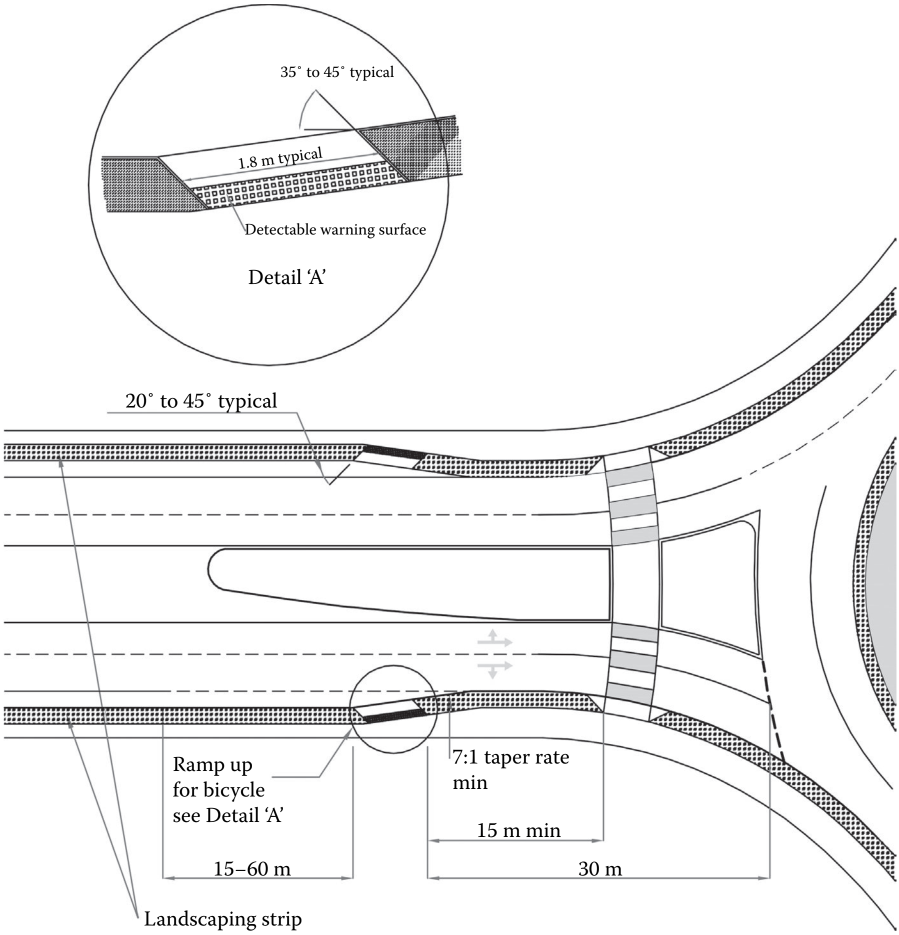


Figure 10.15 Cycle paths at roundabouts. (From Rodegerdts L et al., *Roundabouts: An informational guide*. National Cooperative Highway Research Program [NCHRP] Report 672, Transportation Research Board, Washington, DC, 2010.)

The cycle path is additional to the pedestrian path. If combined, the path should have a minimum width of 3.0 m and be demarcated with a painted centreline, with half allocated to cyclists and the other half allocated to pedestrians. It would be necessary to provide painted legends on the surface to indicate that the centreline is not meant to indicate opposing flow directions but rather to differentiate between the two modes of travel. If there are significant volumes of pedestrians and/or cyclists and movement in opposing directions it would be necessary to increase the overall width of the path by a further 1.0 m for each mode of travel in which the heavy flow occurs. Cyclists would cross the lane(s) of the circulatory roadway as pedestrians.

Wherever pedestrians are present other than infrequently, sidewalks in the form of pedestrian paths that are separated from the circulatory roadway by landscaping strips should be provided. Landscaping strips create a more comfortable environment for pedestrians than when the sidewalk is immediately adjacent to the circulatory roadway. These strips provide a buffer for the overhang of vehicles, which in the case of busses can be substantial. They also allow for landscaping and street furniture and, in colder climes, for storage of

Table 10.10 Accommodation of nonmotorised road users

Road user	*Characteristic*	*Dimension (m)*	*Affected roundabout feature*
Cyclist	Length	1.8	Splitter island width at crosswalk
	Operating width	1.2	Bike lane on approach roadway
Pedestrian	Width	0.5	Pedestrian lane, crosswalk
Wheelchair user	Minimum width	0.75	Pedestrian lane, crosswalk
	Operating width	0.9	Pedestrian lane, crosswalk
Person pushing stroller, pushchair or pram	Length	1.7	Splitter island width at crosswalk
In-line skaters	Typical operating width	1.8	Splitter island width at crosswalk

snow. Furthermore, the plantings on the landscaping strips can discourage pedestrians from crossing to or cutting across the central island. Finally, the landscaping strips serve to guide pedestrians to the designated crosswalks.

Visually impaired pedestrians have a serious problem at roundabouts. At conventional intersections, it is possible to pick up which traffic stream is moving from the changes in the direction from which the traffic noise is coming. At signalised intersections, different sounds can be used to provide useful information. For example, the sound of a dove cooing for two of the legs of a four-legged intersection and the sound of a cuckoo for the other two would provide an indication of when it is safe to cross the road. At roundabouts, the noise made by circulating vehicles is continuous and it is thus difficult to establish the presence of gaps in the traffic stream to be crossed. Finding the location of the crosswalk can also be difficult for a visually impaired pedestrian. The landscaping strip, provided it is kerbed or has some other distinctive edge treatment, can, however, provide useful guidance to the location of the crosswalk.

Pedestrians include people in wheelchairs. The camber of the half of the path allocated for their purpose should be not more than about 1 per cent because it would otherwise be difficult for wheelchair users to maintain travel in a straight line without rolling towards the edge of the path.

The dimensions of various elements of the roundabout required to accommodate nonmotorised road users are summarised in Table 10.10.

10.10.10 Landscaping the central island

Apart from the aesthetic benefits of landscaping the central island, it serves to increase the target value of the roundabout to the approaching driver as shown in Figure 10.16. The town planning principle of providing 'gateways' to areas that are fundamentally different from the previous environment is well-served by the application of landscaping of central islands of roundabouts. Furthermore, landscaping also serves the Human Factors Six Second Axiom (PIARC [World Road Association], 2008) discussed in Chapter 4 in giving drivers advance notice of the impending change in circumstances. Examples are the entrances to shopping malls or residential areas from the surrounding road network.

The landscaping of the central island should be properly planned and thereafter well maintained, as a derelict central island would detract from the appearance of the entire area, including even that beyond the roundabout.

It has been pointed out previously that the sight distance provided at roundabouts should be restricted to not more than the minimum required for the selected design

Figure 10.16 Landscaping of the central island. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.)

speed. High plantings on the central island can be a useful way of achieving this. In areas of the central island where line of sight is required, plantings should, however, be lower than 1 metre.

Large shade trees are not recommended as part of the landscaping of central islands. They serve as attractions to pedestrians to cross the circulatory roadway to enjoy the shade offered, whereas it is desirable to discourage the presence of pedestrians on the roadway. Even if there is a clear line of sight under them, going from direct sunlight to shade while traversing the circulating roadway may cause road markings not to be seen.

The design of the roundabout should clearly indicate where higher plantings may be used and where the height of plantings should be restricted.

10.10.11 Design check

In Chapter 4 reference was made to consistency of design with particular reference to speed variations between successive curves. This also applies in the case of roundabout design where the successive curves are so close together that they comprise a continuously curving alignment as illustrated in Figure 10.17.

It is not possible to vary the superelevation to match the radii of the successive curves for two reasons:

- There is insufficient space to effect the change from one rate of superelevation to another.
- The various radii along the fastest path will occur at different places for each of the approaches to the roundabout.

It follows that the crossfall around the circulatory roadway would remain constant at whatever value is applied throughout the design of the road, typically 2 to 2.5 per cent. At the entry and exit to the roundabout the superelevation will be positive, whereas on the circulatory roadway it will be negative.

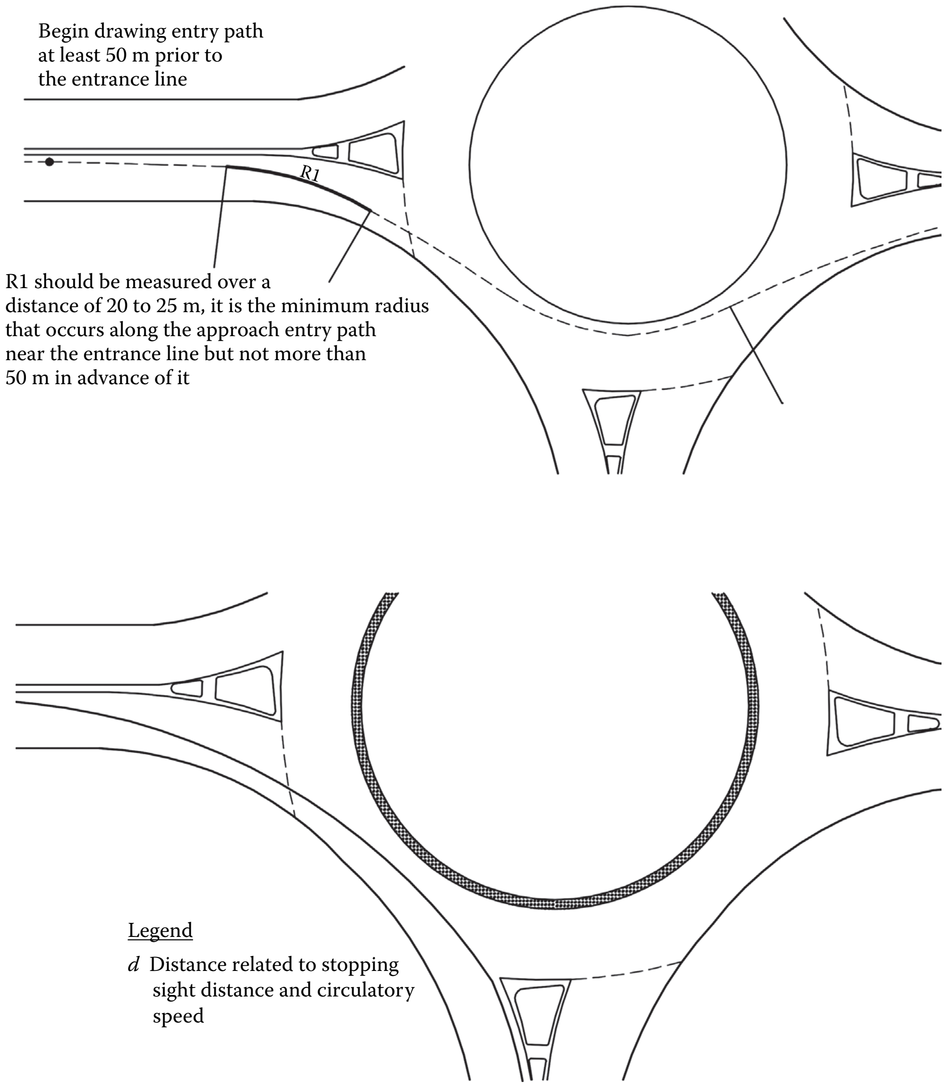


Figure 10.17 The fastest path. (From Rodegerdts L et al., *Roundabouts: An informational guide.* National Cooperative Highway Research Program [NCHRP] Report 672, Transportation Research Board, Washington, DC, 2010.)

The concept of the fastest path comes into play and is a useful tool for checking the consistency of design. The fastest path allowed by the geometry determines the negotiation speed for any of the three or four movements through the roundabout from entry to exit. It is the smoothest, flattest path possible for a single vehicle in the absence of other traffic and ignoring all lane markings. It is the 'best line' adopted by racing drivers who will start negotiating a curve from the marbles on the outside of the track, cut across to the marbles on the inside of the track in the middle of the curve and then wind out again to outside of the curve.

The fastest path is checked for all movements from all approaches. The left-turn movement contains the greatest extent of change of direction and is hence, in general, the slowest of the fastest paths. Depending on the angle of skew applied to a leg, the right turn could be the fastest of the fastest paths. The fastest path is constructed through the centreline of the vehicle traversing the path and has five control points through which it must pass. Assuming a vehicle width of 2.0 m, these are shown in Figure 10.18.

The fastest path does not represent expected vehicle speeds but is useful to establish the theoretically attainable entry speeds for design purposes. Actual speeds vary substantially based on vehicles suspension, individual driving abilities, and tolerance for gravitational forces.

In accelerating from a smaller to a larger radius, the rate of acceleration referred to in Chapter 4 could be applied with the same rate applying also to deceleration. The rate of acceleration or deceleration suggested is 0.85 m/s^2 as determined by the application of

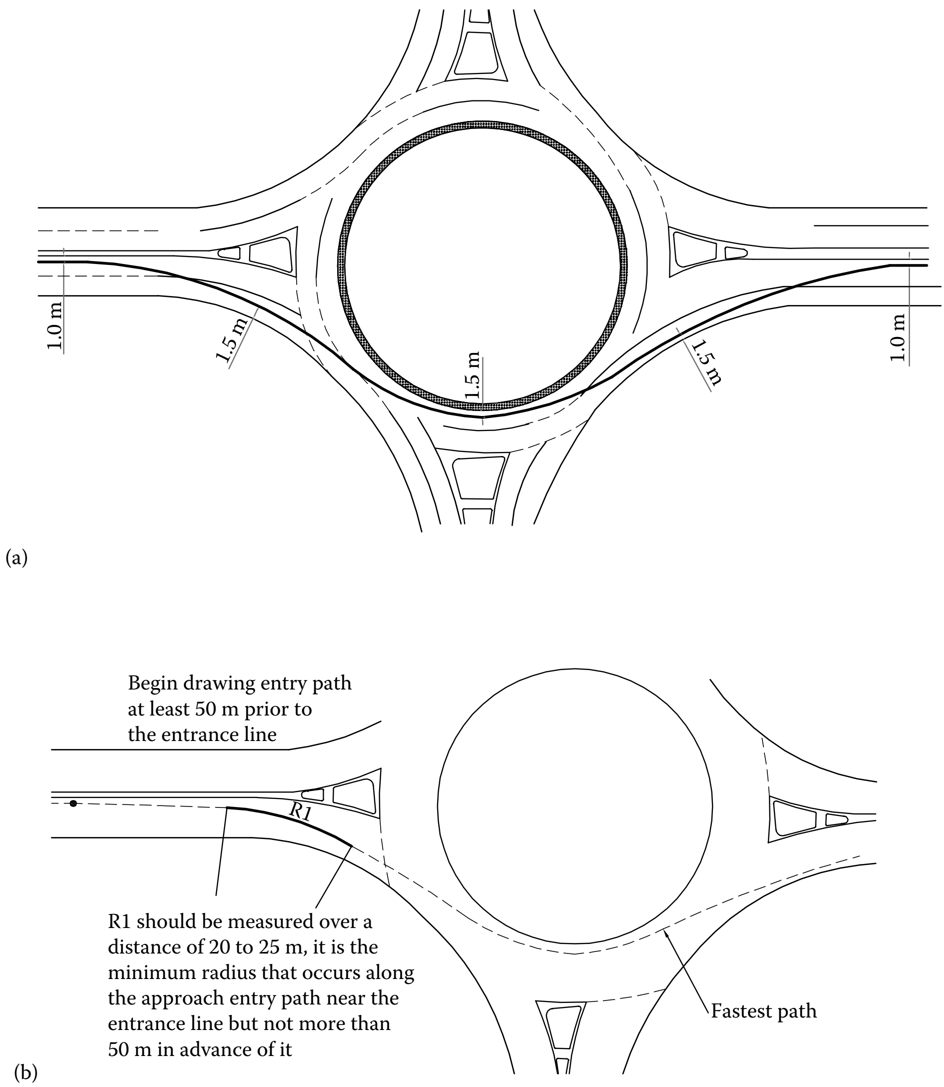


Figure 10.18 Construction of the fastest path: (a) typical offsets and (b) measuring the entry path radius.

car-following techniques in Europe (Lamm et al., 2001). The process to be followed is thus to calculate the speed appropriate to the various radii using the values of the coefficient of friction, *f*, illustrated in Figure 6.3 and then to check whether it is possible to achieve this speed by the end of the larger curve. In the case of deceleration to a smaller radius, the speed should be achieved at the start of the smaller curve.

10.11 SIGNALISED INTERSECTIONS

10.11.1 Introduction

As shown in Figure 10.1, a time comes when traffic volumes are such that priority control at intersections is no longer adequate. In effect, flows on the major road are sufficiently heavy to cause the gaps between vehicles to be inadequate in length or too infrequent to allow vehicles on the minor road to enter or to cross the major road. Although the vehicles on the major road would suffer no delay at all, the minor road traffic could experience serious queuing to the extent that the total delay at the intersection would be unacceptable. Some higher form of control is thus called for.

In this section, the warrants for provision of traffic signals are discussed as well as the geometry, specifically the cross-section, of signalised intersections.

As is the case with other forms of intersection control, the left turn always presents a major problem in terms of queuing and delays to through traffic moving in the same direction. Some unconventional intersection layouts have been developed in an effort to counteract this problem and these are also discussed.

10.11.2 Alternatives to signalisation

It is often believed that all problems at intersections can be resolved by installing traffic signals. In consequence, signals are sometimes installed where they are not needed. Smaller towns and villages seem to regard traffic signals as status symbols and are very quick to install them, especially if they qualify for state grants or subsidies against which the cost of installation can be offset. As traffic signals are expensive to install and maintain, the decision to use them should not be taken lightly.

Unwarranted traffic signals could affect the safety and efficiency of vehicular, cycle and pedestrian traffic. They can also lead to excessive delay which would cause road users to use other routes to bypass the traffic signals. In short, the traffic signal could generate rat running to the detriment of the quality of life in the surrounding residential areas. Because vehicular delay and the frequency of crashes are sometimes greater under signal control than under priority control, the alternative forms of control should be closely studied for application to a given site. The intersection layout should be fine-tuned in terms of site specific conditions even if one or more of the warrants for traffic signals are met. Only after comprehensive study should it be accepted that the only acceptable option is the traffic signal (*Manual of Uniform Traffic Control Devices*, Federal Highway Administration, 2009).

The alternative forms of control and fine-tuning elements include

- Warning signs on the approaches to the intersection
- Improving the sight distance to the crossing legs of the intersection by relocating the stop line as far forward as possible
- Installing traffic calming measures, specifically speed reduction, on the approaches to the intersection

- Installing flashing beacons either as a supplement to Stop control or in advance of the Stop controlled intersection
- Providing auxiliary turning and/or through lanes to reduce the number of vehicles per lane on the approaches to and through the intersection
- Providing channelising islands to define the location of conflict points in the intersection area and to define vehicle paths through the intersection area
- Providing median and refuge islands and neck downs (which are kerb extensions to shorten the distance that pedestrians have to walk to cross the road – also known as ‘bulb outs’)
- Installing street lighting if there are a disproportionate number of crashes at night
- Restricting one or more of the tuning movements either permanently or during certain times of day

Three- or four-way stops are an indiscriminate form of control as they force the vehicles equally on all legs of the intersection to stop. While the queuing problem of the traffic on the minor road is reduced if not eliminated altogether, the traffic on the major road is subjected to delay. Where there is a significant difference between the major and minor road flows, the delay to the major road vehicles could be greater than the time savings accruing to the minor road vehicles. Roundabouts are an equally indiscriminate form of control except that they are normally based on Yield as opposed to Stop control. Without going to the extent of replacing the intersection with an interchange or grade separation, the remaining option is to provide control by means of signalisation.

As in most forms of control, signalisation of an intersection always imposes a penalty on the major flows for the benefit of the minor flows. The only difference between them is that on the three- or four-legged Stop control or the roundabout there is a 50:50 split between entries to the intersection whereas the ratio of the split between the major and minor legs of the intersection can be modified in the case of signalisation to some other ratio supportive of minimisation of the total delay summed across all legs of the intersection.

10.11.3 The safety of signalised intersections

There is a widespread belief that signalisation is safer than other forms of intersection control. This is in fact not the case but what does happen is that the nature of crashes at signalised intersections differs from that at priority controlled intersections. The shunt, where a following vehicle hits the leading vehicle, arises, for example, from the leading vehicle slowing on receiving a yellow signal while the following driver believes that there is sufficient time to get through the intersection before the red phase starts. The 90° impact arises from a vehicle running a red light or the waiting vehicle edging into the intersection prior to its getting a green phase or both.

The dilemma zone is a well-known phenomenon being the situation of a vehicle approaching the signal as it changes to its yellow phase. The dilemma arises when the vehicle is too close to the signal to be able to stop at the stop line but too far from it to clear the intersection before the cross-street traffic starts moving. This phenomenon is typically associated with traffic signals on high-speed roads and is the major reason why traffic signals are an unsuitable form of traffic control where speeds are higher than about 80 km/h.

Studies indicate that more than 40 per cent of personal injury accidents at traffic signals on high-speed roads are rear-end shunts, and between around one fifth and one quarter of personal injury accidents at traffic signals in the United Kingdom are recorded as involving a red-running vehicle. The incidence of red-running tends to increase with approach speed.

Dilemma zone events account for a notable proportion of red-light violations at higher speed sites, and appear to account for a significant proportion of rear-end shunts.

Drivers' decision whether to stop or to continue are found largely to be based on their travel time to the stop line at the onset of the yellow signal. The boundaries of the stop–go decision are at about 2.5 seconds for a speed of 60 km/h and 4.5 seconds for a speed of 90 km/h. In urban areas, speeds are normally restricted to about 60 km/h, hence the general acceptance of a yellow phase of about 3 seconds. At 100 km/h, 90 per cent of drivers stopped if at a distance equating to a travel time of 5 seconds or more from the stop line (Maxwell and Wood, 2006).

10.11.4 Traffic signal warrants

Contrary to popular belief, a newly warranted traffic signal rarely reduces delay, costs, accidents or speed (Sampson, 1999) except possibly during peak hours. The fundamental purpose for the installation of traffic signals is to provide for the orderly control of traffic, and they are normally installed at intersections (New South Wales Road and Traffic Authority, 2011) to

- Provide control where there is a capacity or safety problem
- Provide a form of conflict resolution where there are high traffic flows
- Facilitate access to the road network from and to local areas
- Serve as part of an area-wide system of traffic management

As such, factors influencing the provision of traffic signals include

- Traffic flows and conflicts
- Crash statistics
- Access, vehicular and pedestrian, to major roads from local areas
- Feasibility of installation
- Cost of installation and maintenance in relation to availability of funding

The feasibility of installation includes that the posted speed should not be more than 80 km/h.

Apart from the dictionary definition of a warrant, it is defined as a necessary precondition for the provision of some or other facility. For example, a climbing lane is warranted by a drop in the speed of trucks through 15 km/h. The need for the warrant is that the provision of the facility requires additional funding and it is this that has to be justified. In the case of traffic signals, warrants are based on reduction of total delay at an intersection, maximisation of safety of all road users and network efficiency all of which have cost implications.

Ideally, reduction of total delay is best tested by simulation but, as this software is not always available to a designer, the warrant is often recast into a simple case of comparing actual and anticipated traffic flows with warranting flows. Warranting flows typically include those within a given time period or are experienced at peak hours.

Safety concerns addressed include those of turning traffic, specifically left-turning traffic. This is largely based on crash experience. The safety of other road users, principally pedestrians, also requires attention. Heavy traffic flows make it difficult, if not actually impossible, for pedestrians to cross the road or street. Long delays have been observed to lead to pedestrians attempting to cross streets one lane at a time, pausing on the lane marking while waiting for a gap in the following lane to be essayed. It is suggested that a lane marking does not constitute an acceptable pedestrian refuge. In the vicinity of schools, pedestrian traffic is particularly heavy at the start and end of the school day and, unfortunately, these broadly coincide with peak hour traffic flows.

Minimising of delay can be extended from consideration at single or isolated intersections to highly sophisticated area-wide adaptive signal control system systems such as Split Cycle Offset Optimisation Technique (SCOOT).

10.11.5 Typical signalisation warrants

Warrants for the installation of traffic signals are discussed in depth in the *Manual on Traffic Control Devices for Streets and Highways 2009* (MUTCD) or the latest version thereof. Most authorities worldwide either adopt American practice (MUTCD) or have their own sets of warrants. Examples of these include

- Southern African practice as defined by the Southern African Development Community Road Traffic Signs Manual (SADC RTSM): Volume 3, 2001
- Australian practice typified by those of the New South Wales Road and Traffic Authority (NSW RTA)

SADC RTSM suggests that traffic signals are warranted

- If the minimum vehicle volume entering the intersection from all legs is 8000 vehicles per 12-hour period while the minimum number of vehicles on the crossroad is 2000 for the 12-hour period
- When the intensity of the traffic on the main route is such that traffic on the secondary road experiences difficulties or undue delays entering the main route and when the traffic on the main route exceeds 9000 vehicles and that on the secondary road 1000 vehicles for a 12-hour period
- When the traffic on the main route at an intersection exceeds 600 vehicles per hour in any 4-hour period during the day and the pedestrian traffic exceeds 60 persons per hour during the same 4-hour period
- At consecutive crossings otherwise not conforming to the minimum norms for the sake of promoting progressive traffic flow
- When at least five serious collisions have occurred at the intersection resulting in injuries and/or serious damage to property
- At mid-block pedestrian crossings

NSW RTA offers as a guide that signals may be considered if any one of the following warrants are met for each of four 1-hour periods of an average day.

- Traffic demand
  - The major road flow exceeds 600 vehicles per hour on each direction.
  - The minor road flow exceeds 200 vehicles per hour in one direction.
- Continuous traffic
  - The major road flow exceeds 900 vehicles per hour in each direction.
  - The minor road flow exceeds 100 vehicles per hour in one direction.
  - The speed of traffic on the major road or limited sight distance from the minor road causes undue delay or hazard to vehicles on the minor road.
  - There is no other nearby traffic signal easily accessible to the minor road vehicles.

- Pedestrian safety
  - The pedestrian flow crossing the major road exceeds 150 persons per hour.
  - The major road flow exceeds 600 vehicles per hour in each direction.
  - In the presence of a median that is at least 1.2 m wide, 1000 vehicles per hour in each direction.
- Pedestrian safety on high-speed roads
  - The pedestrian flow crossing the major road exceeds 150 persons/hour.
  - The major road flow exceeds 450 vehicles per hour in each direction.
  - In the presence of a median that is at least 1.2 m wide, 750 vehicles per hour in each direction.
  - The 85th percentile speed exceeds 75 km/h.

In none of the three examples cited – MUTCD, SADC RTSM and NSW RTA – is the rationale behind the selected values offered.

10.11.6 Queue length as a warrant for traffic signals

Queue length has also been proposed as a warrant for the provision of traffic signals (Sampson, 1999). It is argued that queues address

- Vehicle, pedestrian and cyclist delays as well as the effect of mutual interference
- Difficulties that are experienced in crossing or entering traffic streams
- Gap acceptance behaviour
- The effects of different percentages of left and right turns
- The number of lanes to be crossed, with or without median island refuges for pedestrians or turning vehicles
- Geometric variables such as gradients and sight distance
- Human factor variables such as age and youth
- The forced acceptance of inadequate gaps

Practical advantages claimed for queue-based warrants are that

- Whereas vehicle, pedestrian and cyclists volumes cannot be equated or added, the same does not apply to queues. If it is accepted that pedestrians, cyclists and drivers of vehicles should have equal priority and be treated in the same way, then the queues are interchangeable and can be combined in producing a warrant.
- It is not necessary to make adjustments for different numbers of lanes, different speeds or different volumes as these will automatically be reflected in the queue length.
- The application of the queue-based warrant is simple because queue lengths are easily measured and verified.
- The warrant would be applied in peak hours, thus eliminating the need for off-peak measurements.

10.11.6.1 The derivation of the 4Q/6Q warrant

A queue length of 1 vehicle-hour per hour can also be expressed as one vehicle being in the queue for one hour. On this basis, it is suggested that an average queue length of 4.0 vehicle-hours per hour would ensure that the overall delay at the intersection is not worsened by the installation of signals. Queue lengths of four or more vehicles are likely to be seen only during peak hours. During off-peak hours, the delay resulting from the installation of the signals will be higher than at a priority-controlled intersection. This consequence will have to be accepted because it is suggested that an average queue length of four is about the limit of what is acceptable to the public.

Based on the assumption that a queue of one vehicle equates to a queue of one pedestrian, a test using simulation found that a combined queue of 10 suggested that the average delay to a waiting vehicle or person would be in the range of 60 to 500 seconds and a combined queue of eight would result in an average delay of 45 to 200 seconds. A queue of 6 corresponds to an average delay of 35 to 150 seconds in some of the delay warrants, notably MUTCD, South Africa, and Canada including those relating to pedestrians.

In typical conditions, the 6Q warrant would be met at volume conditions similar to those specified by MUTCD 1, 2, 8 and 9 while the 4Q warrant matches the peak hour warrant specified in MUTCD and the Canadian and South African warrants.

While the adoption of the 4Q/6Q warrant is recommended for technical reasons, it also has the benefit of public acceptability. Using traditional warrants, it is sometimes difficult, in the face of public demands, to justify the refusal to provide traffic signals. Congestion manifests itself in queues so that the requirement that queues be present before the provision of signals is easily accepted as being reasonable. Queues or the lack thereof are there for all to see so that the warrants cannot be manipulated for political expediency.

10.11.7 Warrant for pedestrian mid-block crossings

The installation of mid-block pedestrian crossings could be considered on the basis of delay to pedestrians waiting to cross because, as suggested previously, impatient pedestrian may decide to essay unacceptable crossing practices. The warrant would thus be based on the probability of the occurrence of a gap sufficiently long to allow a pedestrian to safely cross from one sidewalk to the opposite sidewalk.

The length of the required gap is dependent on the walking speed of the pedestrian and the width of the road to be crossed. Walking speeds range from 0.8 m/s for the elderly or for people with difficulty in walking to 1.7 m/s with the latter representing a brisk walk. The length of time that pedestrians could be reasonably be required to wait for an acceptable gap to present itself is a decision that should be taken by the road authority concerned and would have to be a compromise between pedestrian impatience and the impact of bringing vehicle flows to a standstill at too frequent intervals. Table 10.11 illustrates waiting times in terms of acceptability expressed as a level of service (LOS).

As an alternative to delay, the decision whether to provide a signalised pedestrian crossing could be based on the number of crossing opportunities that should be provided per hour. Random arrivals by vehicles would provide an indication of the number of crossing opportunities provided, and these can be modelled by a Poisson distribution.

$$P(X_t = x) = \frac{(vt)^x}{x!} e^{-vt}$$

The probability of no vehicles arriving in a period equal to the crossing time is thus, as discussed in Chapter 21, e^{-vt} if v is the average vehicular flow rate in vehicles per hour and t is the required time for crossing the street. The calculation of the flow rate above which traffic signals should be provided for various crossing times, that is, street widths, is shown in that chapter.

If flow is interrupted, by upstream traffic signals, for example, random flow breaks down and the Poisson distribution is no longer applicable. The need for a signalised pedestrian

Table 10.11 Pedestrian levels of service (LOS) for uninterrupted traffic flows

Average time between suitable gaps (s)	*LOS*	*Classification*	*Description*
<10	A	Excellent	Pedestrians are able to cross almost immediately upon arrival at the crossing point
10–20	B	Very good	Most pedestrians are able to cross with little delay Average delay ≤ 10 s 95th percentile worst case delay = 40 s
20–30	C	Satisfactory	Most pedestrians able to cross within an acceptable period Average delay ≤ 15 s 95th percentile worst case delay = 60 s
30–40	D	Some concern	Some pedestrians have to wait longer than desirable for a gap Average delay ≤ 20 s 95 percentile worst case delay = 80 s
40–80	E	Major concern	Most pedestrians have to wait longer than desirable for a gap Average delay ≤ 40 s 95th percentile worst case delay = 160 s
>80	F	Unsatisfactory	Most pedestrians have to wait longer than is acceptable for a gap Average delay ≤ 10 s 95th percentile worst case delay > 40 s

Source: Queensland Department of Transport and Main Roads. *Traffic and road use management manual (TRUM)*. Brisbane, 2010.

crossing should thus be determined by observation but using values of delay as described in Table 10.10. The distance between the upstream signalised intersection and the proposed pedestrian crossing would dictate the extent to which platoons had dispersed prior to reaching the crossing point. If platoons are totally dispersed, random flow has been reinitiated. On the other hand, if traffic is still in a platoon at the crossing point, there would be adequate gaps in which pedestrians could cross without the need for signalisation of the pedestrian crossing.

According to MUTCD the calculation of the warrant for pedestrian signals comprises a two-step process. The pedestrian clearance time, typically given by a flashing or 'Don't Walk' sign, is calculated on the basis of a walking speed of 1.1 m/s over a distance measured from kerb to kerb. The total pedestrian cycle, which includes Walk and Don't Walk, is based on a walking speed of 0.9 m/s for a crossing measured from the top of the sidewalk ramp to the top of the sidewalk ramp at the far side kerb. These walking speeds have implications for geometric design because shortening the crossing distance by using kerb bulb-outs or narrower lanes can reduce the time for the pedestrian walk phase, thereby increasing the time available for opposing vehicular travel.

10.11.8 Geometric design of signalised intersections

The geometric design of signalised intersections does not differ significantly from that of unsignalised intersections.

The angle of skew becomes less of an issue at signalised intersections because opposing traffic would be stopped during times in which the through traffic would be allowed to move. This, however, does not make provision for the possibility that the traffic signals may not be working, for example, because of a power failure, or may be operating in Flashing Yellow mode whereby the intersection would be operate as a three- or four-way Stop controlled intersection. So, although less of an issue than at an unsignalised intersection, a decision to accept an angle of skew outside the normal limit of 60° to 120° should not be taken lightly.

Any road with traffic volumes sufficiently high to warrant traffic signals should not have lane widths less than the maximum adopted by the road authority concerned. And this should be in the range of 3.6 to 3.7 metres. Increasing the lane width significantly beyond this range may result in drivers of passenger cars starting to use a single lane as two lanes. It has been observed that a lane that is 5.0 metres wide often functions in the peak period as two lanes that are 2.5 metres wide. This results in motorists having to signal their intention to move from one side of the lane to the other!

Turning traffic often exerts an undue influence on the efficiency of intersections so that, at the flows warranting the installation of traffic signals, auxiliary lanes will often be required as part of the cross-section of signalised intersections.

10.11.9 Channelisation of a signalised intersection

It doesn't automatically follow that a signalised intersection has to be channelised, as this is mainly a function of the volume of turning traffic at the intersection. On a commuter route, the major flow would be straight through most intersections and the turning volumes from the minor streets, even though possibly a high percentage of the minor street traffic, would probably be low in total.

Channelisation is useful in accommodating turning traffic and, perhaps more importantly, ensuring that turning vehicles do not cause unacceptable delays to through traffic. In heavy flows, a single left-turning vehicle can cause a major queue to develop behind it. Through drivers wishing to escape from being trapped in the queue attempt to merge into the adjacent lane, and this results in significant levels of upstream turbulence. An auxiliary lane accommodating left-turning vehicles would resolve the problem and this is discussed below.

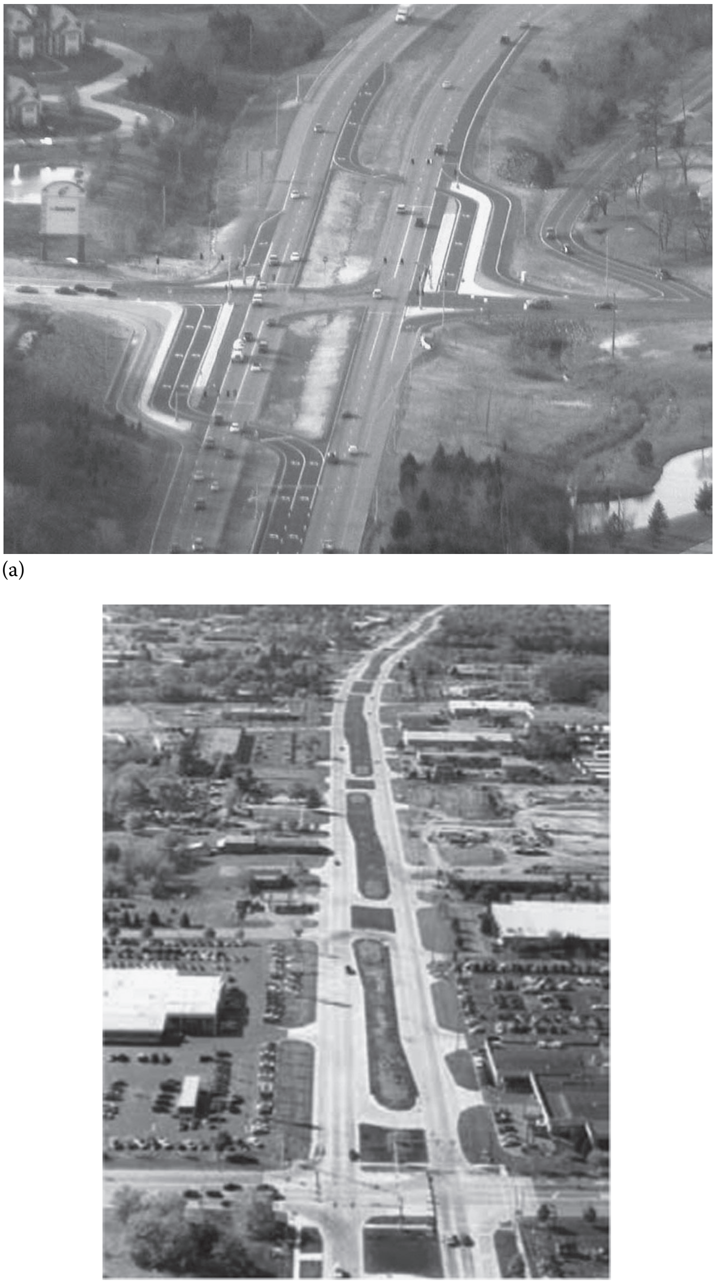
10.11.10 Accommodation of the left turn

The left turn is a major impediment to the efficiency of operation of any intersection. Various ways of improving the situation include leading or lagging signal phases, single or multiple left-tuning lanes and a large variety of unconventional layouts. These include, in addition to the continuous flow intersection (also known as the crossover displaced left-turn intersection), the upstream signalised crossover and the half upstream crossover intersections

- The double left-turn intersection
- The Michigan median U-turn (MUT) design
- The restricted crossing U-turn
- The quadrant roadway intersection

In general, these layouts, illustrated in Figures 10.19 and 10.20, remove left turns from an intersection locating them elsewhere.

The continuous flow intersection moves the left turn upstream or downstream to an adjacent minor intersection where the crossing flows are lighter in volume. As an alternative, in the upstream signalised crossover intersection entire traffic flows are switched so that the driving on the right becomes driving on the left where, obviously, the left turn ceases to be a problem. This is similar to the operating principle of the diverging diamond interchange. Drivers have been known for many years to solve the problem of the left turn by turning



(a)

(b)

Figure 10.19 Unconventional left-turn treatments: (a) double left turn intersection and (b) Michigan median U-turn. (From Hughes W et al., *Alternative intersections/interchanges: Informational report*. Report FHWA-HRT-09-060, Federal Highway Administration, Washington, DC, 2010.)

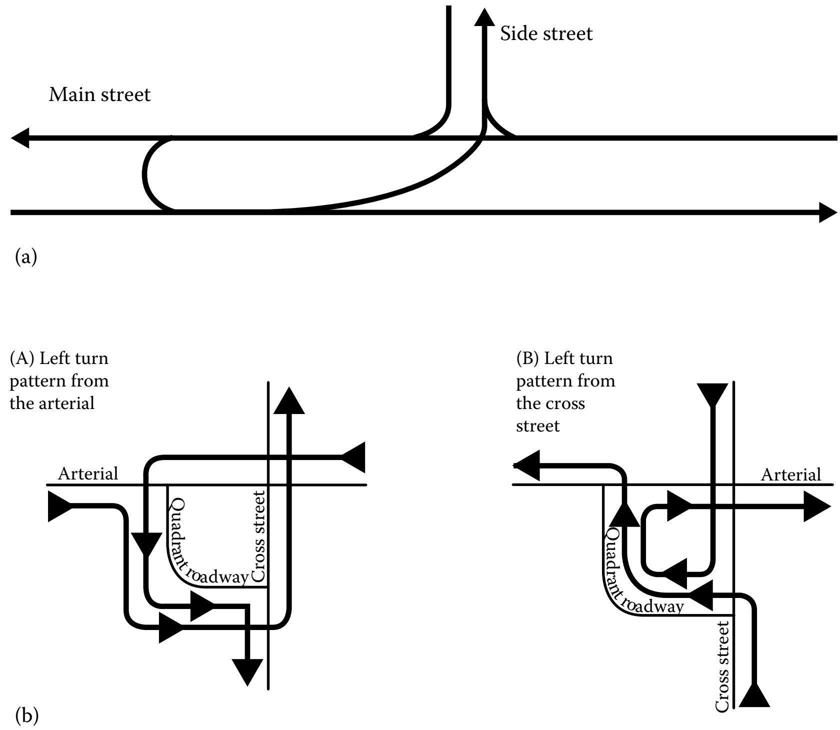


Figure 10.20 Further unconventional left-turn treatments: (a) restricted crossing U-turn and (b) quadrant roadway intersection. (From Hughes W et al., *Alternative intersections/interchanges: Informational report*. Report FHWA-HRT-09-060, Federal Highway Administration, Washington, DC, 2010.)

right onto the cross-street and doing a U-turn there even where these are illegal. Provided city blocks are small, drivers sometimes also convert a left turn into three successive right turns around the block following the intersection at which the left turn is desired.

10.11.10.1 The storage area for left-turn lanes

The components of the left-turn lane include a taper and a storage area. The length of the storage area is dependent on the number of vehicles wishing to turn during a signal cycle. As a general rule, about three or four vehicles can execute their left turn during the yellow and all-red phases of the cycle. The storage area should thus be sufficiently long to accommodate these vehicles. Allowing for a gap of 1 m between waiting vehicles, the storage area should be at least 20 m long.

The Texas Department of Transportation (Lei et al., 2007) offers a rule of thumb method whereby

$$L = 0.3K \frac{V}{N_c} S \qquad \text{for signalised intersections and}$$

$$L = 0.3K \frac{V}{\frac{3600}{I}} S \qquad \text{for unsignalised intersections}$$

where

L = storage length (metres)
V = left-turn flow rate during the peak hour (vehicles per hour)
K = constant to reflect random arrival of vehicles (usually 2)
N_c = number of cycles per hour (for signalised intersection)
I = average vehicle waiting interval in seconds
S = average queue storage length per vehicle (measured from front bumper to front bumper)

The average storage length per vehicle (S) based on the percentage of trucks in the left-turning traffic stream is shown in Table 10.12.

As mentioned previously, random arrivals can be modelled by the Poisson distribution and this applies equally to the arrival of left-turning vehicles at an intersection. Various road authorities have adopted this approach to the determination of the storage length of a turn lane. Some authorities adopt the criterion that the probability of arrivals not exceeding the turn lane capacity during a given time interval will be less than or equal to some or other given percentage. The percentage selected is usually of the order of 95 to 98 per cent. The Poisson distribution is expressed as

$$P(X_t \le k) = \sum_{x=0}^{k} \frac{1}{x!} (\upsilon t)^x \, e^{-\upsilon t}$$

where

k = number of left-turning vehicles
υ = the mean arrival rate
t = the cycle length

To illustrate: If

- The probability, $P(X_t \le k)$, is 0.96
- The mean arrival rate is found to be 100 vehicles per hour (100/60 vehicles per metre)
- The cycle length is 1 minute

the equation becomes

$$0.96 = \sum_{x=0}^{k} \frac{1}{x!} \; \frac{100}{60} \cdot 1 \;^{x} \; e^{-\frac{100}{60} \cdot 1}$$

Table 10.12 Queue storage length per vehicle

Percentage of trucks	*Storage length (m)*
<5	7.6
5–9	9.1
10–14	10.7
15–20	12.2

This has to be solved by trial and error with x taking values between 0 and k, and the relationship summed. If $k = 3$ then the total probability is calculated by summing the probabilities for $x = 1$, $x = 2$ and $x = 3$.

If $k = 3$,

$$P(X \leq 3) = 0.91$$

If $k = 4$

$$P(X \leq 4) = 0.968$$

The left-turn bay must therefore be able to accommodate four vehicles for the design requirement to be satisfied. The queue length per vehicle can be read off Table 10.12. With 5 per cent trucks in the traffic stream it follows that the required storage length is 36 metres.

10.12 RAILWAY AT-GRADE CROSSINGS

Railway at-grade crossings, often referred to as level crossings, are simply another form of priority controlled intersection. Drivers are required to stop or yield right of way to opposing traffic and, at these intersections, right of way is always in favour of the train. The distance required to bring a fast-moving train to a standstill is not trivial and the drivers of road vehicles thus bear most of the responsibility for avoiding collisions with trains (Ogden, 2007).

As a general rule, new grade crossings are not encouraged. This is particularly the case on mainline tracks. In some instances, train speeds are increasing to beyond the flying speeds of light aircraft. Where high-speed trains are operated in the range of 200 to 350 km/h, the probability of drivers having sufficient sight distance along the track is vanishingly small. The consequences of a crash between a high-speed or bullet train and a road vehicle approach the enormity of a massacre and it is thus necessary to provide grade separations over or under the railway lines.

In less dramatic situations, grade separations are warranted in terms of economic analysis. Analysis can be in terms of either benefit/cost ratio or net present worth. Benefits involve prediction of delay and crash costs that would be eliminated at a level crossing over the life of the alternative structure compared to the whole-life cost of a grade separation that would totally eliminate both. There is always the possibility that the grade separation would also have its fair share of crashes such as vehicles leaving the road on a high fill or hitting the bridge balustrades, but including these in the analysis is believed to be an unnecessary refinement.

The probabilities of delay and crashes are functions of the volumes of road and rail traffic. The number of crashes is usually calculated as the product of the rail and road traffic multiplied by some or other prediction factor. Delay is the product of the number of vehicles forced to queue and the duration of these queues. In addition to time, delay costs include the cost of idling engines. It is pointed out that both road and rail traffic volumes increase with time, and the development of reasonably reliable models of total benefits is not a trivial exercise. Many road authorities have attributed unit costs to the value of time and crashes, usually drawing a distinction between fatal, injury (severe and light) and property damage only crashes. And these data have to be acquired as part of the analytical procedure.

10.12.1 Rail crossing geometry

The geometry of railway lines is very restricted. Curve radii are high, being in the range of 2000 to 7000 metres in the case of high-speed operation, although dropping to as little as 700 metres for speeds of less than 120 km/h. Gradients cannot be steep and are usually limited a maximum value of 4 per cent and as little as 1.5 per cent for freight trains. Superelevation of rail tracks is limited to between 2.5 and 5 per cent. The geometric design of highways is significantly more flexible so that it is the geometric constraints of rail design that would dictate the final layout.

A major component of the crossing consists of the physical aspects of the highway on the approach and at the crossing itself. The following roadway characteristics (Ogden, 2007) are relevant to the design of highway-rail grade crossings:

- Traffic volumes
- Sight distance
- Angle of skew
- Horizontal and vertical alignment
- Cross-section
- Nearby intersecting highways

10.12.2 Traffic volumes

Traffic volumes are a significant input into the decision of whether a grade separation should be provided in preference to a rail crossing. Volumes relate both to road and to rail traffic.

In urban areas, speeds are low and volumes high. High volumes, and hence frequency of arrivals, of trains will inevitably disrupt the smooth flow of road traffic, causing significant delay and queues backing up onto the other roads in the network, spreading congestion far and wide. High volumes of road traffic can similarly lead to the formation of queues, particularly if the arrival of trains at the crossing is at short intervals.

In rural areas, while speeds are high, volumes are low. Rail traffic is infrequent and may be as little as one or two trains per day. Under these circumstances, it is unlikely that grade separations would be economically justified. The high speeds would generate a need for generously sized sight distance triangles, particularly if vehicles are not forced to stop at the rail crossing.

10.12.3 Sight distance

As is the case with other intersections, as discussed in Section 10.6.2, two sight triangles are of interest. The first, called for convenience Case A, is the approach sight triangle. This includes the distance from the rail crossing at which a driver must begin to decelerate in order to stop at the stop line short of the rail crossing and the distance between the approaching train and the crossing. The legs of the sight triangle vary according to the speeds of the road vehicle and the train. Figure 10.21 illustrates the sight triangle to be provided at a railway grade crossing for Case A.

In Figure 10.21, the various symbols have the following significance:

d_H = sight distance leg along the highway allowing a vehicle to safely stop without encroachment of the crossing area (m)
d_T = sight distance leg along the railway for approaching vehicle (m)
D = distance from the stop line to the nearest rail (assumed to be 5.0 m)
d_e = distance of driver behind the front of the vehicle (assumed to be 3 m)

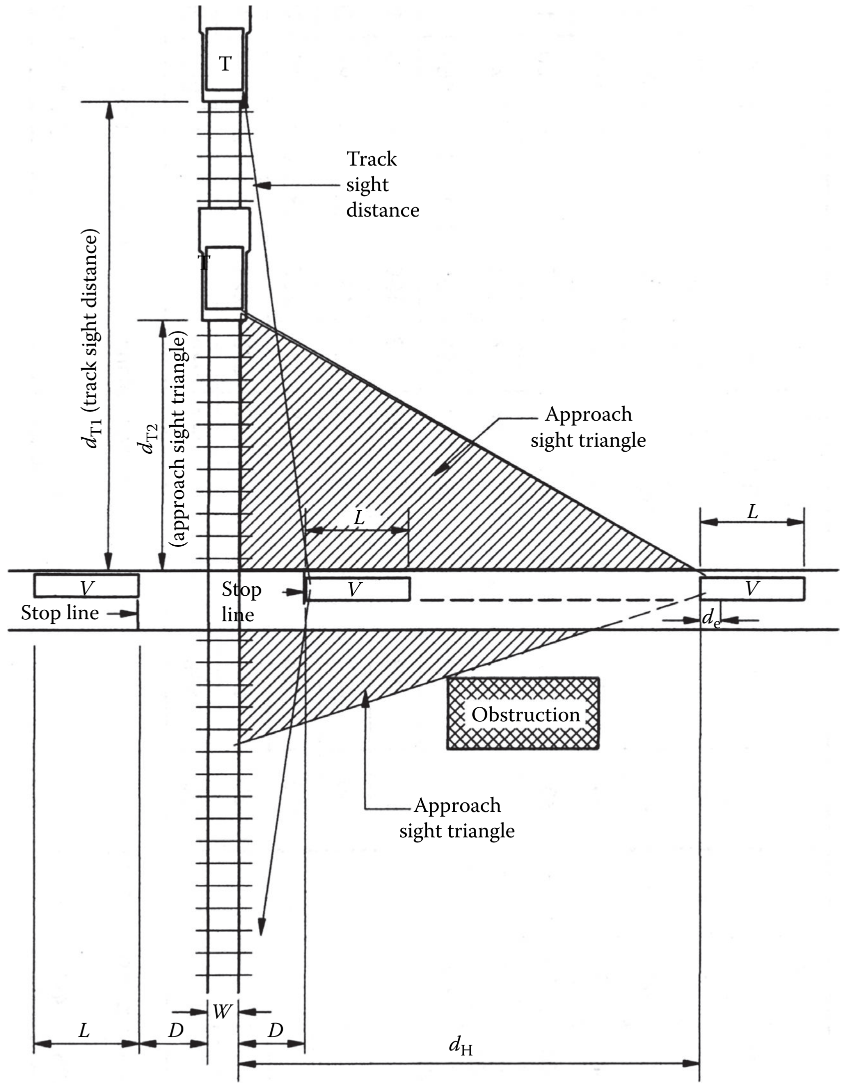


Figure 10.21 Sight distances at railway grade crossings for Case A. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

The values of the sides of the sides of the sight distance triangles are shown in Table 10.13.

The second case, Case B, is the departure sight triangle. This case presumes that the road vehicle pulls away from a stationary position at the stop line and has to cover the sum of the distance

- From this stop line to the nearest rail
- The gauge of the track (in the case of a single track – for multiple tracks the distance to be cleared is the sum of the widths of all the tracks plus the sum of the distances between them)
- The length of the vehicle
- The distance between the furthest rail and the far stop line

Table 10.13 Values of Case A sight distance legs for various vehicle and train speeds

Train speed (km/h)	Distance along railroad from crossing (m)									
	Vehicle speed (km/h)									
	40	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*	*130*
10	14	15	16	17	18	20	21	22	23	24
20	28	30	32	34	37	39	42	44	46	49
30	42	45	48	52	55	59	62	66	70	73
40	57	60	64	69	73	78	83	88	93	98
50	71	75	80	86	92	98	104	110	116	122
60	85	90	97	103	110	117	125	132	139	147
70	99	105	113	120	128	137	145	154	163	171
80	113	120	129	138	147	156	166	176	186	196
90	127	135	145	155	165	176	187	198	209	220
100	142	151	161	172	184	195	208	220	232	245
110	156	166	177	189	202	215	228	242	255	269
120	170	181	193	206	220	234	249	264	279	294
130	184	196	209	224	239	254	270	286	302	318
Distance along highway from crossing (m)										
	57	75	97	120	147	176	208	242	279	318

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets.* Washington, DC, 2011a.

The values offered in Table 10.14 are based on the assumptions of

- A track gauge (W) of 1.5 metres
- A distance (D) of 5 metres from the stop line to the nearest rail
- The length (L) of the vehicle, assumed to be 22 metres
- The maximum speed (V) of a vehicle in first gear of 2.7 m/s
- A rate of acceleration (a) of 0.45 m/s^2
- A time (J) of 2.0 seconds required to get the vehicle in motion from rest

The total distance the vehicle has to travel to clear the track is the sum of:

• The distance from the stop line to the first rail	5.0 m
• The gauge of the track	1.5 m
• The distance from the second rail to the far stop line	5.0 m
• The length of the vehicle	22.0 m
	33.5 m

The time taken to cover this distance is the sum of the time taken to

• Get the vehicle in motion	2.0 s
• Accelerate to 2.7 m/s from rest	6.0 s
• Complete the balance of the distance	9.4 s
	17.4 s

The balance of the distance derives from the fact that during acceleration the vehicle will travel a distance of 8.1 metres, leaving a balance of 25.4 metres to be covered at a speed of

Table 10.14 Values of Case B sight distance legs for various vehicle and train speeds

Train speed (km/h)	*Sight distance along track (m)*
10	48
20	97
30	145
40	193
50	242
60	290
70	338
80	387
90	435
100	484
110	532
120	580
130	629

2.7 m/s. The distance travelled by the train for during this time for various speeds is offered for convenience in Table 10.14.

10.12.4 Angle of skew

The road should cross the railway line(s) at an angle as close as possible to a right angle. As is the case, in fact, in an intersection between two roads and for the same reason, to wit to ensure that drivers can easily observe the rail track in both directions. The angle of skew should ideally fall within the range of 75° to 105°.

10.12.5 Horizontal alignment

The road approaching the crossing should preferably not be on a curve. The superelevation required for the curve is unlikely to match the gradient of the railway. In consequence, the cross-section of the road would have to be warped causing an uncomfortable crossing of the railway line. In the extreme case of a 10 per cent superelevation and a 4 per cent down gradient in the opposite direction on the railway line, the loss of control of the vehicle is almost guaranteed. Although a broken-back curve, albeit unsightly, could resolve this problem, the length of tangent would have to be sufficient to be perceived as such.

A curve that is such that the superelevation at the crossing point, either on the curve itself or on the superelevation development, precisely matches the gradient of the railway would, of course, be eminently acceptable.

Regardless of what the gradient of the railway is, the horizontal alignment of the road would have to include a section analogous to the superelevation development described in Section 6.4.7. It would be necessary to rotate the cross-section to replace the camber with a crossfall and then to rotate this crossfall to match the gradient of the railway. If the gradient of the railway is less than the camber of the highway, the two sides of the camber could be rotated simultaneously in opposite directions to achieve the appropriate crossfall.

10.12.6 Vertical alignment

Ground clearance (sometimes also referred to as ride height) is the lowest height between the bottom of the tyres and the lowest point on the body of the vehicle. In the case of passenger cars, manufacturers typically quote a ground clearance of the order of 150 to 200 mm. This is, however, measured without any load or passengers (including the driver) in the vehicle and an empty fuel tank. Any car in normal service would thus be 25 or even 50 millimetres lower.

The probability of a vehicle scraping its bottom is also dependent on the wheel base. The longer the wheelbase for a given ground clearance, the greater is the probability of its scraping. This largely has to do with the likelihood of the difference between two gradients being of sufficient magnitude to cause the vehicle even to hang up on the change of gradient.

The two rails define a plane surface, and the vertical alignment of the road must precisely match this surface, which should also extend a short distance on either side of the track. Typically a distance of 600 to 1000 millimetres would be sufficient to ensure that the rails are not proud of the road surface. The material used in the construction of this length of the road should, for preference be concrete both for durability of the road surface and for the ease of construction with a minimum of damage to the rail tracks. Outside the limits of this plane surface, a modest gradient of less than 1 per cent could be considered. This grade should extend for the length of the wheel base of the longest vehicle likely to use the crossing. It is offered that, in the case of a WB-19 semitrailer, the distance between the rear wheels of the tractor and the rearmost wheel of the rear bogie of the trailer would be of the order of 8 to 10 metres. A grade of this length would be sufficient to ensure that vehicles crossing the track would not scrape or hang up on it. Outside the bounds of the rail crossing, the normal standards of vertical alignment would apply.

For abnormal vehicles, such as those with 50 or more wheels under the trailer and used for the transport of massive loads such as stator casings found in power stations, a template of the vehicle would have to be compared with the profile of the rail crossing. It is pointed out that if one of these vehicles were to hang up on the railway crossing, removing it would be virtually impossible.

10.12.7 Cross-section

The cross-section of the road outside the bounds of the rail crossing would apply equally to the crossing itself. Where booms are used as a traffic control device, it is, however, noted that some drivers are inclined to move onto the wrong side of the road to drive around them. This can be eliminated be the provision of booms long enough to span the entire width of the road but these would require bases sufficiently sturdy to counteract the bending moment generated by booms of this length. A more practical option, which also has the benefit of creating a target value for the crossing, making it more visible to approaching drivers, is to include a kerbed median island in the layout of the crossing. This island should extend to a point sufficiently clear of the tracks to ensure the safe passage of trains but not so far from the tracks that drivers can still manoeuvre around the end of the island.

10.12.8 Network considerations

The space between a rail crossing and other intersections in the network requires careful consideration. Transportation corridors often have railways and roads parallel to each other over considerable distances. This creates the possibility of minor roads crossing the railway and very shortly thereafter terminating at or crossing the major road. It is essential to ensure that there is sufficient space between the crossing and the following intersection to provide adequate storage space for vehicles waiting to enter or cross the major road, that is, to avoid their being trapped on the rail crossing.

Chapter 11

Freeways and interchange design

CONTENTS

11.1 INTRODUCTION

Roads are classified in various ways as described in Chapter 3. To recapitulate, these are

- Functional classification, which is the grouping of highways by the character of service they provide across the spectrum of movement ranging from mobility, the highest order of which is the freeway, to direct land access.
- Design type classification, which is based on the major geometric features of roads, that is, freeways, conventional highways or streets.
- Administrative classification, which is classification by route numbering and, in addition to defining the authority responsible for a particular road, includes application to route guidance for navigation by tourists and others unfamiliar with the layout of the road network. With regard to guidance, reference is made to numbered and unnumbered routes. All freeways are numbered routes.

Freeways are defined as a class of design types and provide the highest order of mobility in the functional spectrum of movement. They are discussed in Section 11.2. Interchanges, as the only form of permitted access to freeways, are discussed in the sections that follow thereafter.

11.2 FREEWAYS

11.2.1 Introduction

The prime function of freeways is to provide safe and convenient movement of large volumes of vehicles at high speeds over extended distances (AASHTO, 2011a). They are characterised by full control of access so that the word ‘freeway’ signifies an absence of interruption of flow caused by intersections and not, as widely believed, an absence of tolling. Toll roads are thus also considered to be freeways even though the process of tolling, if not electronic, may cause interruption of free flow. Control of access refers to access to and from abutting properties. Access to freeways is restricted to entry via a limited number of roads and streets, which are generally high in the spectrum of functionality in terms of mobility.

The integrity of local road networks is maintained by the provision of grade separations where crossings are necessary. Streets other than those from which access to the freeway is provided or where grade separations are planned to be provided either have to be converted into culs de sac or linked by frontage roads (Bonneson et al., 2003).

Frontage roads are normally one-way and preferably in the same direction as the adjacent carriageway of the freeway. The reason for this preference is that drivers, having misread the outer separator as being the median island, may be confused or possibly even stressed by having traffic approaching on their right. The glare of the approaching headlights at night could compound the confusion. The normal human factors expectation is for approaching traffic to be to their left.

The service road is parallel to the freeway and usually has a row of properties between it and the freeway. It could therefore accommodate two-way traffic flow like any other street.

As freeways are intended to serve uninterrupted flows of traffic, it follows that, whether the freeway is located in an urban or a rural area, railroad crossings would be by means of grade separations.

In the distribution of functionality between mobility and accessibility, the freeway is, as illustrated in Figure 4.1 in Chapter 4, the highest form of mobility. This is achieved by the replacement of the stop-and-go flows at intersections by merging and diverging manoeuvres at high absolute, but very low relative, speeds.

The design of freeways involves consideration of the elements of concern also of the roads described in the preceding chapters, namely the horizontal and vertical alignments and the cross-section. They are thus repeated here only to the extent that they differ in the case of freeways from those of other roads. In addition, freeways include ramps providing access to and from freeways to the local road networks.

11.2.2 General design considerations

11.2.2.1 Design speed

Given that freeways are intended to serve high volumes of traffic moving at high speeds, it follows that the selected values of design speeds should also be high. As a general rule, design speeds should be as high as are achievable in terms of economic restraints. When applying economic analysis to the selection of an appropriate design speed, two aspects of the analysis should be borne in mind:

- The whole-life benefits and costs are to be included in the analysis.
- There isn't a direct correlation between design standard and construction cost, although, very loosely, higher standards may result in higher construction costs.

In the case of rural freeways, design speeds typically are in the range of 120 to 130 km/h, although in mountainous areas lower design speeds may be necessary for economic reasons. In these areas, design speeds of 80 to 110 km/h would have to be accepted but only reluctantly.

In the more constricted urban areas, lower design speeds are often acceptable and reference would once again be made to the range of 80 to 110 km/h. Freeways should not have a design speed lower than 80 km/h. Where a design speed of less than 100 km/h has to be accepted for whatever reason, a posted speed limit should be applied, with the signposting being repeated at about 5-kilometre intervals as a reminder to drivers. Low design speeds on urban freeways are acceptable because travel distances are relatively short so that little time is saved by providing a high design speed. Furthermore, urban interchanges are often closely spaced, resulting in substantial levels of turbulence. These would require increased vigilance by drivers. Forcing slow travel speeds by design would have benefits with respect to road safety in terms of the additional reaction time provided. Lower design speeds would also have the effect of reducing the speed differential between nonweaving and weaving vehicles, which would also have a favourable impact on road safety.

11.2.2.2 Levels of service

Levels of service (LOS) are dealt with exhaustively in the *Highway Capacity Manual* (HCM; Transportation Research Board, 2010). Rural freeways are usually designed to achieve LOS B in the design year, with this typically being 20 years in the future. The higher traffic volumes in urban areas make this LOS difficult, if not actually impossible, to achieve and the best that can be hoped for is LOS C. In practice, particularly closer to the central business district (CBD), LOS D would probably have to be accepted as the design LOS.

11.2.2.3 Horizontal alignment

The horizontal alignment of rural freeways should be characterised by long tangents and gentle curves. If possible, a curvilinear alignment with a high design speed should be aimed for. Aesthetically, independent alignments of the carriageways are visually pleasing, and a variable width of median of 50 metres or more supports the independence both horizontally and vertically of the carriageways. The natural topography, trees and rocks and even streams can be retained and help create a parklike appearance for the freeway.

In the extreme, and where space and land use permits, the carriageways could be totally independent of each other to the extent of being contained within individual road reserves. The opposing carriageway should, however, be visible from time to time to reassure drivers that they are on a dual carriageway. Although this may be aesthetically pleasing, it could create havoc in areas of pastoral or intensive agricultural activity. Furthermore, as discussed in Chapter 21, road networks cut the natural environment into smaller and smaller parcels as they develop. This is inimical to the preservation of wildlife, as animals need to be able to range freely. Being confined to increasingly restricted areas could threaten them with extinction. Separate road reserves would exacerbate the problem, thus calling for careful consideration of the maintenance of the integrity of animal habitats.

This freedom of choice regarding the location of the two carriageways relative to each other is invariably not available in the case of the urban freeway and the median width would normally remain constant.

11.2.2.4 Vertical alignment

The vertical alignment of rural freeways is dictated largely by topographic conditions and drainage and to a lesser extent by the need to provide frequent vertical separations to crossing roads and interchanges. They can thus be constructed at or near ground level.

In broken topography, long and high viaducts may, however, span valleys if traffic volumes are sufficiently high to make them economically viable. An example is the Millau Viaduct, a cable-stayed bridge that is shown in Figure 11.1. It is on the national route between Paris and Montpellier and is the tallest bridge in the world, with one of its piers situated 343 metres above its base. The deck is, on average, 270 metres above ground level and has a total length of 2460 metres supported by seven piers.

The vertical alignment of urban freeways, on the other hand, is controlled by the need to fit in with the existing local street networks. The need to provide access or systems

Figure 11.1 Milau Viaduct. (From Wikipedia 2013.)

interchanges at relatively short intervals, which can be as low as 2 to 3 kilometres between the centrelines of crossing roads, is a further constraint on the vertical alignment of the freeway.

As in the case of conventional roads, stopping sight distance and headlight sight distance, where freeways are unlit, will be major controls of the vertical alignment.

11.2.2.5 Cross-section

Freeways are usually dual carriageways although, in a phased development, initial construction may be limited to only one of the carriageways. The carriageway selected for construction should be offset to allow for symmetrical doubling up for the second carriageway and also for the width of the median. Phasing could include interchanges being replaced by at-grade intersections as an interim measure but the design should be such that, when the interchange is to be constructed, the location of the interim intersection should make construction of the interchange possible with a minimum of abortive cost.

Being intended for high-speed travel, lane widths should be in the range of 3.6 to 3.7 metres. Being a dual carriageway, each carriageway should have a minimum of two lanes.

The outside shoulder should be 3.0 metres wide. If surfaced, it could be used by vehicles moving onto the shoulder to allow faster vehicles to overtake them. Obviously, if surfaced, the full pavement design layers should extend across the shoulder to accommodate the trucks that are most likely to move onto the shoulder for the benefit of overtaking traffic. In some parts of the world, driving on the shoulder is not permitted by law and, in others, it is permitted but only during the hours of daylight. A third practice is to open the shoulder as an additional traffic lane during peak hours. This practice is not popular with emergency services, who claim the need for use of the shoulder as an emergency lane, allowing them to move rapidly to crash sites, for example.

Inner shoulders could be as little as 1.5 metres wide. Where a carriageway has three or more lanes, the inner shoulder width should be increased to 3.0 metres. This is to provide a refuge for disabled vehicles without the driver having to negotiate two lanes of fast-moving traffic, possibly with a dead engine, in seeking the sanctuary of the outer shoulder.

As a general rule, out-of-control vehicles leaving the road seldom travel farther than about 9 metres beyond the edge of the travelled way. Medians wider than this would thus have little merit in terms of enhancement of road safety. As stated previously, the argument has been raised that two vehicles travelling in opposite directions and simultaneously leaving their respective travelled ways to achieve a head-on crash would need a median that is 18 metres wide. The increase in the total area of the road reserve over a distance of 10 kilometres amounts to 9 hectares lost to farming forever. The probability of this head-on crash ever occurring is small and can safely be ignored.

If the topography requires that a split grading, that is, the two carriageways being independently graded, be provided, the median width required to accommodate the height difference between the two carriageways would have to be determined. A width of 5 metres would be required to accommodate two shoulders each 1.5 metres wide, with a foreslope and a side drain each 1 metre wide on the lower carriageway. A side slope of 1*V*:6*H* would require an additional 6 metres of median width and hence a total median width of 11 metres to accommodate a height difference of only 1 metre.

Medians with a side slope always raise the prospect of the driver of an out-of-control vehicle not being able to avoid entering the opposing carriageway. A crash at a relative speed of 240 km/h is guaranteed to result in several casualties, many of them fatal. Split gradings should thus always be provided with safety barriers as discussed in Chapter 13. If safety barriers are provided, steeper side slopes could be considered. A median width of 11 metres providing the minimum shoulder widths, foreslope and side drain suggested previously and with a side slope of 1*V*:1.5*H* could accommodate a height difference of 4 metres.

Horizontal and vertical clearances should be as discussed in Chapter 8.

11.3 DEPRESSED AND ELEVATED FREEWAYS

11.3.1 Introduction

In urban areas, the local street system that has evolved over several years will largely be at or as close as possible to the natural ground level. Pushing a freeway through this system at ground level could require those local streets, which have to cross the freeway, to be located on grade separations either over or under the freeway. Because of the distance required to achieve a height difference of about 6 to 7 metres, several properties that previously had direct access to the crossroad would either lose part of their frontage to make provision for a cross-section on fill or in cut or have to be acquired in their entirety. The alternative would be to provide the freeway with an undulating or roller coaster alignment as it climbs over or drops under successive crossroads. As suggested in Chapter 9, this is unacceptable.

The cross-section of the crossroad would usually not be as wide as that of the freeway and the design speed lower. The cost of construction of the earthworks and structure over a grade separation with the crossroad passing over the freeway would thus be less than that of the freeway passing over the crossroad. Depending on the market value of properties to be acquired either in toto or to the extent of loss of frontage, it is possible that, even including the cost of land acquisition, having the crossroad passing over the freeway would be the more economical solution.

Considerations other than the economic may result in the need for the freeway either to be depressed below or elevated above the local street system over a considerable distance.

11.3.2 Depressed freeways

Depressed freeways are visually less intrusive than elevated or ground level freeways and result in less noise in the surrounding area. In consequence, the splitting of a previously cohesive community by the provision of the freeway is thus significantly reduced.

The local street network is as disturbed by provision of the depressed freeway, as in the case of a ground-level freeway. Local streets either have to be linked across the freeway by grade separations or served by frontage or service roads. Alternatively, they are converted into culs-de-sac.

A problem typically encountered in the design of depressed freeways is that of disposal of storm water. Storm water drainage for the road reserve by gravity requires the freeway to be above ground level somewhere along its length, which may not always be feasible. Pumping stations would then be required.

Torrential downpours of rain are often accompanied by power outages, possibly resulting in a depressed freeway becoming flooded. It would be prudent to provide pumping stations with auxiliary power supplies.

A depressed freeway would typically have to be at a depth of about 6 to 7 metres below that of the local street network. If entire city blocks are expropriated in the land acquisition process, the road reserve would, in all probability be wide enough to contain normal cut slopes. This would support landscaping of the freeway reserve to the benefit of the environment as a whole. Given the typical width of 40 to 60 metres of a freeway reserve in relation to the size of a city block, it is more likely that acquisition of full city blocks would not be necessary and that some useful application could be found for the alienated portions of these blocks.

A typical freeway cross-section could comprise

- Six 3.6 to 3.7 metre wide lanes for moving traffic, with 3 metre outer and 1.5 metre inner shoulders
- A median island sufficiently wide to allow for the central pier of a structure passing over the freeway with a clear space on either side of it
- A 1 metre wide foreslope and a side drain with a width of 1 metre at the bottom of the trapezoidal section

This is illustrated in Figure 11.2.

In the case of a 40 metre wide reserve, these dimensions would not leave any space to accommodate the back slopes. The earthworks would thus necessitate the use of retain-

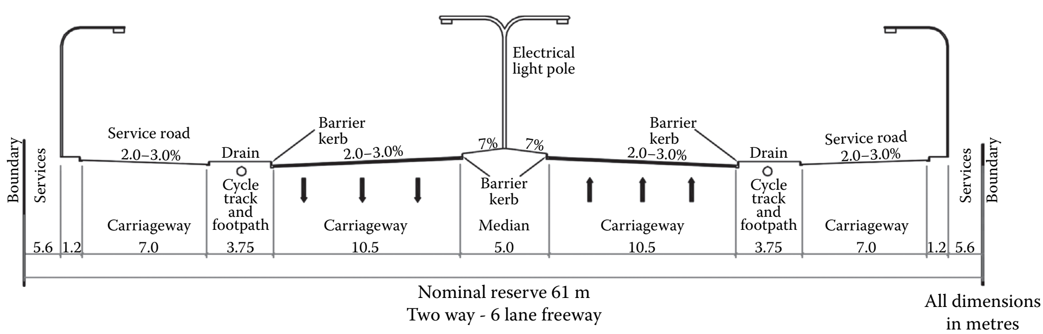


Figure 11.2 Six-lane freeway cross-section.

ing walls, and it is suggested that a 40 metre width is the absolute minimum that may be contemplated for a six-lane freeway.

The actual width of the road reserve in urban areas is typically dictated by issues such as the location of property boundaries relative to the freeway, the need for frontage roads and provision of space for public utilities adjacent to the frontage roads. There shouldn't be any utilities at the level of the freeway. The reserve width could thus vary along the length of the freeway. At some places it may be necessary to provide retaining walls, and, at others, space would permit the use of relatively flat and aesthetically pleasing landscaped side slopes or batters.

If they are required, retaining walls need not necessarily extend to the full depth of the excavation. They are required only to balance the extent of the cut batter to the width of the reserve. For example, a 60 metre wide reserve could accommodate a cross-section comprising six lanes, a 9-metre median central island with 2 metre wide shoulders and 3 metre outer shoulders, with some allowance for slope batters and verges. A narrower reserve would, however, require a batter steeper than 1*V*:6*H* or, alternatively, retention of the 1*V*:6*H* batter and retaining walls to make up the height difference between ground level and the top of the cut slopes at the reserve boundaries.

The question then is where in the cross-section the retaining walls should be located. They could be at the reserve boundary, at the base of the cut slope or at any point in between. From a purely geometric perspective, road safety would dictate that the retaining walls should be as far as possible away from the carriageways, that is, at the reserve boundaries. Furthermore, having the retaining walls located at the base of the cut slope may engender a feeling of claustrophobia, whereas a sense of openness would be strengthened by location of the retaining wall at the reserve boundary. Unfortunately, having a retaining wall located at the top of a cut slope may cause problems of slope stability. A geotechnical investigation should thus be undertaken to determine the measures that would have to be put in place to ensure the integrity of the cut slope.

11.3.3 Elevated freeways

Although the depressed freeway may be, for various reasons, the preferred option, the cost of an extended cut, particularly through rock, could be exorbitantly high. If having the freeway at ground level is not an option, it follows that all that is left is to elevate the freeway above the local street system. The freeway could then be provided either on a continuous fill or embankment or on a viaduct, both of which are visually highly intrusive.

The embankment option could have the undesirable effect of being a wall between two portions of a previously cohesive community. On the other hand, it could provide a screen between two areas that are fundamentally different. A residential area could, for example, be protected from the visual impact of an industrial area. This option, like the ground level freeway, is seriously disruptive of the local street network. It requires frontage or service roads to reinstate the continuity of the street network and the conversion of previously continuous streets into culs-de-sac on either side of the freeway road reserve.

Provided an architectural sensitivity is brought to bear on its design, the viaduct option could be aesthetically pleasing, as illustrated in Figure 11.1. The viaduct also ensures that the local street network can be left virtually undisturbed, as illustrated in Figure 11.3. The surface area below the viaduct could be used for parking, commercial activities or community activities such as children's playgrounds. This area is also attractive to homeless people as a shelter from the elements and thus may require additional policing.

Where space is limited, the viaduct offers the option of a double decker cross-section, as illustrated in Figure 11.4, with one of the carriageways being located above the other. The

Figure 11.3 Viaduct cross-sections. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

Figure 11.4 Double-decker viaduct. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

decks could also be cantilevered off central piers, thus significantly reducing the street level footprint of the viaduct.

A problem normally encountered in the case of the double decker viaduct is the length of the ramps required to provide access to and from the local street system from the upper deck. These ramps usually have to be threaded between the two carriageways, so the location of the supporting piers can also be a major problem. Securing adequate vertical clearance between the decks and the ramps and determination of the location of the points at which minimum clearances occur can also present major difficulties.

11.4 INTERCHANGES

11.4.1 Introduction

It was noted (Collins and Hart, 1935) that, at most intersections, only 10 to 20 per cent of all vehicles entering the intersection intended to turn: 'If means could be adopted to separate the roads at such crossings, about eighty to ninety per cent of the traffic would benefit. Unfortunately, the cost of such flying junctions will usually preclude their adoption, but their value nevertheless cannot be denied'.

In the meantime, over in the United States, New Jersey was developing its cloverleaf interchange.

Turning volumes higher than can be accommodated at signalised intersections are indeed one of the warrants for the provision of interchanges. There are other and more significant warrants for their provision, and these are discussed further in the sections that follow.

Discussion in this section includes

- The principles underlying the design of interchanges
- The location and spacing of interchanges
- Typical and unconventional interchange layouts
- Ramp design

11.4.2 Design principles

The principal difference between interchanges and intersections is that, in the case of interchanges, crossing movements are separated in space, whereas in the latter they are separated in time. At-grade intersections accommodate turning movements within in the widths of the crossing roadways or through the application of turning roadways. Turning movements at interchanges are accommodated on ramps. These ramps replace the low-speed turns through approximately 90° at intersections with high-speed merges and diverges on the major road of the interchange and at-grade intersections on the crossing road if this is not also an interchange.

Grade separations also provide a spatial separation between the two crossing roads but do not make provision for turning movements. They are, in effect, interchanges without ramps and thus do not qualify for consideration as a form of intersection.

The speed of movement through an interchange on the major road has the consequence of drivers not having much time in which to make up their minds as to their desired actions within the interchange area. It follows that the decision-making process has to be simplified as far as possible. At a four-legged interchange the driver has the choice of proceeding through the interchange or turning to the right or to the left. For drivers familiar with the freeway and their normal route along it, this does not constitute a problem. They do not require the support of navigational signs either. An unfamiliar driver, on the other hand, has to spend time reading the directional signs as they come up and then to decide at each one whether to stay on the freeway or leave it and, in the latter case, the direction in which to turn once having left it. In executing the design of an interchange, it is important that the designer takes cognisance of the problems of the unfamiliar driver to avoid the problem of wrong-way driving, specifically wrong-way entry onto the freeway, the consequences of which are likely to include multivehicle crashes and fatalities.

There are numerous decisions that the driver has to take and several tasks that the driver has to execute successfully to achieve the desired end result without being a hazard to other traffic. These include

- Deciding on the lane appropriate to the intended action
- Carrying out the weaving manoeuvre necessary to give effect to the lane change if necessary
- Selecting the appropriate speed for the required manoeuvre
- Accelerating or decelerating to this speed from the speed currently being travelled
- Diverging towards an off-ramp or merging from an on-ramp with the through traffic

In the latter case, while still on the on-ramp the driver has to

- Locate a suitable gap in the through traffic
- Adjust the position of his or her vehicle relative to this gap
- Adjust the speed of his or her vehicle to that of the freeway traffic
- Merge with the through traffic, bearing in mind the gap between the leading vehicle and the merging vehicle; and simultaneously remaining aware of the location and actions of the following vehicle

The level of concentration required is significant. It follows that the design of a safe and effective interchange must be a response to all of the above points. For example, a basic principle of interchange design is to provide drivers with structured binary rather than multiple decisions. The decision facing the driver on the freeway is, 'At this interchange do I go straight or turn to the left or turn to the right?' is replaced by the decision, 'At this interchange, do I leave the freeway or stay on it?' followed some seconds after having left it by the decision regarding the direction in which to turn. The decision and judgment required of the merging driver is significantly more complex, as it contains most of the decisions required of the diverging driver in addition to the further decisions and actions described previously.

To be able to carry out all these actions in safety, the driver must be in a position to understand the operation of the interchange and not be surprised or misled by an unusual design feature. Understanding is based on the human factor of expectancy and this is supported by consistency of design and uniformity in the selection of the components of the interchange.

Part of this expectancy involves the ongoing debate of over versus under. Should the crossing road be taken over the freeway or under it? Frequently, this decision is automatically resolved by considerations of the topography and cost of construction. However, if this is not the case, having the crossing road over the freeway means that the driver on the freeway is made aware while still some distance away from the interchange of its presence and location. Furthermore, particularly on narrow diamond interchanges (also known as tight urban diamond interchanges [TUDI]), the crossing road over the freeway means that off-ramps will usually be on upgrades and on-ramps on downgrades. Where space is at a premium, deceleration to a stop from a speed of 120 km/h at an off-ramp and acceleration from crawl speed at the crossroad ramp terminal to the freeway speed on the on-ramp are significantly eased. This would make it possible to reduce the lengths of the ramps if space limitations makes this necessary. Because of the narrower cross-section of the crossing road and also the lower design speed on it, raising the crossing road over the freeway would involve a lesser quantity of earthworks than the alternative and hence a significant reduction of construction cost. The principal advantage of having the crossing road under the freeway means that abnormally high loads would not have the problem of limited vertical clearance under the freeway structure.

The diverging diamond interchange, discussed later, represents a conflict with driver expectations insofar it require causes drivers accustomed to driving on the right side of the road to suddenly, and for a short distance only, drive on the left. In spite of the research (Bared, 2007) that demonstrated that drivers managed in just about all cases, with the notable exception of one elderly female driver, to negotiate the interchange correctly, some reservations must remain. The nano interchange (Moon et al., 2010), also discussed later, conflicts with driver expectations insofar as it requires both freeways to be double deckers, thus facilitating exits and entrances to be on both sides of each carriageway. Exiting from a freeway from both sides of a carriageway runs counter to driver expectations and would, in any event, probably result in serious levels of turbulence.

It is always possible that, by the exercise of some imagination, a new and unconventional design can be created. It is equally possible that, by adequate and possibly elaborate signs and markings, drivers can be guided through the interchange to reach their desired destination. A major problem that needs to be addressed in the adoption of unconventional designs is the damage that these designs could inflict on driver expectation, particularly when they address a one-off situation in a series of otherwise conventional interchanges. A further question is whether the unconventional design is necessarily any better than what it replaces.

Consistency of design is, because of the issues discussed earlier, important if an interchange is to operate safely and efficiently. Of particular importance is that the design should match driver expectations. For example, drivers have come to expect that on- and off-ramps will be on the right. In addition, they also expect that there will be a high level of turbulence on the outside lane of the freeway and, in consequence, would tend to move out of this lane if intending to move through the interchange. They also expect that leading vehicles will move at fairly constant speeds over considerable distances without the punctuation of sudden sharp changes in travel speed. Headways between vehicles in a platoon thus tend to be shorter on freeways than on other roads. Although this is a partial explanation for the higher capacity on a freeway compared to that on a dual carriageway non-freeway, it has the unfortunate drawback that a sudden sharp change of speed is likely to result in a multi-vehicle pile-up and total chaos for an extended period of time.

Uniformity of signing practice is an important aspect of consistent design and, in the United States, reference should be had to the *Manual of Uniform Traffic Control Devices*. Most other countries have their own signs manuals and, in southern Africa, the Southern African Development Community comprising Angola, Botswana, Democratic Republic of the Congo, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia and Zimbabwe has its *Road Traffic Signs Manual* common to all countries.

A very important principle is that of route continuity. Drivers wishing to remain on the freeway while passing through a town or city should not be required to leave the freeway and join another to continue their journey. This has the effect of forcing the unfamiliar drive into having to take a decision on route selection where none should have been necessary. Examples of route continuity and lack thereof are illustrated in Figure 11.5.

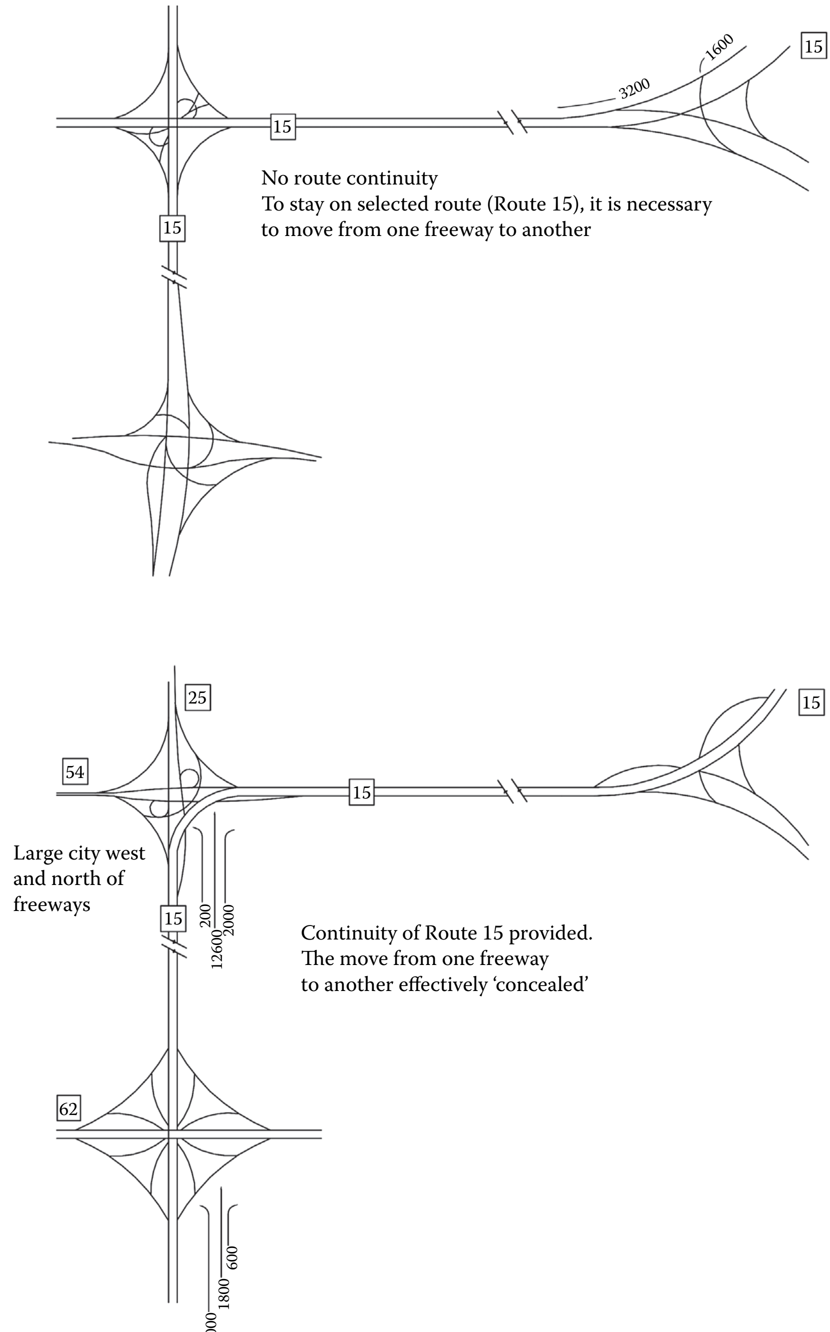


Figure 11.5 Route continuity. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

11.4.3 Interchange warrants

11.4.3.1 Freeways

As previously stated, a freeway is not free from payment or tolling; it is free from interference caused predominantly by entering and exiting vehicles and, to a lesser extent, by vehicles changing lanes for whatever reason. The freeway is, in the spectrum of mobility to accessibility, the highest manifestation of mobility. It is intended for travel in safety at high speeds over relatively long distances. In support of this goal, a total control of access is applied to its operation. In short, if a road is designated as a freeway, the only permissible access to it is via an interchange. This is the most common warrant for the provision of an interchange.

11.4.3.2 Traffic volumes

To refer back to the opening paragraph of this section, the 'flying junction' mentioned by Messrs Collins and Hart was intended to accommodate turning vehicles away from the through traffic to eliminate unacceptable spot congestion at intersections and thus to ensure smoothly flowing streams of traffic on the major road of the intersection. This would become necessary when all the possibilities of the at-grade intersection had been exhausted. These include spatial separation of turning vehicles onto turning roadways or slipways that are clear of the through lanes and temporal separation by means of leading or lagging phases in the signal pattern.

Unconventional intersection layouts, such as the median U-turn design, change the left turn into a right turn onto the crossing street followed by a U-turn across the median. This results in the original left turn being replaced by a straight through movement on the crossing street. As such, it may delay the changeover from a highly sophisticated intersection to an interchange but there is a limit to the volume of traffic that could be accommodated by a U-turn. At some or other stage, upgrading from an intersection to an interchange would probably become almost inevitable.

Heavy traffic volumes have a negative impact on road user costs owing to the length of time that a vehicle is stationary with its engine idling, not to mention the waste of time that could have been more usefully employed elsewhere. Tyre wear is higher during acceleration and deceleration than during travel at a constant speed, as is fuel consumption. Before taking a decision to replace an intersection by an interchange for reasons of congestion, a financial analysis would have to be carried out to establish whether the cost of provision of the interchange is less than the savings it generates. If this proves not to be the case, a policy decision would have to be taken on the price that road users would be prepared to pay for convenience.

11.4.3.3 Safety

It is possible that an at-grade intersection may have an unacceptably high frequency of serious crashes. The accident black spot is not unknown. This could sometimes be caused by a misinterpretation of the information the driver perceives from the road, for example, an optical illusion creating the impression of a straight road when the intersection is actually curving. It could also arise from an information overload. When all the possible options aimed at improving the safety of the intersection have been exhausted, the only remaining alternative is to replace the intersection by an interchange.

11.4.3.4 Topography

It is possible that the site topography could be so adverse that construction of an at-grade intersection would be prohibitively expensive if appropriate design criteria are to be met and an alternative site is not available. For example, there might be a very large height difference between the gradelines of the crossing roads, necessitating excavation through hard rock including substantial daylighting of the intersection to ensure adequate intersection sight distance.

The implication of the use of this warrant is that neither of the crossing roads needs to be a freeway. If traffic is light, a single two-way ramp, also referred to as a jug handle because of its shape, would probably be adequate. This is similar to the quadrant roadway intersection discussed in Chapter 10.

11.4.4 Spacing of interchanges

The spacing of interchanges is based primarily on service to adjacent land uses. On rural freeways passing small communities the provision of a single interchange may be sufficient. Obviously, in this case, spacing would be an irrelevant concept. Larger communities would require more interchanges. Two interchanges, with one on each side of the village, would be adequate for most rural settlements. If the CBD were to grow to the point of its being a destination in its own right, a central interchange may also be necessary. It is not possible to lay down hard and fast rules on the number of interchanges required by any community, and each case would have to be considered on its own merits. The precise location and hence the spacing of interchanges would depend on the needs of the community being served by them but, as a general rule, would be on roads recognised as being major components of the local street system (Burrell et al., 2002).

Rural interchanges are typically spaced at distances of 8 kilometres or more apart. This distance is usually measured from centreline to centreline of the crossing roads. For planning purposes, centreline to centreline spacing is usually sufficiently accurate.

At spacing appropriate to the urban environment, reference to a centreline to centreline spacing is too coarse to be practical. Weaving, discussed further in the text that follows, takes place between interchanges, and the available distance for weaving is a function of the layout of successive interchanges. For a common centreline-to-centreline spacing, the weaving length between two diamond interchanges is significantly different from that between a parclo A followed by a parclo B.

It is suggested that centreline to centreline spacing be replaced by weaving distance or white line break point (WLBP) distance.

Weaving distance as defined in the HCM is measured from the point where the distance between the left edge of the on-ramp and the edge of the freeway is 0.5 metre to the point where the distance from the left edge of the off-ramp is 3.7 metres from the edge of the freeway. In consequence, the weaving length becomes a function of the taper rates applied to the on- and off-ramps. The WLBP distance, on the other hand, is not subject to extraneous variables and is totally unambiguous. It is thus the recommended option in the measurement of the spacing between successive interchanges and is used throughout in this Handbook. The comparison between the WLBP distance and weaving distance as defined by the HCM is shown in Figure 11.6.

From considerations of geometry as illustrated in Figure 11.6, it follows that the overall length of a weaving section could be aggregated as shown in Table 11.1 for a 3.7 metre wide auxiliary lane, with a 1:50 entrance taper and a 1:15 exit taper, these being fairly typical dimensions of a weaving section.

There are three criteria for the spacing of interchanges that can be considered:

- The distance required for adequate signage of the freeway
- The achieving of acceptable traffic operation on the freeway
- Minimisation of turbulence on the freeway

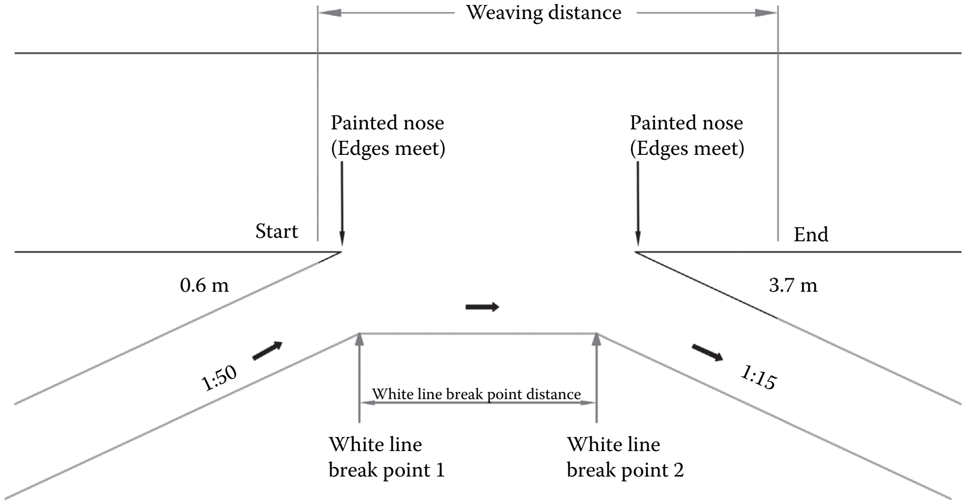


Figure 11.6 White line break point distance. (From Wolhuter KM and Garner D. *Spacing of interchanges in Gauteng*. Department of Public Transport, Roads and Works, Johannesburg, 2003.)

Table 11.1 Comparison between white line break point (WLBP) and weaving length

Component[a]	*Without auxiliary lane (m)*	*With auxiliary lane (m)*
Start to painted nose (edges meet)	30	30
Painted nose to WLBP1	185	15
WLBP distance	X	X
WLBP2 to painted nose (edges meet)	56	5
Painted nose to end	55	55
Total weaving length	*X* + 326	*X* + 105

Source: Wolhuter KM and Garner D. *Spacing of interchanges in Gauteng*. Department of Public Transport, Roads and Works, Johannesburg, 2003.

[a] See Figure 5.2.

11.4.4.1 Adequate distance for signage

The signposting requirements on freeways should be such that drivers are adequately informed of the presence of interchanges and sufficient time allowed them to manoeuvre safely and unhurriedly to the position required for a safe exit from the freeway. Southern African practice provides for three countdown or braking markers located at 100-metre intervals in advance of the WLBP, and these may be a useful aid to drivers, especially under conditions of poor visibility.

The full sign sequence between interchanges commences with post interchange signs. These signs principally serve a confirmatory function so that the driver knows that, having passed through the interchange, he or she is still on the correct road to the desired destination, or alternatively of course, that the required exit has been missed. If space allows, anything up to three signs can be used – a confirmatory or route sign followed by a speed restriction reminder and a distance sign. The route sign is mounted 500 feet (150 m) downstream of the on-ramp WLBP and the following two signs at 1000 foot (300 m) intervals thereafter. According to the *Manual for Uniform Traffic Control Devices*, advance guide signs, as shown in Figure E2-22 of that manual, should be located 0.5 mile (0.8 km)

Table 11.2 Suggested interchange spacing in terms of signage requirements

Country	*Qualification*	*Minimum distance between ramp noses*
United States	Systems to access interchange	600 m (absolute minimum)
Germany	Access to access interchange	480 m (absolute minimum)
	System interchange	2700 m (desirable) 600 m (absolute minimum)
	Access interchange (high traffic load)	2200 m (desirable) 600 m (absolute minimum)
	Access interchange (low traffic load)	1700 m (desirable) 600 m (absolute minimum)
France	Except for very exceptional cases	1000 m (absolute minimum)
United Kingdom	120 km/h design speed	450 m (absolute minimum)

Source: Wolhuter KM and Garner D. *Spacing of interchanges in Gauteng*. Department of Public Transport, Roads and Works, Johannesburg, 2003.

and 1 mile (1.6 km) in advance of the off-ramp WLBP. It follows that providing the full sign sequence between successive interchanges would require a WLBP distance of 2300 metres between them plus a gap between the last post interchange sign and the first of the advance guide signs. In terms of the 6-second axiom (World Road Association, 2008) an additional gap of 6 seconds travel time should be provided between the on- and off-ramp sign sequences. At a normal operating speed of 120 km/h, this requires a further 200 metres and, hence, a total distance of 2500 metres between WLBPs.

This distance can be reduced if space does not permit use of the full sign sequence. The post interchange signs can be reduced to providing only the route sign 150 metres beyond the on-ramp WLBP and leaving out the first advance guide sign located 1600 metres upstream of the off-ramp WLBP. This has the effect of reducing the interchange spacing from a rural 2500 metres to an urban 1150 metres. Interchange spacings in terms of WLBP distances are suggested for various interchange sequences in Table 11.2.

The meanings of the terms 'access interchange' and 'systems interchange' used in this table are explained in Section 11.4.8. Briefly, an access interchange provides access to and from local areas to freeways whereas a systems interchange is the link between two freeways.

11.4.4.2 Required distance in terms of acceptable traffic operations

In terms of the design domain concept, the distances suggested in Table 11.2 constitute desirable minima. The distances required for operational efficiency may be considered to be absolute minima. It has to be noted that there has been a major change in the description of weaving configurations between the 2000 and the 2010 versions of the HCM as discussed further in Section 11.4.6.

Operational efficiency is described in terms of LOS, which are, as is the case of the other segments of a freeway, defined by density. Density is measured in units of passenger cars per lane per kilometre and the boundaries of the various levels are given in Table 11.3.

The process of calculation of the density achieved commences with the assumption that a new interchange on the freeway would be accommodated by two merge–diverge operations, one upstream of the proposed interchange and the other downstream. The calculation methodology is described in the HCM 2010. As shown in Table 11.3, breakdown, that is, the boundary between LOS E and LOS F, is likely to be at a density of 27 passenger cars per per lane per kilometre. It would not be wise to accept breakdown as a design limitation, as this constitutes a totally unacceptable level of congestion. As a general rule, if density is at

Table 11.3 Boundary densities for various levels of service

Level of service	*Density (passenger cars per lane per kilometre)*
A	6
B	13
C	18
D	22
E	27

Source: Transportation Research Board. *Highway capacity manual 2010*. Washington, DC, 2010.

LOS D or better, the merge–diverge is acceptable. If this should prove not to be the case, it would be necessary to adopt one or another of the two weaving operations.

As the ramp weave requires all weaving vehicles to undertake a lane change, the ramp weave cannot accommodate as many weaving vehicles as the major weave but it could accommodate higher through flows on the freeway. It would thus be suitable as an access to small communities. The major weave not only accommodates higher weaving volumes than the ramp weave, but it is also the location for the addition of a further basic lane on the freeway or the dropping of a basic lane. Lane balance, discussed in Section 11.4.5, thus requires that the upstream ramp should have more than one lane in the case of addition of a basic lane drop or the downstream ramp have more than one lane to balance a lane drop.

A problem that all road authorities encounter on a regular basis is the developer who considers access to a freeway absolutely essential for the profitability of his or her development and is prepared to leave no stone unturned to achieve this. Particularly in urban areas where interchanges are already closely spaced, the impact that the demanded additional interchange could have on flow conditions on the freeway could be catastrophic. Unfortunately, this would be a matter of indifference to the developer. With the growing litigiousness of society at large, a rejection of the application for access to the freeway would almost inevitably be followed by the developer seeking a court order against the road authority. The road authority would have to display that it had 'applied its mind' in rejecting the application. A calculation of the LOS without and with the requested interchange should go a long way towards setting the legal mind at ease regarding the validity of the road authority's argument.

11.4.4.3 Turbulence

According to Roess and Ulerio (1993), the area of turbulence downstream of an on-ramp and upstream of an off-ramp is of the order of 450 metres measured from the painted nose so that weaving occurs, by definition, in a turbulent zone. Turbulence is characterised by large speed differentials and fluctuations in speed. It follows that conditions within a weaving section are inherently unstable so that, even where the calculations suggest that a weaving section will operate satisfactorily, the possibility of disturbances in the traffic flow cannot be excluded. These disturbances could result in a local bottleneck developing and may even lead to a heightened likelihood of crashes. Consequently, a spacing of greater than 900 metres between successive noses is to be preferred.

This translates to a WLBP distance of 880 metres for ramp weaves as shown in Table 11.2. In the case of major weaves, four possible combinations of auxiliary lane configuration and ramps arise: An upstream auxiliary lane is one that terminates at a single-lane off-ramp or, if lane balance is to be achieved, at a two-lane off-ramp. In these cases, the WLBP distances that are equivalent to a distance of 900 metres between painted noses are 695 metres and 645 metres respectively. Similarly, a downstream auxiliary lane could commence at either a one- or a two-lane on-ramp with equivalent WLBP distances of 825 metres or 655 metres respectively. It should be clear that the nomenclature of 'upstream' and 'downstream' auxiliary lanes is relative to the ramp of interest.

11.4.5 Basic lanes and lane balance

Two types of lane are of interest: basic lanes and auxiliary lanes. As in the case of intersections, basic lanes are those that continue en route past several interchanges. Typically, on the outskirts of a town or city, a freeway may have two lanes in each direction. As the freeway gets closer to the CBD, traffic volumes increase and, to maintain an acceptable LOS, it is necessary to increase the basic lanes to three or four or even more. Once past the CBD, traffic volumes decline in the outbound direction and it becomes possible to reduce the number of basic lanes and still maintain an acceptable LOS. Auxiliary lanes, as discussed previously, are those that are added to the cross-section to serve a specific need. Once the need has fallen away, the auxiliary lane is dropped.

At off-ramps, lane balance is intended to give drivers a choice of leaving or staying on the freeways without having to change lanes in order to stay on the freeway. Without lane balance, a driver in the outside lane is inexorably sucked into a local area where he or she did not necessarily wish to be. At on-ramps, choice is irrelevant but, if the number of basic lanes is not to increase at every interchange, drivers on the on-ramp are forced to merge with the freeway traffic. Lane balance is illustrated in Figure 11.7.

11.4.6 HCM changes in the classification of weaves

The merge–diverge was defined in HCM 2000 by the fact that the distance between an on-ramp and the following off-ramp was greater than 750 metres. Weaving was between ramps and the freeway and a distinction was drawn between types A, B and C weaves on the basis of the number of weaving movements required and the distribution of these between the various traffic streams involved.

In HCM 2010, the distance between the successive ramps of a merge–diverge is considered irrelevant and the principal criterion is the existence or otherwise of an auxiliary lane. Weaves are classified as either ramp weaves or major weaves and also one- and two-sided weaves. A one-sided ramp weaving segment is between a single on-ramp lane followed by a single lane off-ramp *and connected by means of an auxiliary lane*. Every weaving vehicle must make one lane change and the lane changing turbulence occurs between the auxiliary lane and the outermost through lane of the freeway. This is the type A weave of HCM 2000. A major weave has one or other of the ramps bounding the weaving section having two lanes. This is the type B weave of HCM 2000. In the case of a two-lane off ramp, the on-ramp to freeway movement requires one lane change whereas the freeway to off-ramp movement can be made without a lane change. In the case of a two-lane on-ramp, it is the freeway to ramp movement that has to make the lane change to remain in the outer lane. High volumes of merging or diverging traffic may require two lanes to accommodate the ramp traffic. The type C weave of HCM 2000, which required one of the movements to make more than one lane change, has been abandoned although, in terms of this definition, it could be analogous to the two-sided weave.

The various forms of weaves are illustrated in Figure 11.8.

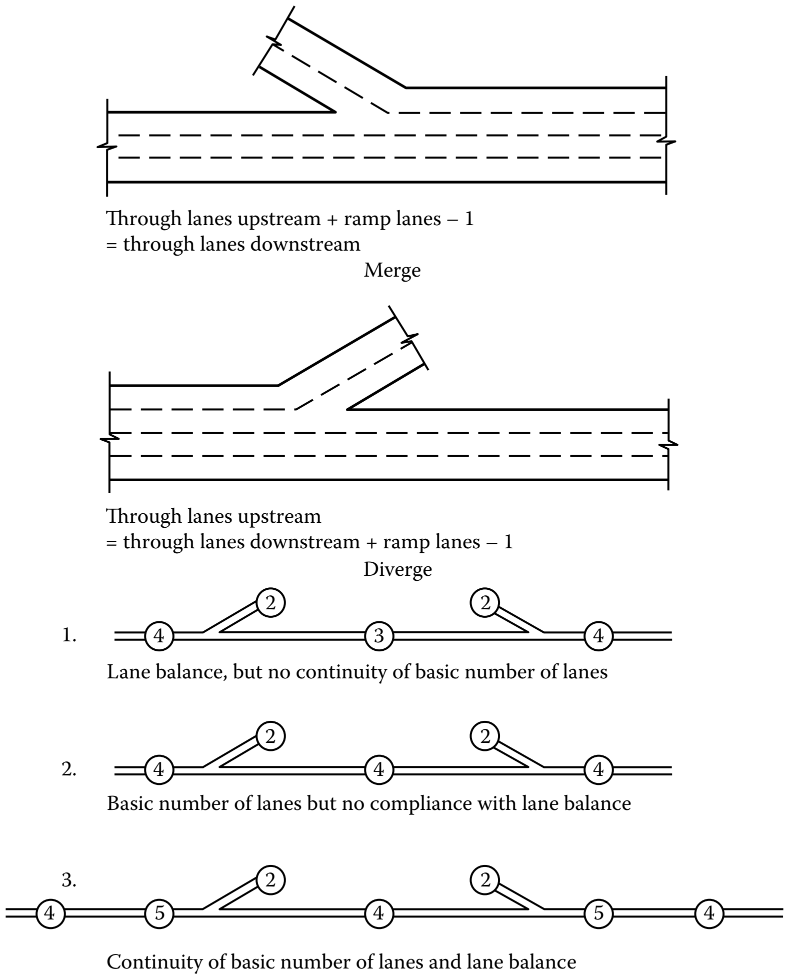


Figure 11.7 Lane balance.

11.4.7 Selection of interchange type

Designers, particularly those without much experience of interchange design, are sometimes inclined to take a decision as the type of interchange to be designed as the first step in the design process. They could say, 'We have already designed three diamond interchanges on this freeway. How about a parclo for a change?' This is a contradiction of the whole philosophy of design because, like the cross-section, an interchange is not the whole and indivisible entity that developers and their town planners often imagine it to be.

An interchange comprises many elements, each with their own functions and dimensions. Design is principally the process of selecting the elements appropriate to the operating conditions and topography prevailing at a particular site, assembling them in sequences corresponding to the needs of each movement through the interchange and, finally, sizing these elements in response to operational requirements such as turning volumes and design speed. It will usually be found that, having gone through this process, a previously well-known typical layout, such as one or other of the various forms of the diamond interchange, will emerge.

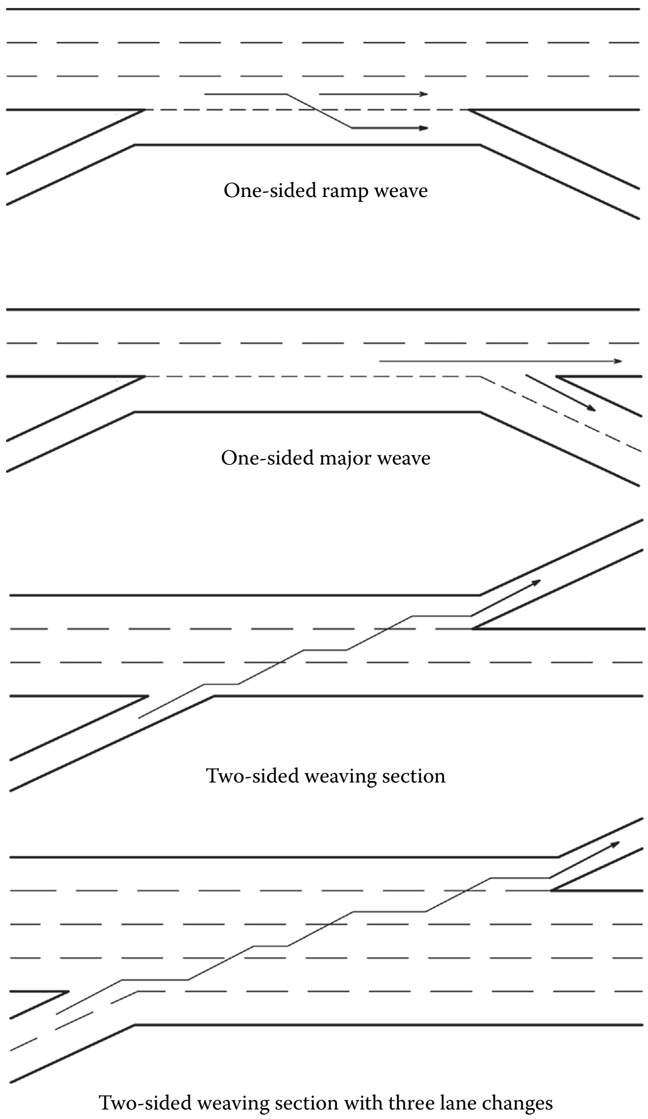


Figure 11.8 Types of weave.

As an alternative to initiating the design without any preconceptions regarding the required type of interchange, an interchange type study should be carried out (Illinois Department of Transportation, 2010). This would serve as a guide to selection of a type of interchange suitable to a particular site and would be helpful in the process of detailed design. Information typically acquired during the type study includes

- Operational issues such as
  - Through and turning volumes of traffic
  - Traffic composition, that is, the numbers of the various vehicle types present in the traffic streams
  - Weaving analyses
  - Capacity analyses of entrance and exit terminals

- Planning issues such as
  - Adjacent land use
  - Existing cultural developments
  - Budgetary constraints
  - The availability of power for signalisation
- Geometric considerations such as
  - Route continuity
  - Lane balance
  - Topography
  - Current mainline and crossing road profiles
  - Preliminary signing plans

Study of this information would lead to informed conclusions being drawn regarding the most suitable top of interchange to be applied at the site of interest.

11.4.8 Systems interchanges

Particularly in urban areas, freeways do not occur in isolation. There is a network or system of freeways. Any system of freeways comprises links and nodes, with the links being the basic freeway segments and the nodes the individual interchanges. These interchanges are thus referred to as systems or freeway-to-freeway interchanges, with the former being the customary nomenclature.

With freeways being designed for high-speed travel, it follows that the connections between them should also be designed to cater to high speeds as far as is practicable. The terminals at both ends of the ramps should thus be designed to cater to merging and diverging manoeuvres, and the radii between the ramp terminals should also be generously sized.

To discuss briefly a matter of nomenclature: In conventional layouts, the directional turn to the left is replaced by the semidirect turn whereby vehicles exit from the right-hand side of the freeway, cross over the opposing freeway and then the initial freeway and come back to earth to join the opposing freeway once again from the right. The semidirect ramp is illustrated in Figure 11.9, with other basic ramp forms being the outer connection (for turns to the right) and the loop (for indirect turns to the left). It is pointed out that the appellation 'semi' has largely fallen away in practice so that, when reference is made to a directional ramp, it is often the semidirectional movement that is intended.

In the mid-1920s the New Jersey State Highway Department developed the first systems interchange. It provided off-ramps with exits exclusively from the outside or right hand lane in the form of diverges. Ramps turning to the right took the form of direct turns and those turning to the left were indirect or loop ramps. Because of its resulting overall shape the interchange became known as the cloverleaf interchange. Figure 11.10 illustrates that the choice of name was not entirely fanciful.

This particular design shows an interesting departure from the norm. The outer connectors have links not only between the two freeways but also to the crossing road in the foreground of the photograph. In effect, this layout serves not only as a systems interchange but also, partially, as an access interchange.

It was soon realised that the cloverleaf had a serious drawback in that vehicles leaving the freeway had to weave across the paths of vehicles wishing to enter it and that this manoeuvre had to take place over a very short distance. If the freeway of interest passed under the crossroad, there would be a sudden and short change from light to dark and back again, making it difficult for drivers to see the opposing flow of vehicles. Furthermore, this weave took place on the freeway itself and created severe turbulence affecting the smooth flow of

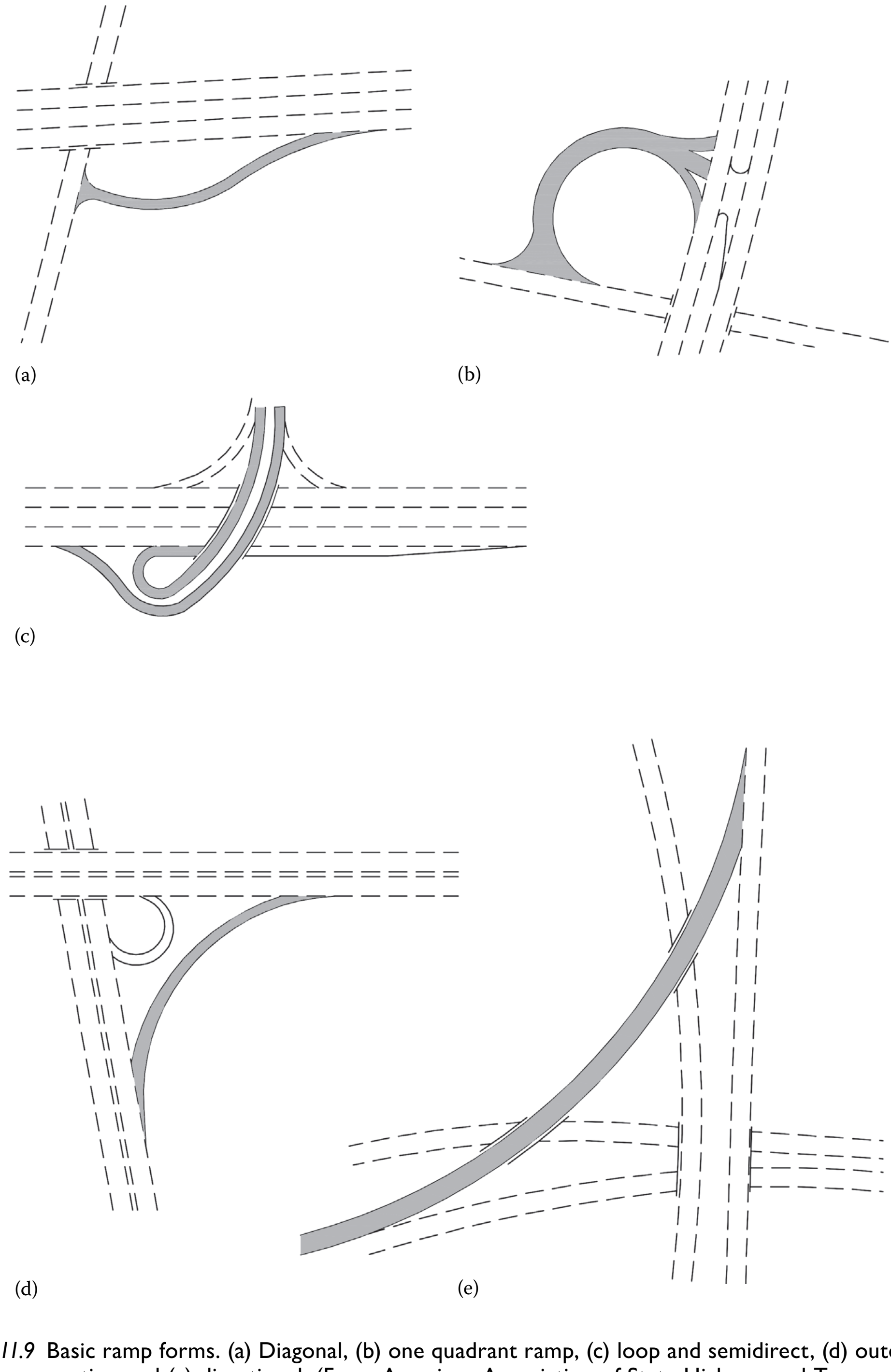


Figure 11.9 Basic ramp forms. (a) Diagonal, (b) one quadrant ramp, (c) loop and semidirect, (d) outer connection and (e) directional. (From American Association of State Highway and Transportation Officials [AASHTO]. *A policy on the geometric design of highways and streets*. Washington, DC, 2011a.)

Figure 11.10 The cloverleaf interchange. (From Michigan Department of Transportation, 2004.)

through traffic. The solution to the problem was to move the weaving manoeuvre off the main carriageway and onto an adjacent lane, the collector–distributor (CD) road. If the crossing road was also a freeway, it also acquired CD roads.

The directional movement of right-turning vehicles allows for high-speed movements, as is to be desired on freeways. As always, however, the left turn is problematic. If this were also to be directional, vehicles would have to exit from the inside or fast lane and merge back into the fast lane of the crossing freeway. Apart from this conflict with driver expectations, this would result in the mixing of a fast through flow of vehicles with a slow flow of turning vehicles.

The designer would also be confronted with the problem of extracting vehicles from the median and developing a sufficient height difference to be able to take the flow over the opposing freeway, which is already one level up from the initial freeway. The flow would also have to clear the opposing carriageway and this height difference would have to be developed within the compass of the median. The problem would be repeated, but in reverse, on the crossing freeway. In consequence, the purely directional left turn was abandoned although it has now reemerged in the nano interchange (Moon et al., 2010) discussed in Section 11.4.11

Systems interchanges can be either three-legged or four-legged. A three-legged interchange is the point at which one freeway terminates at its connection with another. It can have a directional layout or have one directional ramp and a loop ramp serving the freeway carriageway farthest away from the terminating freeway. Because of its overall layout, the latter is referred to as a trumpet interchange. If the loop is in advance of the structure, reference is to a trumpet-A layout and, if beyond the structure, as a trumpet-B interchange (A for *A*dvance and B for *B*eyond). Typical three-legged interchanges are illustrated in Figure 11.11.

Four-legged systems interchanges occur whenever two freeways cross each other. Ideally, these interchanges should be fully directional, both to accommodate all movements at relatively high speeds and also because directional ramps occupy a smaller footprint than loop ramps. As they involve four-level structures, they have the drawback of being the most expensive of the four-legged interchange layouts. In consequence, one or more of the turning movements

(a)

(b)

Figure 11.11 Three-legged systems interchanges. (a) Trumpet A interchange and (b) directional interchange.

Figure 11.12 A typical four-legged directional interchange.

are often located on loop ramps. In spite of this, reference is still made to directional interchanges. A typical layout of a four-legged directional interchange is shown in Figure 11.12.

11.4.9 Access interchanges

In American practice, these interchanges are referred to as service interchanges. The term 'access interchange' is, however, to be preferred, with the term 'Service interchange' reserved for interchanges that are intended to provide service to road users on the freeway and who do not have any intention of leaving it. Services provided include rest areas, filling stations, restaurants, overnight accommodation, truck stops, and so forth.

As can be gathered from their collective name, access interchanges are intended to provide access to and from a freeway and serve the land use on one or both sides of the freeway. There are three basic layouts that can be used as access interchanges. These are

- The diamond interchange
- The partial cloverleaf or parclo
- The trumpet interchange

Although it can be used as an access interchange, the trumpet has long since been in disfavour as an access interchange. This is because urban growth causes what was previously an urban bypass to be swallowed up by the proliferation of suburbs beyond it. The trumpet interchange must then be converted from a three-legged to a four-legged interchange. The cost of this conversion is not trivial and the preferred option is to build the four-legged interchange in the first instance. The crossing road as left as a stub beyond the interchange area until required to be further developed.

11.4.9.1 The diamond interchange

There are numerous variations of the diamond interchange, including the

- Narrow or tight urban diamond interchange (TUDI)
- Wide diamond
- Split diamond
- Roundabout
- Single-point urban interchange (SPUI)
- The diverging diamond

The narrow layout is illustrated in Figure 11.13, with the split diamond, the roundabout and the SPUI illustrated in Figures 11.14, 11.15 and 11.17 respectively. The diverging diamond is discussed in Section 11.4.11.

The narrow diamond, often referred to as a tight urban diamond interchange (TUDI), is normally employed in urban areas because it is possible to fit it into the road reserve of the freeway to which it is granting access with, if necessary, only a modest increase in the width of the reserve. A weakness of this layout is that, if the crossing road is over the freeway, the sight distance afforded drivers on the off-ramps is possibly restricted by the crest curve over the freeway and also by the bridge balustrade to their left. The crossing road under the freeway would, in all probability, have a sag curve in cut. This could cause a drainage problem. Operational efficiency of the narrow diamond is dependent on synchronisation of the traffic signals at the crossing road ramp terminals.

The difference between the narrow and the wide diamonds is that the ramps in the latter case join the crossing road at ground level at points remote from the freeway so that the sight distance problem falls away. Obviously, the road reserve width has to be considerably increased to accommodate the wide diamond ramps. The wide diamond is not a layout purely in its own right but is actually the first stage of the development of a cloverleaf interchange. This would come into play if it became necessary to upgrade the crossing road to a dual carriageway urban

Figure 11.13 The typical diamond interchange in a rural area. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics*. Transportation System Planning, Boise, 2002.)

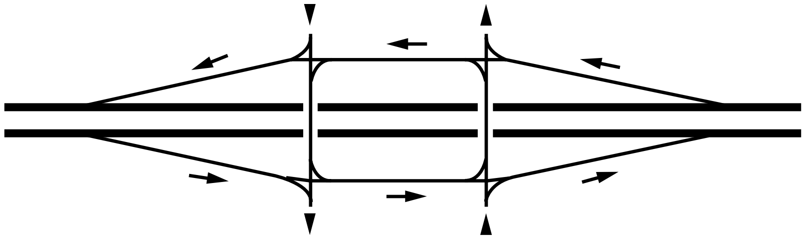


Figure 11.14 The split diamond. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics*. Transportation System Planning, Boise, 2002.)

Figure 11.15 The roundabout diamond.

arterial cross-section where the left turn would be accommodated by a loop ramp. The location of the diamond ramps, which would become outer connectors in the upgraded layout, would be determined by design of the full cloverleaf in the first instance. The upgrade would typically take place several years after completion of the construction of the wide diamond. Experience has shown that standards often change during the intervening years so that the radii originally selected for the loop ramps may have become substandard. However, if significant additional expenditure is to be avoided, it will be necessary to accept the sub-minimum radii, thus invoking the processes of design exceptions, variances and waivers discussed in Chapter 4.

The split diamond illustrated in Figure 11.14 is an accommodation of the crossing road, which is usually a one-way urban pair. Assuming that the one-way streets are correctly located relative to each other, it requires one-way links between the two one-way streets to accommodate left-turning traffic.

It is possible to replace the crossing road ramp terminals of a diamond interchange with roundabouts. Two forms of roundabout diamonds can be considered. In the first case, the diamond interchange is sometimes referred to as a 'dumbbell' interchange. The dumbbell has the effect of changing the operational rules in force at the intersections that constitute the crossing road ramp terminals. Depending on the turning and through volumes at these terminals, the dumbbell can offer operational advantages, for example, by providing precedence to left-turning traffic off the freeway over through traffic on the crossing road.

The roundabout layout can also take the form of a three-level layout with the roundabout at the level between the two crossing roads. This layout, sometimes referred to as the ‘island in the sky’, is frequently encountered in the United Kingdom, which has a long and successful history of the application of roundabouts. Although, in principle, it is a variation of the diamond interchange which is primarily intended for access, it thus also finds application as a systems interchange.

The circular central island of a roundabout can be changed to a teardrop configuration as shown in Figure 11.16 to reduce the possibility of wrong-way driving resulting in a vehicle entering the freeway at an off-ramp.

In a single-point urban interchange (SPUI) all four ramps meet at a common point, usually at the intersection of the freeway and crossing road centrelines. The appellation urban interchange applies simply because the spatial constraints of the urban environment may require a reduction in footprint, which does not normally apply in the rural environment. Furthermore, signalisation is not customarily applied in the rural environment so that being confronted by an unexpected signal may result in an unanticipated reaction by a driver. There is, of course, also a possibility that, in the rural area, power would not be available to drive a signalised system although this problem could be overcome by solar power cells.

Replacing two sets of traffic signals with one reduces the complexity of phasing of the signalisation control and results in a more efficient throughput of traffic on the crossing road (Figure 11.17).

The three-point diamond refers to the geometry of the crossing road and applies to the situation of a skew crossing (Figure 11.18).

Of the three variations of the basic diamond, the SPUI has less conflict points than the conventional diamond interchange, with the diverging diamond having the least number of conflict

Figure 11.16 Teardrop crossing road ramp terminal. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics.* Transportation System Planning, Boise, 2002.)

Figure 11.17 The single-point interchange. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics.* Transportation System Planning, Boise, 2002.)

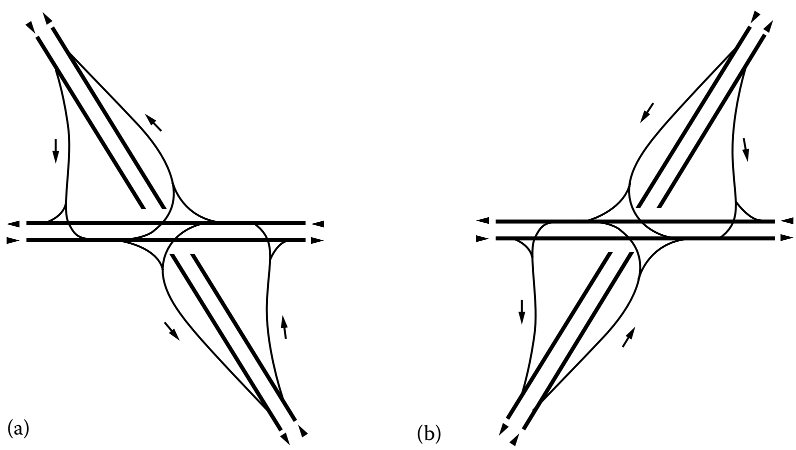


Figure 11.18 The three-point diamond. (a) Angle of skew less than 90° and (b) angle of skew greater than 90°. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics.* Transportation System Planning, Boise, 2002.)

points. This is illustrated in Table 11.4 (Spath, 2008). It is reasonable to assume that the safety of the various layouts and their capacity are a function of the number of conflict points present.

11.4.9.2 The parclo interchange

The parclo interchange was developed by the Ontario Ministry of Transportation and has been well received not only in North America but also worldwide. 'Parclo' is a contraction

Table 11.4 Comparison of conflict points in various diamond interchange layouts

	Number of conflict points		
Manoeuvre	*Conventional*	*SPUI*	*Diverging diamond*
Crossing	10	8	2
Merging	8	8	6
Diverging	8	8	6

Source: Spath SB. *A comparative analysis of diverging diamond interchanges*, 2008.

of the term partial cloverleaf and arises from these interchanges being a combination of loop ramps and outer connections, albeit not a complete set addressing all possible turning movements. An example of the parclo is offered in Figure 11.19.

There are three basic parclo configurations: the parclo A, the parclo B and the parclo AB. As in the case of the trumpet interchange, the letters A and B refer to the loop ramp being either in *A*dvance of or *B*eyond the structure over the freeway. A further distinction is drawn as being a parclo A2 or A4 or B2 or B4. The numbers refer to the number of outer connections provided so that a parclo A2 has the loops in advance of the structure and is provided with two outer connections, both of which are in the same quadrants as the loops. A parclo A4 has outer connections in all four quadrants, as does the parclo B4.

The various forms of parclo are illustrated in Figure 11.20.

The parclo has the effect of converting the troublesome left turn into a right turn with the further possibility of this right turn having a continuous merging flow as opposed to being stop or signal controlled. The main advantage of the parclo 2 layouts is that they leave the

Figure 11.19 The parclo interchange.

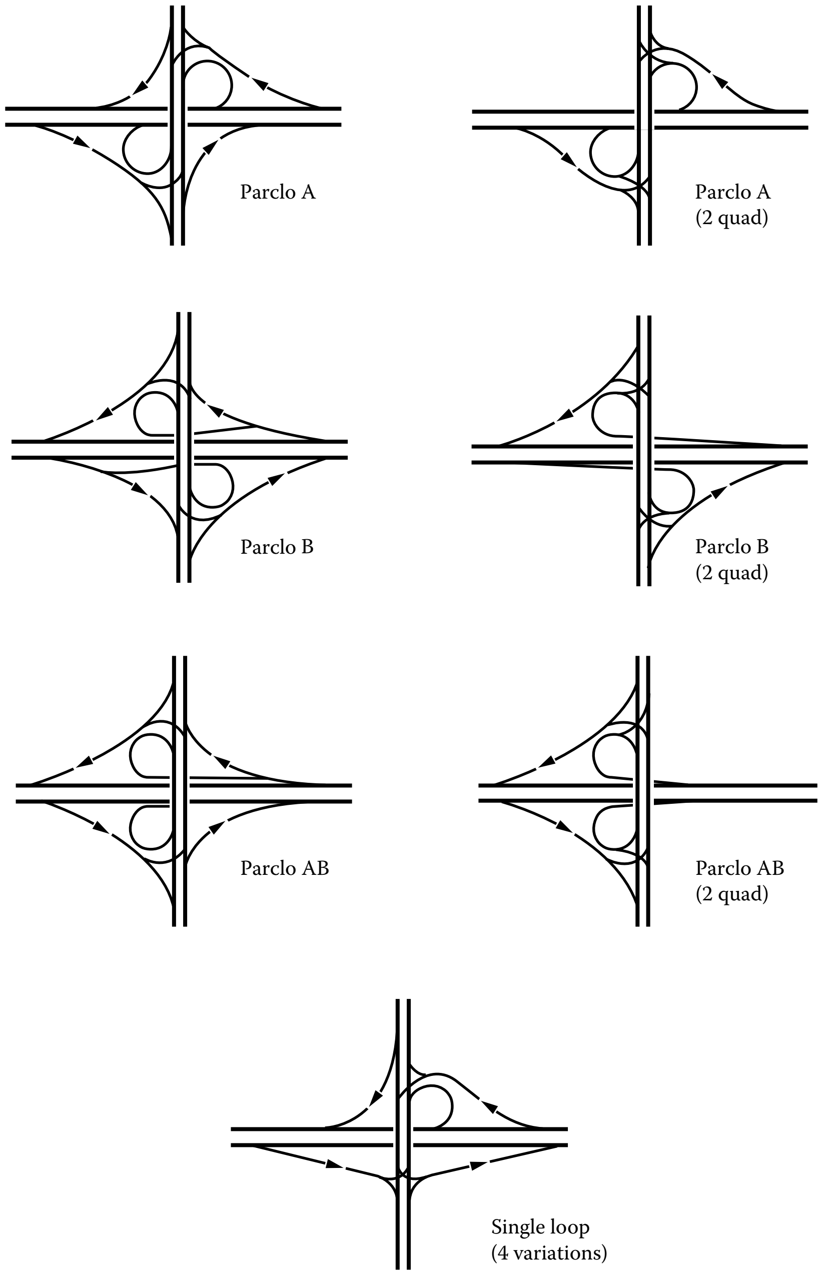


Figure 11.20 The basic parclo interchanges. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics*. Transportation System Planning, Boise, 2002.)

quadrants not containing the loops unoccupied, and this is particularly advantageous when opposing land uses such as large buildings have to be accommodated.

The parclo AB has both loops on the same side of the crossing road. This is often applied in a transportation corridor having both rail and road links. These corridors are usually not wide enough to accommodate loops between the road and the railway without the loops needing further structures carrying them over the railway line. This also has the consequence of the parclo not having outer connections in all four quadrants.

The parclo B has a significant weakness insofar it could comprise a 40 km/h loop being located between two tangents of considerable length, with vehicles on the approach to the loop travelling at a speed of 120 km/h or more. Drivers who have been on the freeway for any length of time will have become desensitised to their speed of travel, with a distinct possibility of misjudging their speed on entry to the loop. Run-off-the-road (ROR) crashes become a distinct possibility.

11.4.10 Service interchanges

Referring again to the distinction drawn between access and service interchanges, service interchanges provide services to passing motorists, specifically the provision of food, fuel or rest facilities. They have on- and off-ramps similar in geometric standard to those on other interchanges but do not necessarily include a structure over the freeway. Normally, the service facilities are duplicated on either side of the freeway but these may also be provided on a structure spanning the freeway.

Apart from the benefit to the travelling public, a hidden advantage is that service interchanges can readily be converted into access interchanges should further development of adjacent land uses makes this necessary. In the case of the service interchange including a structure over the freeway, this benefit stems from the nature of the contract between the transport authority and the entrepreneur who has been granted the use of the air space above the freeway. Permission for this usage is withdrawn when the need for the crossing road manifests itself. Payment for it usually takes the form of the developer either providing or contributing to the cost of a structure that can be converted into a road bridge capable of carrying the design vehicle with a cross-section to match that required to be provided for the new crossing road.

11.4.11 Unconventional interchanges

Study of the literature will reveal that many unconventional interchange layouts have been proposed in the past and many more will, no doubt, be proposed in the future. Only a limited selection is discussed here. These interchanges are normally innovative attempts to resolve the problem of the left turn. They are unconventional in that they tend to take liberties with driver expectations with the consequential possibility of high crash rates.

Reference has already been made to the diverging diamond interchange. This is illustrated in Figure 11.21.

The diverging diamond has all crossing road traffic moving to the other side of the road so motorists drive on the left for a short distance. This results in the left turn onto the freeway not being a problem and can, in fact, flow continuously. The right turn onto the freeway is located upstream of the crossover and can thus also flow continuously. The departure from the conventional situation lies in traffic being required to proceed on the left, which is the norm only in the United Kingdom and some of its previous colonies. Various states have built diverging diamond interchanges and, contrary to expectations, they have been found to be successful. As shown in Table 11.4, these interchanges have less conflict points than any of the other diamond interchange layouts and it can thus be inferred that they would have lower accident rates and higher capacities than the others. It has also been found that drivers do not experience any difficulty in navigating their way around what is an unusual layout.

Figure 11.21 The diverging diamond interchange. (From Bared JA. *Drivers' evaluation of the diverging diamond interchange*. Techbrief FHWA-HRT-07-048, Federal Highway Administration, Washington, DC, 2007.)

11.4.11.1 The Michigan diamond interchange

The Michigan diamond interchange is also known as the median U-turn (MUT) interchange. Vehicles turning to the left onto the freeway are required to make a right turn onto a service road which ultimately becomes an on-ramp followed by two left turns to join the service road which becomes the on-ramp in the opposite direction. This is broadly similar in operation to the concept of drivers travelling beyond the road into which they wish to make a left turn and then executing successive right turns to achieve the desired change of direction followed by proceeding through the intersection onto the leg which would have been the original target of the left turn. This operation does, of course, require a significantly longer travel time than is the case with the MUT so that the analogy could be pushed too far.

11.4.11.2 The double trumpet

This interchange arises as a retrofit of a cloverleaf interchange (Hummer, 2010). As can be seen in Figure 11.22, it has the weakness of having two closely spaced exit ramps in quick succession followed by two closely spaced successive entrance ramps. The signposting required to ensure that drivers do not take the wrong exit could be problematical given the length of time drivers would have in which to decide which of the off ramps to take. The

Figure 11.22 The double trumpet interchange. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics*. Transportation System Planning, Boise, 2002.)

problem could, of course, be resolved by the use of C–D roads although weaving, which is the normal application of C–D roads, is not an issue. In this case, the C–D road would purely be a device to give drivers more decision time.

It is not known at the time of writing (November 2013) whether any interchanges having this layout have been built so far.

11.4.11.3 The single quadrant or jug handle interchange

This interchange can be used to resolve a height difference between two roads at the logical point of intersection between them. Its name derives from the appearance of the interchange when it comprises two three-legged intersections, one on each of the crossing roads, linked by a two-way ramp. This application is shown in Figure 11.23, which illustrates that this interchange need not be applied only to the crossing of a freeway by another road.

It is pointed out that this layout is well known and could perhaps be excluded from the list of unconventional layouts. It is, however, not very frequently encountered and it is for this reason that it has been included in this category of layouts.

11.4.11.4 The nano interchange

This interchange was introduced at the 4th International Symposium of Highway Geometric Design held in Valencia in June 2010 (Moon et al., 2010). It is intended to resolve the problem of the left turn at systems interchanges and various forms of this layout are offered. Basically, it requires both freeways to be converted to double decker cross-sections. This makes it possible for left turns to be accommodated by directional ramps.

In the mid-1920s Collins and Hart suggested that, although the ‘flying junction’ resolved the problem of turning traffic at intersections, the cost of its construction rendered it

Figure 11.23 Jug handle interchange. (From Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics*. Transportation System Planning, Boise, 2002.)

unlikely that it would ever gain general acceptance. With respect to the nano interchange, it is thus with reluctance that this prediction is repeated. However, like the 'flying junction', the construction of a four-level interchange would be a costly exercise. The highest level of the structure would be approximately 30 metres above the carriageway located at ground level and two of the ramps would have to climb through this height difference, suggesting that they would be long and correspondingly costly to build.

More importantly, ramps exiting both from the left and the right at more or less similar locations on a carriageway would, in all probability, generate significant levels of turbulence. This is because, where ramps generally enter and exit from the right, through traffic tends to move to the left to avoid the turbulent area. In the nano interchange, changing lanes to the left simply results in vehicles moving into a different turbulent area, with the added complication of following vehicles now being in a driver's blind spot. The developers of this interchange would probably have to demonstrate a solution to this problem before the nano interchange could gain general acceptance.

11.5 GRADE SEPARATIONS

The need for grade separations generally arises from a freeway being planned through an existing street system. Access management requires that some streets simply be cut off and converted to dead-ends or, if a turning area is provided, cul-de-sac. Others may be joined up by means of service roads parallel to the proposed freeway. Where a service interchange could fancifully be described as an interchange without a crossing road, the grade separation could perhaps be described as being an interchange without ramps.

The continuity of the existing road network may, however, partially be retained by taking major roads over or under the proposed freeway without their being linked to it by ramps. The over versus under debate referred to in the case of interchanges applies also to grade separations. The absence of any connection between the two roads does, however, simplify the problem. The fact that abnormally high loads on the freeway are not inconvenienced by limitations of vertical clearance creates a distinct bias in favour of the crossing road being depressed below the freeway. Unfortunately, if it is necessary to provide a sag curve in a cut section in the vertical alignment of the crossing road, the resulting drainage problem may be difficult and possibly costly to resolve.

11.6 RAMP DESIGN

11.6.1 Overview

As is clear from the preceding discussion, the names attached to the various forms of interchange derive from the configuration of the ramps providing the connections between the freeway and its crossing road. The diamond, the cloverleaf, the parclo and the trumpet all derive their names from the shapes of the ramps that form them. To recapitulate, ramps serving the right turn are sometimes known as outer connectors. This name stems from their role in the cloverleaf interchange which was, historically, the first interchange developed. Fifteen years later, the diamond interchange was developed (Leisch, 2002). As there weren't 'inner connectors' on these interchanges with diamond ramps serving turns both to the left and to the right, this nomenclature fell away. Ramps serving the left turn may be either loop ramps or directional ramps. Loop ramps require a 270° change of direction to complete the left turn whereas directional ramps require a change of approximately 90°. It was also pointed out previously that the distinction between 'directional' and 'semidirectional' ramps is no longer drawn. This distinction arose from the directional ramps commencing and terminating to the left of the carriageway whereas the semidirectional ramps required a slight deviation to the right from the rightmost lane of the carriageway before turning to the left and joining the next carriageway, again from the right.

The various ramp types are illustrated in Figure 11.9.

11.6.2 Design speed

A feature of ramps is that, at one end, they should address the design speed of the freeway and, at the other, the design speed of the crossing road. The midsection of the ramp is thus characterised by a transition from the one design speed to the other. Freeways are typically designed for speeds in the range of 100 to 130 km/h and usually of the order of 120 km/h. Where the crossing road is an urban street, the design speed could be 60 km/h or, if an urban arterial, 70 to 80 km/h. Typical ramp design speeds are shown in Table 11.5.

As vehicles undertake the change from the one design speed to the other, it is necessary that the ramp midsection should have a design speed at least equal to that likely to be travelled at any point along its length. This may be a fairly linear interpolation of the speeds at each end of the ramp but, in the case of the loop ramp, the smallest radius on it may be much lower than either of these speeds. For this reason, great care must be applied to the design of

Table 11.5 Ramp design speeds in relation to freeway speeds

Freeway design speed (km/h)	*70*	*80*	*90*	*100*	*110*	*120*	*130*
Ramp design speed (km/h)							
Upper range (85%)	60	70	80	90	100	110	120
Midrange (70%)	50	60	60	70	80	90	100
Lower range (50%)	40	40	50	50	60	70	80

Source: Modified from American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets*, Table 10-1. Washington, DC, 2011a.

an off-ramp employing a loop configuration. Drivers on a freeway have 'acclimatised' to a speed of 120 km/h or higher and a speed of, say, 40 km/h, may feel akin to getting out and walking. In consequence, there is a likelihood that the speed appropriate to the loop may be totally misjudged. A speed far higher than is actually appropriate could thus be selected with the possibility of a resulting ROR crash.

Right turns are addressed by directional ramps, diamond ramps and, at cloverleaf interchanges, outer connections. The directional ramp and the outer connection usually have high-speed tapers at either end and the ramp midsection invariably includes a long circular curve. The upper range of design speeds should thus be accepted for the ramp midsection. It may, however, be necessary to accept a radius falling in the midrange of design speeds for these ramps if the reserve width is in any way constrained.

If loop ramps were to be designed for the upper or even midrange design speed, the travel distance would be so long that the travel time around the loop would be greater than that at a lower speed around a shorter radius loop. Furthermore, the long radius curve would be more expensive to construct than the shorter curve. It would also require a much greater reserve width than the short radius loop, which obviously also has cost implications, particularly in urban areas. Although, for the loop ramp, the lower range design speed would be acceptable, it may be necessary to accept a radius for an even lower design speed but it is suggested that a design speed of 40 km/h should be considered to be the irreducible minimum.

Historically, loop ramps were designed with a single minimum radius curve being placed between two tangents. This resulted in a proliferation of ROR crashes, as drivers could still be travelling at or near 120 km/h when confronted by a 40 km/h curve. The current option is to provide either a spiral or a compound curve that would provide a comfortable transition between the freeway speed and the minimum radius curve provided on the loop. In the case of a compound curve, the ratio between the longer and the shorter radii should not exceed 1.5:1. The design speed of each curve should equal the speed likely to be encountered at the start of the curve and the length of each curve should be sufficient to allow deceleration or acceleration at a comfortable rate to achieve the design speed of the next radius curve at its start.

The calculation of the successive radii of a compound curve and the lengths of the successive curves are illustrated in Table 11.6.

The assumptions made for this illustration are that the design speed of the freeway is 120 km/h and the minimum radius of the loop corresponds to a speed of 40 km/h. The deceleration and acceleration rates are taken as 3 m/s^2 and the maximum superelevation rate, e_{max}, as 10 per cent. The ratio between the radius of successive curves is assumed to be 1.5:1. These values provide a fair indication of typical values likely to be encountered in practice.

The sequence of calculation is shown for deceleration and commences with the radius of curvature appropriate to the design speed of the freeway and thereafter calculation of the

Table 11.6 The successive radii of a compound curve

Design speed			
km/h	*m/s*	*Radius (m)*	*Curve length (m)*
120	33.3	600	43
105	29.2	400	38
90	25.0	270	28
77	21.4	180	25
63	17.5	120	15
53	14.7	80	12
40	11.1	40	–

radii of the following curves based on the assumed ratio of successive radii. The maximum speed appropriate to each radius is interpolated from those tabulated in Table 6.1.

Should the loop increase from its minimum radii to a higher design speed, for example, where the receiving road could be a freeway justifying a design speed of 120 km/h, the radii calculated for the deceleration phase could, for symmetry, be applied in reverse order for the acceleration phase of the turning manoeuvre.

11.6.3 Sight distance

As is the case with all other roads, stopping sight distance must be provided along the full length of the ramp. Values of stopping sight distance are provided in Table 5.6 for desirable and minimum conditions.

The crossing road ramp terminal is usually an at-grade intersection where drivers may elect to turn either to the right or to the left. Very often, dedicated lanes are provided for each of the two directions of turn. Drivers must be able to see the road markings indicating the appropriate lane for each direction of turn.

Where ramps have two lanes and taper down to a single lane, for example, in advance of the freeway on-ramp entrance, drivers would have to see clearly the reduction in ramp width in sufficient time to merge with the vehicles in the other lane. Decision sight distance would therefore have to be provided to the start of the lane drop.

Interchange ramps are geared towards one-way flow so that it is not necessary to make provision for passing sight distance.

11.6.4 Horizontal alignment

As a general rule, the guideline values for the various elements of the horizontal alignment as defined in Chapter 6 apply equally to the design of ramps. A problem arises in that a ramp is a very constricted environment with horizontal (and vertical) curves following each other in quick succession. It could thus be difficult to apply the normal guideline values relating to rates of superelevation development and design exceptions may have to be requested. (Design exceptions are discussed in Chapter 4.)

A particular problem is the tangent between reverse curves, which may be very short in terms of the winding of the superelevation from a curve in one direction to that in the opposite direction. The end result could be an unaesthetic and discontinuous road edge. This problem may be compounded by the presence of a vertical curve. About the only way the problem can be resolved is by using spline curves to define the inner and outer edges of the ramp by means of graphical grading.

A curvilinear freeway alignment may require that an exit ramp terminal should be on a curve to the left. If this curve has a short radius, vehicles exiting from the freeway could be confronted by a seriously sized negative superelevation. One alternative solution could be to extend this superelevation across the off-ramp taper and then to lengthen the ramp midsection, incorporating a section in which the superelevation is rolled over to a normal crossfall before attempting to develop superelevation for the curve that almost inevitably would be in a direction opposite to that on the freeway.

Another solution would be by employing crossover crown lines. On a normally cambered road the difference between the two slopes is of the order of 4 to 5 per cent. Overtaking vehicles can manage this difference comfortably, although trucks with high loads may sway rather alarmingly. A 5 per cent crossover crown between the outer edge of the outside freeway lane and the exit taper could dramatically reduce the extent of the negative superelevation by just about half, and a further crossover crown of the same magnitude across the gore area could reduce the negative superelevation still further. Any remaining runoff would then have to be absorbed on the midsection of the ramp. This may require a fairly long vertical spline curve on the outside edge of the ramp.

11.6.5 Vertical alignment

As implied in Chapter 7, a distinction is drawn between grades and gradients although AASHTO (2011a) uses the terms interchangeably. The grade is to the vertical alignment as the tangent is to the horizontal alignment whereas the gradient, which may be indicated as 1:*X* or, customarily, as a percentage representing the number of metres climbed or descended in 100 metres measured horizontally, refers to the steepness of the grade. This distinction is drawn throughout this book.

When the crossing road is over the freeway, the off-ramp generally has a positive, that is, climbing, gradient and the on-ramp a negative, that is, descending gradient. These are helpful in easing the braking effort required on the off-ramp and increasing the possible rate of acceleration of vehicles on the on-ramp. This does not, however, mean that there is unlimited licence to increase these positive and negative gradients. Controls that limit the maximum gradients include the effect of steep upgrades on truck traffic and their propensity to run away on steep downgrades. Maximum gradients suggested for upgrades are shown in Table 11.7. Where forced by topographic or other constraints, gradients steeper than these may have to be adopted but only after careful consideration of the consequences. The same gradients should apply to downgrades but, if need be, may be 2 per cent steeper.

Ramp profiles generally comprise sag curves at their lower ends and crest curves at their upper ends. Ideally, these curves should be located on the ramp midsection, that is, clear of the ramp terminals. The gradients of the terminals are generally dictated by the alignment of the immediately adjacent roads and, if possible, the added complication of the presence of a vertical curve on the ramp terminals should be avoided. In spite of the fact that, from

Table 11.7 Suggested maximum gradients for upgrades

Design speed (km/h)	*Gradients (%)*
70–80	3–5
60	4–6
40–50	5–7
30–40	6–8

a mathematical point of view, there is no reason why these two curves should not abut each other, there should be a length of tangent, preferably 50 to 60 metres long, between them purely for aesthetic reasons. The *K*-values of these curves should be as indicated in Chapter 7, Tables 7.3 and 7.4.

11.6.6 Cross-sections

The design approach to the width of the travelled way of a ramp is similar to that of a turning roadway at an intersection. It considers various cases, specifically

- Case 1: One-lane operation with no provision for passing a stopped vehicle
- Case 2: One-lane operation with provision for passing a stopped vehicle by another of the same type
- Case 3: Two-lane operation with provision for passing a stopped vehicle by another of the same type

It also considers various traffic conditions:

- Traffic condition A: Predominantly passenger vehicles but with some consideration of rigid chassis trucks
- Traffic condition B: There are sufficient rigid chassis trucks in the traffic stream to govern design but some consideration is also given to semitrailers
- Traffic condition C: There are sufficient semitrailers to govern design

Travelled way widths for the various cases and traffic conditions are shown in Chapter 10, Table 10.7. These widths may be reduced or have to be increased on the basis of the edge conditions of the ramp, specifically whether or not kerbs or paved shoulders have been provided. The modification is shown in Table 11.8.

11.6.7 Ramp terminals

11.6.7.1 Crossing road ramp terminals

Crossing road ramp terminals are usually at-grade intersections which may be priority controlled, signalised or roundabout layouts and are designed as described in Chapter 10.

Table 11.8 Width modification for edge conditions

Edge condition	*Case 1*	*Case 2*	*Case 3*
No stabilised shoulder	None	None	None
Stabilised shoulder one or both sides	Lane width for Conditions B and C on tangent may be reduced to 3.6 where shoulder is 1.2 m or wider	Deduct shoulder width(s) Minimum pavement width as under Case 1	Deduct 0.6 m where shoulder is 1.2 m or wider
Sloping kerb	None	None	None
Vertical kerb			
One side	Add 0.3 m	None	Add 0.3 m
Two sides	Add 0.6 m	Add 0.3 m	Add 0.6 m

Source: American Association of State Highway and Transportation Officials (AASHTO). *A policy on the geometric design of highways and streets*, Table 3-29. Washington, DC, 2011a.

A major difference between normal intersections and crossing road ramp terminals is that, although the latter may resemble four-legged intersections, they are actually operationally more in the nature of two three-legged or T-intersections back to back. The cross-piece of the T is normally a two-way road, whether divided or not and the stem of the T is a one-way road either arriving or departing from the intersection.

The design of the crossing road ramp terminal has to take cognisance of the possibility of a driver entering an off-ramp in the belief that it is an on-ramp. This problem is addressed in Section 11.7.

11.6.7.2 Freeway ramp terminals

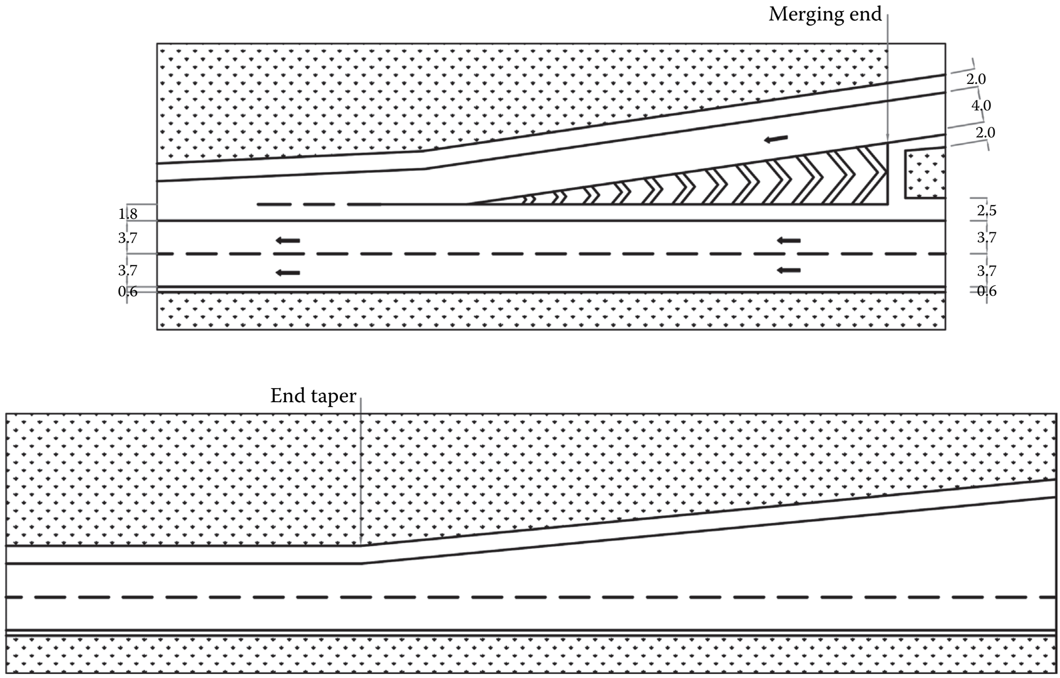
Freeway ramp terminals typically take the form of high-speed tapers in support of diverging and merging manoeuvres at freeway speeds but may include parallel sections, effectively auxiliary lanes abutting the freeway outside lanes. Tapers should not be preceded by curves so that the break point between the edge of the freeway outside lane and the ramp edge is a highly visible discontinuity. This is a form of guidance to the driver that a different flow regime is being initiated. The taper rate selected should, however, be such that, in changing direction from the freeway onto the ramp or from the ramp to the freeway, a vehicle travelling at least at the design speed could follow a smoothly curving path without invoking an excessive side force friction or encroaching either onto the shoulder or onto the gore area (Wolhuter and Garner, 2003).

Drivers on the on-ramp need to select a suitable gap in the through traffic on the freeway. The distance required to select the gap and then to move into it depends to a certain extent on the gap acceptance characteristics of the driver. It also depends on the availability of acceptable gaps in the freeway traffic stream. The selection process takes place in the time it takes the vehicle to travel from the merging end to the point where the ramp left edge intersects the right edge of the freeway, the WLBP previously discussed. Allowing the driver 10 seconds to complete the selection process and assuming that a freeway speed of 120 km/h has already been achieved, a distance of 330 metres would be travelled during gap selection. At the gore, the lane and ramp edges are typically separated by a distance of 6.5 metres, comprising the 2.5 metres of the freeway shoulder, the 2.0 metre wide nose area and the 2.0 metres of the ramp shoulder. Reducing this gap to zero over a distance of 300 metres leads to the widely accepted taper rate of 1:50.

Gap acceptance is not an issue at off-ramps but the path actually followed by the vehicle in leaving the freeway needs to be modelled. The model offered assumes that

- The vehicle will be traveling at the design speed of the freeway.
- Its path will commence and terminate with the vehicle positioned in the centre of lane 1 of the freeway and the ramp respectively, which are taken as being 3.0 metres and 4.0 metres wide respectively.
- At no stage will the vehicle encroach on the shoulders of the lane or the ramp.
- Only normal crossfall taken as being 2.5 per cent is available.

Values of superelevation listed in Table 6.7 and e_{max} = 10 per cent suggest that a radius of 3000 metres would be comfortable at a design speed of 120 km/h. For a design speed of 100 km/h the radius could be reduced to 2000 metres. Simple geometry demonstrates that the taper rate could be of the order of 1:20 for a design speed of 120 km/h and 1:15 for design speed of 100 km/h.



Note: For 50 m in advance of the merging end, the ramp should not be more than 0.75 metres lower than freeway.

All dimensions in metres unless otherwise stated.

Figure 11.24 One-lane on ramp.

Obviously, where

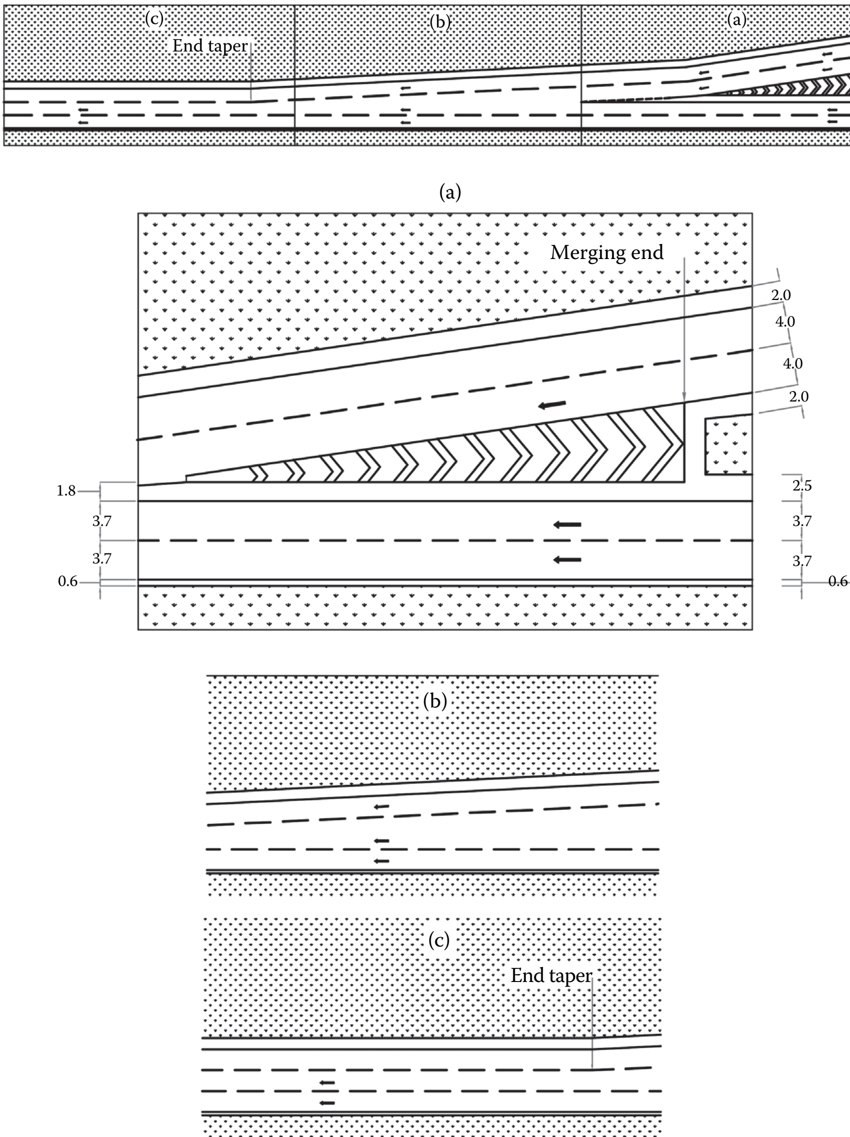
- Design speeds can vary from 80 km/h to 120 km/h or more
- Maximum superelevation can vary from 4 per cent to 10 per cent
- Ramp widths can vary over a wide range as shown in Table 10.7 as modified in Table 11.8
- The dimensions of design vehicles are country dependent

It should be clear that the taper rates offered in the preceding text are intended purely to be illustrative.

The detailed dimensions of one- and two-lane on- and off-ramps are illustrated in Figures 11.24 to 11.28 (AASHTO, 2011a).

11.7 WRONG-WAY DRIVING

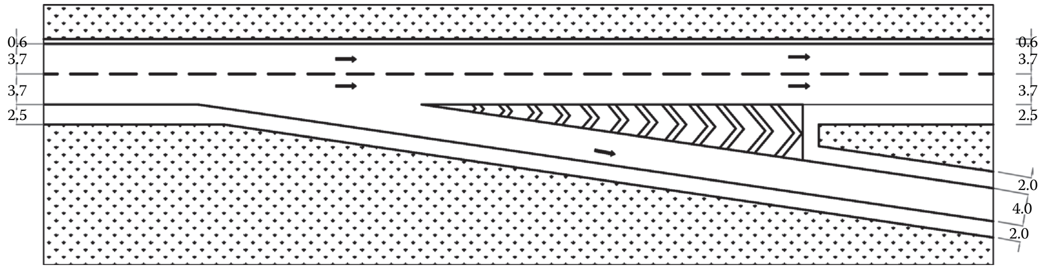
Ever since freeway construction commenced in the United States in the 1950s, there have been concerns about the problem of wrong-way driving. If a vehicle enters the main line carriageway travelling in the wrong direction, the likelihood is that a crash is almost inevitable



Note: For 50 m in advance of the merging end, the ramp should not be more than 0.75 metres lower than freeway.

All dimensions in metres unless otherwise stated.

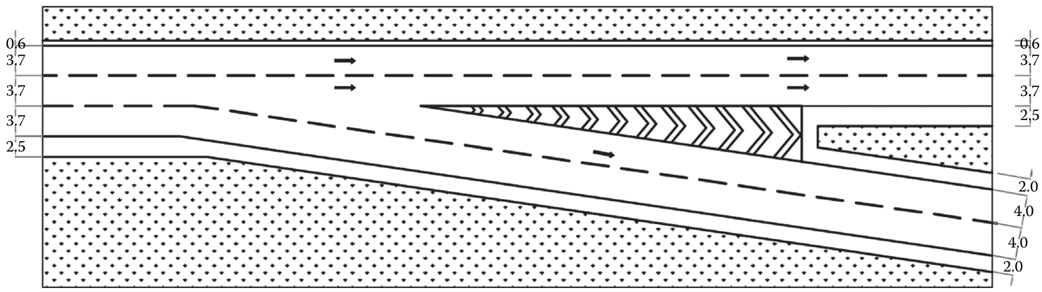
Figure 11.25 Two-lane on ramp.



Note: Sight distance measured from an eye-height of 1.05 metres to the road surface should be provided in advance of the nose.

All dimensions in metres unless otherwise stated.

Figure 11.26 One-lane off-ramp.



Note: Sight distance measured from an eye-height of 1.05 metres to the road surface should be provided in advance of the nose.

All dimensions in metres unless otherwise stated.

Figure 11.27 Two-lane off-ramp.

Figure 11.28 A median inhibiting wrong-way driving.

and there is a strong likelihood that it will also be fatal. In the United States wrong-way driving accounts for, on average, 350 fatalities a year nationwide. Sixty per cent of wrong-way crashes are attributed to substance abuse and 80 per cent take place after dark. It is possible that these two factors are not unrelated.

As previously stated, the accident black spot is well known and it has been found that some interchange layouts are more prone to be the sites of wrong-way driving incidents than others. In general, cloverleaf interchanges have a good safety record. It deserves to be noted that, at the diverging diamond interchange, the wrong-way movement is impossible. Research has shown that the two-quadrant parclo is the prime offender in terms of wrong-way incidents. This is very surprising because the crossing road ramp terminals look exactly like conventional three-legged intersections so that finding the correct turning path through them should not present any problems. This, of course, assumes that, on the approach to the ramp terminal, both directions of flow at the intersection are simultaneously visible to the driver.

In the early days of freeway development, diamond interchanges were a major site for wrong-way driving incidents. This was attributed to drivers who were unfamiliar with driving on freeways assuming that the first ramp they came to was the way to the freeway. With time, it was learnt that one had to cross over the freeway before essaying a left turn and the frequency of wrong-way driving incidents plummeted. It is, however, still possible that drivers on the crossing road could mistake an off-ramp for their intended destination, the following frontage road, and turn prematurely.

Most states seem to subscribe to the belief that providing drivers with appropriate information through signing and road markings is the best way to go in eliminating the wrong-way problem. And have safety programmes set up on this basis. Laudable though this approach is, it would appear that relatively little effort has been put into making interchanges fool-proof by making it difficult, if not actually impossible, to execute an incorrect turning movement.

All wrong-way driving incidents are initiated at the crossing road ramp terminals, specifically at the terminals of off-ramps. The driver of a vehicle on the crossing road arrives at an off-ramp terminal and mistakes it for an on-ramp terminal. The real problem is that the crossing road ramp terminal may look like a conventional four-legged intersection, particularly if the off-ramp has a two-lane cross-section at the terminal. All drivers are very familiar with four-legged intersections where turns can be executed either to the right or to the left so that the likelihood of an incorrect manoeuvre exists. It needs to be made abundantly clear to the driver that this is NOT a conventional four-way intersection and is actually two T-intersections operating back to back.

If the crossing road ramp terminal is, in fact, a T-intersection, as is the case with a parclo, the choice of turns to the driver on the cross-leg of the T should be limited to turning to the right if the approach to the T is from the left or turning to the left if travelling in the opposite direction. If the stem of the T-intersection happens to be an off-ramp, the alternative turn would result in wrong-way driving. It should therefore be made very difficult for drivers from either direction to enter the off-ramp. At the same time, drivers legitimately on the off-ramp must have the ability to turn either to the left or to the right onto the crossing road and they should not be inconvenienced by efforts to avoid incorrect movements by the opposing flows of traffic.

The solution to the problem of wrong-way driving in both cases, the apparent four-legged intersection and the T-intersection, lies in addressing the layout of the T-intersection supporting it by clearly providing the information required by the driver to make correct decisions. The most obvious source of information is the road itself. If the driver is able to see the intersection in its entirety, it is unlikely that he or she would enter the off-ramp by mistake. Obscuring the view of the on-ramp, for example, by barriers on the median island, could create the impression that the off-ramp is the only entrance to the freeway, resulting in a wrong-way turn. Barriers on the ramp median and any other obstructions to this sight line should be terminated well in advance of the terminal.

A frontage road that is very close to the off-ramp may create the false impression of the off-ramp being part of a parclo two-directional T-intersection. The solution in this case is the reverse of the aforementioned by deliberately obscuring the sight line to the frontage road by the provision of a berm between it and the off-ramp.

In addition to adjustments to the intersection sight distance, a further technique is simply to make the incorrect turning manoeuvre difficult to execute. Kerb roundings should be replaced by kerb lines meeting at sharp angles. To negotiate the incorrect turn may require the driver to turn so wide that the vehicle could possibly end up on the far shoulder of the ramp. This should not only send a clear message to the driver about the correctness or otherwise of the turn that has just been completed but would have the additional benefit of at least placing the vehicle out of harm's way.

A teardrop island on the crossing road would offer clear guidance to the driver leaving the off-ramp towards a turn to the left. Simultaneously, the ramp could be angled towards the right, forcing the wrong-way manoeuvre to be through more than 90°.

Techniques whereby wrong-way driving can be counteracted are illustrated in Figures 11.28 and 11.29 (Zhou and Rouhlamin, 2014). Figure 11.28 shows a raised median within the interchange area even if the crossing road is undivided elsewhere. The median makes a wrong-way right turn onto the exit ramp impossible, particularly if the opening in the median for the left turn from the exit ramp is narrow and displaced towards the left.

As shown in Figure 11.29, the right turn onto the exit ramp can be made more difficult by skewing the ramp towards the downstream or right side and not providing a curve in the

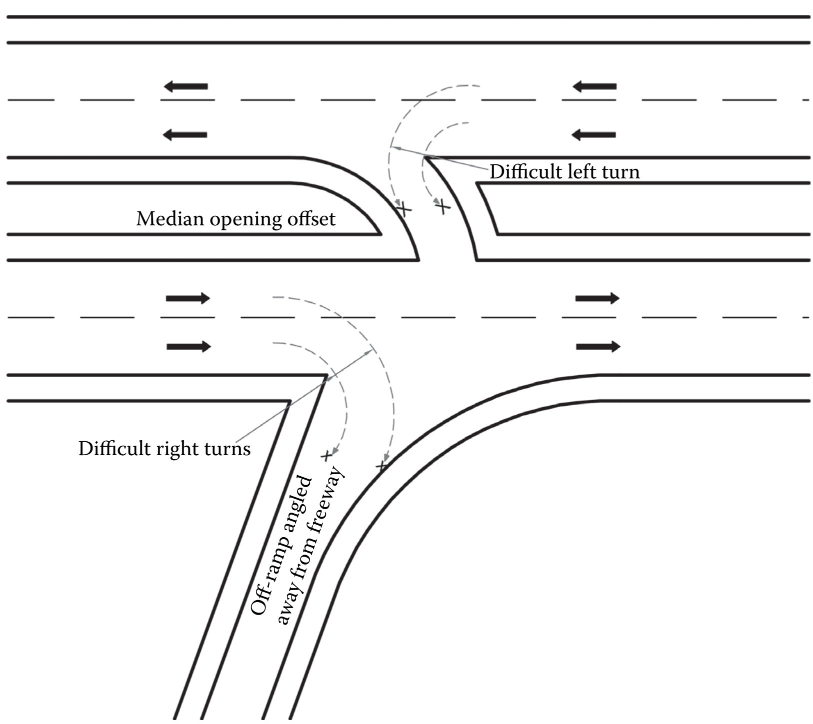


Figure 11.29 Further techniques to prevent wrong-way driving.

kerb line at the intersection of the ramp and the crossing road. In addition, street lighting provides a clear view of the terminal area at night so that drivers can easily see the path they are required to follow. The crossing road ramp terminals should be well clear of any local network intersections to ensure that they are not perceived as being part of the local street network with the possibility of confusion about their one-way status in an otherwise two-way road network.

Chapter 12

Low-volume roads

CONTENTS

12.1 INTRODUCTION

In most countries, the total length of road network includes a high percentage of unpaved roads. American practice suggests that, because the traffic counts normally encountered on these roads fall in the range of 400 to 500 vehicles per day or less, they are referred to as very low-volume roads (Neumann, 2001). It is estimated that 80 per cent of the roads in the American road network fall into this category. However, any American who has visited a Third World country will appreciate that the difference between a very low-volume road and a conventional two-lane two-way road in America is minute when seen in the context of most African countries.

In these countries, there generally is a dearth of professional skills and financing of road construction is usually through the medium of foreign aid. In consequence, many African roads are designed by international consultants, some of whom are inclined to transplant the design standards of their own countries into the new environment, little realising the inappropriateness of their actions. This chapter therefore does not address the First World environment and focusses rather on the Third World situation in the hope that it may offer some guidance on what to most professionals would be unfamiliar territory.

Simply adopting standards from developed countries does not automatically result in acceptable levels of safety on Third World rural roads. Furthermore, Esra Hauer in the opening address to the Third International Symposium on Highway Geometrics pointed out that a road designed to standards was not necessarily safe. In these countries, the traffic mix is totally different from that of First World countries. It includes relatively old, slow-moving, usually overloaded and unroadworthy vehicles; large numbers of pedestrians; animal-drawn vehicles, not to mention untended animals such cattle, donkeys, water buffalo and the occasional camel; and, possibly, small motor cycles. Traffic movement is characterised by poor driver behaviour and negligible enforcement of regulations. Methods to improve safety through engineering design are thus extremely important (Ministry of Roads and Bridges, South Sudan, 2013).

A high percentage of the populations of Third World countries live at significantly below what elsewhere would be described as the poverty datum line of $1 per day. Privately owned passenger cars are thus a rarity, and most travel is by public transport, bicycles or walking. The *Guideline: Low-Volume Paved Roads* notes,

> The criteria for defining a Low-Volume road vary significantly in various parts of the World. In the SADC region, such roads may be primary, secondary or tertiary/access roads. They typically carry less than 200 vehicles per day and often include non-motorised traffic, particularly near populated areas.
>
> The successful provision of a low-volume road requires ingenuity, imagination and innovation. It entails 'working with nature' and using locally available, non-standard materials and other resources in an optimal and environmentally sustainable manner.
>
> (Pinard et al., 2003)

In the preceding extract, SADC is the acronym for the Southern Africa Development Community. The Community comprises Angola, Botswana, Democratic Republic of the Congo, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia and Zimbabwe.

In Third World countries, the majority of the network, including major links between towns and cities, carries very low volumes of traffic. The *Guideline* referred to earlier focusses on surfaced roads because of their role in the functional classification of roads rather than because of traffic volumes on them.

Traffic volumes are often too low to warrant the possible increase in construction and maintenance cost attaching to the higher standards associated with surfaced roads. It is suggested, however, that the defining characteristics of these roads are not their function in the network or even the traffic volumes on them. Low-volume roads are often either gravelled or earth roads. Where gravel or earth roads are discussed they will be referred to generically as 'unpaved roads'. A typical unpaved road in good condition is shown in Figure 12.1.

Materials engineers question the practice of surfacing low-volume roads partly on economic grounds but also because bitumen pavements require the kneading action of traffic to preserve their flexibility. Without this flexibility, cracks develop, allowing the ingress of water into the base course with the consequent failure of the pavement.

This chapter is aimed primarily at the provision of low-volume roads in Third World countries where labour-based and labour-intensive construction are often employed in support of job creation. A distinction is drawn among labour-based construction, labour-intensive construction, and community-based construction (Watermeyer, 2003). Labour-based construction incorporates a mix of small items of plant with labour whereas labour-intensive construction seeks, as far as is practicable, to replace capital plant with human sweat. Community-based construction is focussed on imparting managerial and administrative skills to the community and to promote the emergence of local contractors.

The adoption of these techniques stems largely from socioeconomic imperatives that accept that labour-based and labour-intensive constructions are typically more expensive than capital-intensive construction with its focus on the use of large plant items. The cost

Figure 12.1 Typical unpaved road. (Photo: RC Burrell.)

of social grants aimed at supporting unemployed people and their often large and extended families has to be added to the cost of capital-intensive construction to draw a true comparison between it and the cost of labour-based construction. With this added cost component, the difference between the two forms of construction declines quite dramatically.

12.2 LABOUR-BASED CONSTRUCTION

Given that the level of development of Third World countries is low, it follows that best use must be made of limited resources to rapidly provide strategic routes that offer year-round all-weather access and also routes that provide basic access to the majority of the population most of the time. All roads should be designed to have the minimum of undesirable side effects on the environment and be constructed in a manner that will be supportive of the development of local capacity.

Factors that have to be considered in the decision to adopt labour-based construction are summarised in Table 12.1, which is modified from Table 3.3 of the SADC *Guideline: Low-Volume Paved Roads*.

As indicated in Table 12.1, labour- based construction requires a philosophy and approach to design completely different from the way one looks at conventional design. In fact it can even be questioned whether design speed is appropriate in the selection of design standards in the case of very low traffic volumes. Excavation by pick and shovel and transport of material by means of wheelbarrows cannot compete with front-end loaders and 20-tonne trucks in terms of productivity and speed of construction. It follows that reducing earthworks quantities by keeping the gradeline as close as possible to the natural ground level makes sense. As discussed further in Section 12.9, this has an impact on the drainage of the road.

Particularly where towns are remote from each other, traffic on the roads between them move at high speeds. These roads, which often would be the main links between towns,

Table 12.1 Factors impinging on the application of labour-based construction

Environment	*Factor and implication*
Political	Government policy • Influences practice. Covers issues such as poverty alleviation, sustainable socioeconomic development, technology choice, employment creation, geometric and materials standards, sources of funding. Political perceptions • The tendency to favour conventional approaches and standards with perceived minimal risk attached to them needs to be combatted. Innovative approaches and nontraditional standards need to be communicated effectively. Political involvement • Political involvement tends to influence decision making. It is necessary to highlight the pros and cons in a balanced, transparent manner and maintain a continuous dialogue with stakeholders.
Social	Poverty alleviation • This implies the use of labour-based or labour-intensive construction rather than capital-intensive methods. Sustainable livelihood • Local participation and resource mobilisation would be enhanced by involving the people who will ultimately benefit from the projects. Gender considerations • Gender biases could be removed by integrating the transport needs of women in the mainstream of policy and planning. • Participation of women in labour-based construction and maintenance programmes should be promoted as well as training provided for women to assume supervisory roles.
Institutional	Organisation • There is a growing trend towards the establishment of more autonomous central and local roads authorities. • There is a need for a greater scope of generating accountability for results in roads programmes and moving towards the contracting of work out to the private sector.
Technological	Technology choice • Technology supportive of labour-based construction needs to be developed and refined.
Economic	Evaluation • Road benefits are not limited to the use of the road but also include the way in which roads are financed, designed, constructed and maintained. Monetary and irreducible factors (defined in Chapter 16 as benefits to which a cash value cannot be attached) need to be captured.
Financial	Funding • Because funding is usually scarce there is a need to look at minimum standards, limited donor funding and local funding of recurring maintenance costs. Sustainability • There is a need to commercialise operations where possible and to involve stakeholders in the maintenance of facilities.
Environmental	Impact • Social as well as environmental impacts need to be captured in the evaluation of unsurfaced roads. • Health-threatening impacts need to be addressed as a high priority.

Source: Pinard MI, Ellis C-H, Johansen R, Toole T et al. *Guideline: Low-volume paved roads*. Southern African Development Community, Maputo, 2003.

whether or not they are surfaced, should therefore be designed to the standards appropriate to paved roads. It should also be borne in mind that they are more likely to require an upgrade to surfacing than most other unpaved roads. It follows that the design standards should be sufficiently high to ensure that upgrading would be limited to the addition of base course and paving to the existing gravel wearing course which, most probably, would already be to, at least, sub-base specifications.

The strip roads widely adopted in Zimbabwe (formerly Southern Rhodesia) as illustrated in Figure 12.2 are an interesting interim stage between gravel and bituminous surfacing.

These roads were built from 1933 onwards to provide links between towns in Zimbabwe because there wasn't sufficient funding to build surfaced two-lane roads between them. By the end of World War II a total of more than 3300 kilometres of strip roads had been built including from Harare to the southern border town of Beit Bridge, a distance of about 500 kilometres. Beit Bridge crosses the Limpopo River, which is the boundary between Zimbabwe and South Africa.

When vehicles approached each other on a strip road, each would move away from the centre of the road and use only one of the strips until the other vehicle had been passed. From personal experience in the mid-1950s, this manoeuvre was surprisingly difficult to achieve because the gravel on either side of each of the strips invariably was somewhat lower than the strips. Vehicles could rock quite alarmingly while moving across, and ending the manoeuvre with the inside wheels actually on one of the strips was sometimes more the

Figure 12.2 Rhodesian strip road. (From Rustyproof, 1963.)

result of blind luck rather than of skilful driving. Strip roads in Zambia usually comprise a single strip wide enough to accommodate a single vehicle in the middle of the road.

12.3 THE PRINCIPLES OF DESIGN OF LOW-VOLUME ROADS

For unsurfaced roads, an argument could be made in support of geometric standards higher than the minima recommended for surfaced roads. This argument is based on the level and variability of the coefficient of friction of a gravel surface and also on the susceptibility of gravel and earth surfaces to scour. The coefficient of friction on gravel is as variable as the quality of the gravel surface, which could range from excellent riding quality on a firm smooth surface to a corrugated surface comprising rounded aggregate without the benefit of a binder. The latter is akin to driving on ball bearings. It has been found that the coefficient of friction of a gravel surface is roughly half that of a paved surface, typically in the range of 0.3 compared to the 0.6 to 0.7 of a dry bitumen surface. This is similar to the coefficient adopted for design purposes for paved roads in order to build a safety factor into the design. If applied to a gravel road, it follows that there is, in fact, no safety factor at all. Theoretically speaking, geometric standards should therefore be higher on these roads than on paved roads. The problem of scour is discussed further in Section 12.5.

Unfortunately this argument fails in the face of economic reality and dwindles to the level of a nice-to-have.

The whole-life cost structure of a road has the components of construction cost, maintenance cost and road user cost. On roads where traffic volumes are high, savings in road user costs, principally in the form of time savings and crash reduction, would justify the use of high standards of geometric alignment. The breakdown of the whole-life cost of a high-volume road typically has the breakdown of 75 to 80 per cent road user cost, with the balance being split roughly equally between construction and maintenance. With a usage of as few as 100 vehicles per day, the bulk of expenditure no longer lies in road user costs and is mainly in the areas of construction and maintenance, predominantly in maintenance. Furthermore, even with a crash rate per 100,000 vehicle kilometres comparable to that on high-volume roads, the volume of traffic on a gravel road would be so low that the frequency of crashes would be minuscule. Savings arising from a reduction in what is already a very low number thus simply would not justify the expenditure involved.

The traditional or First World approach to the selection of geometric design standards is therefore totally inappropriate to roads carrying low volumes of traffic regardless of whether they are surfaced or gravelled. The First World approach was based primarily on considerations based on safety factors that would be reasonable given the overall costs and benefits of road provision but without going through the rigour of a proper economic evaluation. In the Third World, roads should be designed with consideration of

- Available materials (and the modification of unsuitable materials where this is possible)
- Minimisation of steep gradients because of the risk of scour resulting in a road being prematurely totally damaged
- Flood risk
- The development of construction capacity
- The intended maintenance regime

The development of construction capacity has been mentioned earlier. This would usually take the form of a government initiative and comprise a graded programme whereby a person could be trained as a basic labourer and then, with experience, further training and skills development, move up the scale of competence towards becoming a plant operator, ultimately to move into the realms of capital-intensive construction with his or her own contracting business.

Poor maintenance has been the nemesis of many African roads. Gravel roads require directing effort to ongoing or recurring maintenance to ensure that damage does not occur rather than repairing damage that has occurred. The development of maintenance capacity should be built into the design process by the incorporation of the assessment of maintenance capacity and the development of capacity enhancement initiatives during construction.

As an example of these, the lengthman programme applied in Kenya and other African countries employs previous construction workers who live in close proximity to a road as contractors to maintain sections of road typically 1.5 to 2.0 kilometres in length. They are provided with wheelbarrows and simple hand tools and are required to work for 3 days a week, thus allowing time to work on their own lands or on whatever other opportunities present themselves (Jones and Petts, 1991). Figure 12.3 illustrates the tools at the start of a day of labour-based construction.

Design has to be flexible in the sense that a changing circumstance may require a change in design so that these factors can vary not only from road to road but even from place to place along a road. For example, a gravel road may be provided with a bitumen surface over a section where the gradient is in excess of about 6 per cent and likely to cause scour. In bad weather, rain on a gravel surface may result in a loss of traction, effectively stalling vehicles on an upgrade and causing vehicle travelling in the opposite direction to slide out of control.

Major roads often include a fair proportion of tourist traffic and hence unfamiliar drivers in the vehicle mix. These drivers could brake sharply and unexpectedly to avoid overrunning an intersection at which they would have wished to turn. This normally is a result of inadequate or badly weathered signage or even the theft of road signs for use as building material for shacks. Generous provision of intersection sight distance and signage in addition to high standards of horizontal and vertical alignment in the vicinity of intersections should be provided to minimise the likelihood of this occurring.

Figure 12.3 Typical lengthman tools. (Photo: RC Burrell 2005.)

Minor roads mainly serve drivers familiar with their geometry whereas other drivers would be inclined to proceed cautiously on entering unfamiliar territory, especially if it comprises a narrow gravel road. They would drive fairly slowly until driver expectancy and some history on driving in the area had been acquired, at which stage they may increase their speed. It is thus possible that design speed is not an entirely appropriate criterion on which to base the selection of geometric design alignment elements. As stated in Chapter 4, a major weakness of design speed as a measure of consistency is that it only defines the minimum geometric standards that could be adopted in the design of a road. The question should be asked whether a section of road comprising tangents that are 10 or more kilometres long linked by a 60 km/h curve can truly be described as having a design speed of 60 km/h. True consistency of design could thus be achieved only if the entire road had geometric standards that are close to minimum values. Design consistency is discussed in Chapter 4.

The risk attaching to the adoption of low design speeds or the selection of subminimum design standards is relatively low because it applies only to those invariably short sections of road where these standards have been adopted. Drivers familiar with the road would know where these sections are and would proceed with caution where required. However, where a road has a primarily recreational function, such as access to rivers, lakes and other scenic areas, the percentage of unfamiliar drivers would be high. They could also be towing caravans, boats, and so on. Standards of horizontal curvature should therefore be above minimum wherever possible and adequate signposting, including chevrons and advisory speed limits, installed and maintained.

12.4 THE SELECTION OF GEOMETRIC GUIDELINE VALUES

There is a worldwide misconception that, as stated previously, a road designed to standards is safe with the corollary that, if not to standard, it is unsafe. In the latter case, the road authority is automatically exposed to litigation with the prosecutorial accusation of 'sub-standard, ergo unsafe' and is then in the situation of having to prove that it has 'applied its mind' to the design. The misconception relates to the belief that a standard is tacitly a quality of perfection. Any departure from perfection must, by definition, have less than perfection as its end result. The South African judge who ruled that a bridge should be 'designed and built to withstand the worst storm of all time' obviously overlooked the issue of affordability. The word 'standard' is deeply entrenched in the geometric design process but should be eradicated. It has the sense of some or other value of a parameter that may not be exceeded. It can be a maximum value, such as gradient, or a minimum value such as radius of horizontal curvature. Either way, it is not really set in concrete although many departmental officials regard their geometric design manual as a bible to be slavishly followed without the freedom to select design values appropriate to a given circumstance.

The preferred substitute for 'standard' is 'guideline'. Designers are expected to think and not to simply look up a value in a book and drop it into their design. This philosophy applies particularly to low-volume or unsurfaced roads. These roads are typically right at the limit of what is acceptable and are correspondingly more difficult to design than conventional roads. The level of judgment required in the design of an affordable and reasonably safe, durable low-volume road can come only after years of experience.

In the sections that follow, it is to be understood that values of 'standards' that are offered are not measures of excellence or perfection. They are guidelines that have been found either empirically or through rigorous research to be workable and affordable – in other words, practical. Where a designer finds that, for whatever reason, it is not practical to apply the recommended value, it may be necessary to request a design waiver as described in detail in Chapter 4.

Assuming that the flexibility tacit in the adoption of the use of guidelines as opposed to rigidly enforced standards is accepted, some thought needs to be applied to the selection of appropriate values of the various parameters and elements to be embodied in the design. The process of selection of these is illustrated in Figure 12.4.

Step 1 reflects the need to reflect traffic growth over the design life of the road. This is based on the judgment of the designer because gravel roads do not justify the expense of

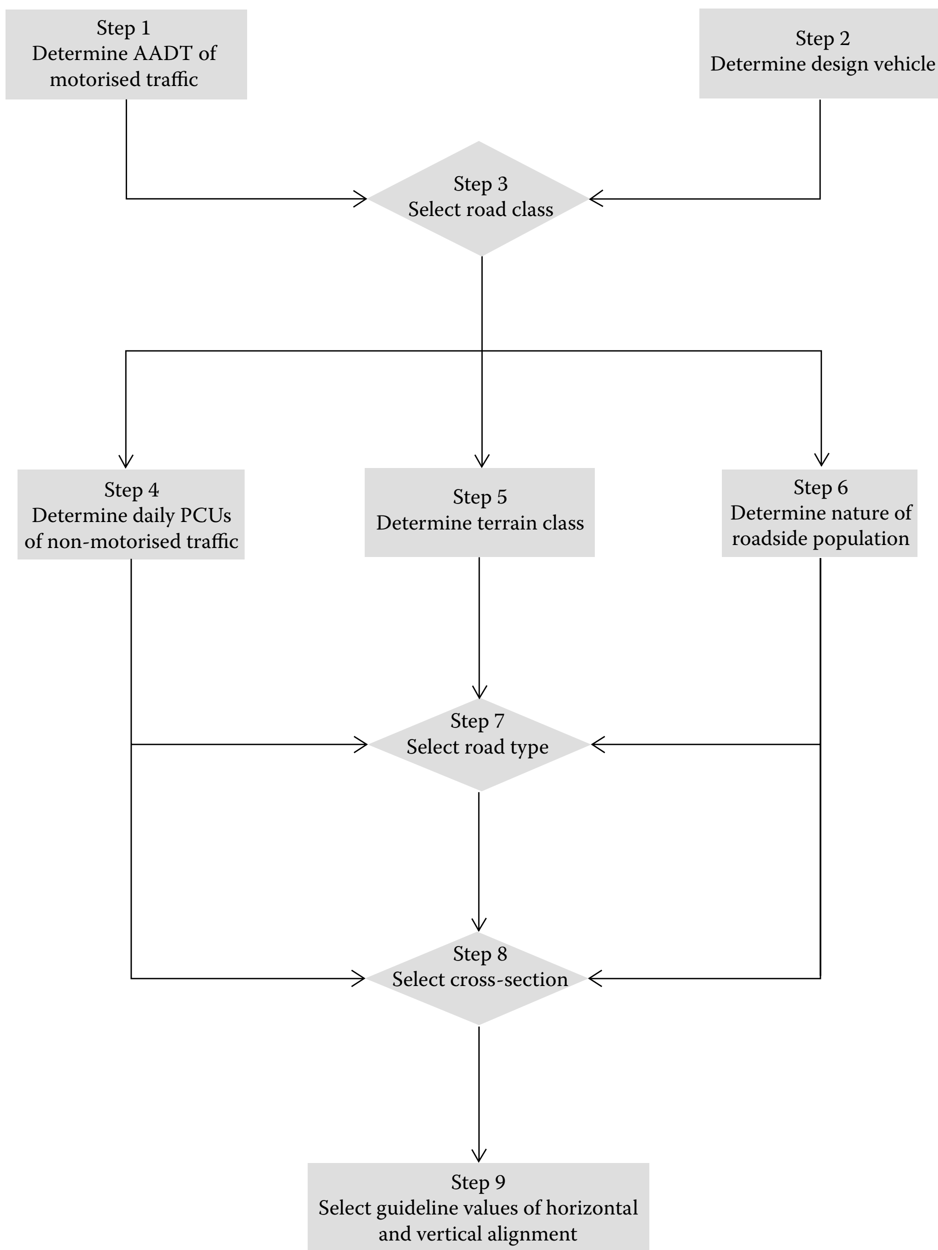


Figure 12.4 The selection of appropriate guidelines. Abbreviations: AADT, annual average daily traffic; PCU, passenger car unit. (Adapted from Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013.)

a transportation study. A point to bear in mind is that through roads are likely to have a higher growth rate than access roads.

Step 2 is based on the composition of the traffic stream. In developing countries reliance on public transport by impoverished people causes there to be a high percentage of buses in the traffic stream, with trucks often also being pressed into service as people movers. These vehicles are generally slow moving and can create considerable congestion. On low-volume roads congestion is not a problem but, if vehicles are present in sufficient numbers, their ability to pass each other could become a safety issue. Furthermore, trucks and buses in Third World countries are often poorly maintained and are inclined to break down frequently, creating a further obstruction to moving vehicles.

Step 3 follows directly from steps 1 and 2 on the basis of traffic volumes. Although low-volume roads are all those with traffic volumes of fewer than 400 vehicles per day, it is necessary to divide this group of roads into smaller subgroups because the very lowest traffic volumes do not require the guideline values applicable to the higher volumes. This subdivision is shown in Table 12.2.

Step 4 defines the impact of pedestrians and nonmotorised vehicles on safety. If these are present in high numbers, it may be desirable to widen the cross-section. If the road is surfaced, the widening is applied to the shoulders to allow for animal-drawn vehicles without their being mixed with the motorised traffic stream.

Step 5 is particularly important, as the terrain has a direct bearing on the cost of construction and maintenance of the road. Four classes of terrain are defined in Table 12.3.

Step 6 refers to the nature of off-road activities, principally trading and other business activities but also would-be passengers waiting for busses and other modes of public transport. Roadside trading usually takes the form of food, fruit and curios for tourists such as hand-carved wooden animals, beads and the like.

Special provision needs to be made if traders, in any numbers, occupy a large amount of space as well as for parking by tourist coaches and passenger cars. People wandering around and looking at what is on offer would not have their minds or eyes on passing traffic. Off-road parking should thus be provided for preference between the trading area and the roads so that pedestrians are accommodated beyond the parking area. It follows that a large area may have to be provided. If the parking area requires surfacing, the surface should either be brick, interlocking road stones or tar because oil spills would dissolve bitumen, leading to

Table 12.2 Subgroups of traffic volumes per day for selection of guideline values

Subgroup	*AADT (vehicles per day)*
1	150–400
2	75–150
3	25–75
4	<25

Note: AADT, annual average daily traffic.

Table 12.3 Type of terrain

Type of terrain	*Typical number of 5-metre contour spacings per kilometre*	*Average ground slope perpendicular to contours (%)*
Flat	0–10	<3
Rolling	11–25	3–25
Mountainous	26–50	>25
Escarpment	>50	>25

the creation of potholes. These would typically not be reinstated, leading ultimately to the total destruction of the parking area. Furthermore, parked busses and trucks would require properly compacted base course and sub-base if the parking area is to have a reasonable life span.

Step 7 is the major guide to required pavement designs and guideline values of horizontal and vertical alignment. As pointed out previously, the lack of durability of gravel surfaces may result in poor riding quality of the road and could require higher standards of geometry than surfaced roads to avoid scour of the road surface. At the same time, the economics of the matter suggests that standards should be held low to achieve savings in construction costs. The dilemma of simultaneously meeting these diametrically opposed requirements does not normally present itself in the design of conventional roads and is one of the reasons why the design of low-volume surfaced or unsurfaced roads is significantly more complex than the design of conventional roads.

Step 8 defines the cross-section needed to match the requirements discussed in the preceding seven steps. Cross-section design is discussed in Section 12.7.

12.5 BASIC DESIGN PARAMETERS

12.5.1 Design speed

Design speed was previously defined as the maximum speed at which a road could be negotiated when weather, traffic and light conditions were so favourable that the design features governed the speed. The 'maximum' was actually taken as the 85th percentile speed. Because of the weaknesses of the design speed as previously discussed, this definition has been universally abandoned. AASHTO (2011a) defines design speed as the speed selected for design. It is, in fact, no longer a speed and should rather be considered simply to be the collective name for a group of guideline values.

Design speeds normally used for various topographies and ranges of traffic volumes are given in Table 12.4.

12.5.2 Stopping sight distance

As shown in Chapter 5, stopping sight distance is measured from an eye height of 1.05 metres to various object heights. It is based on a reaction time of 2.5 seconds and a recommended rate of deceleration of 3.0 m/s^2 (with various authorities proposing rates higher than this). On gravel roads it is necessary to adopt a lower rate of deceleration and it is suggested that a deceleration rate of 2.5 m/s^2 would be more appropriate. The stopping sight distance on a gravel road calculated on this basis is compared to that on a paved road in Table 12.5.

Table 12.4 Design speeds

Traffic volume (vehicles per day)	*Design speed for various qualities of topography (km/h)*				
	Flat	*Rolling*	*Mountainous*	*Escarpment*	*Urban*
150–400	70	60	50	40	50
75–150	70	60	50	30	50
25–75	60	50	40	30	50
<25	50	40	30	20	40

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual.* Juba, 2013.

Table 12.5 Comparison of stopping sight distances

Design speed (km/h)	*Paved roads*	*Gravel roads*
20	20	20
30	30	35
40	50	60
50	65	75
60	90	100
70	110	125
80	140	150
90	170	190
100	200	225
110	230	265
120	270	310

12.5.3 Meeting sight distance

Meeting sight distance is based on the ability of two vehicles approaching each other to be able to decelerate to a stop without hitting each other. This addresses the problem of the single-lane road. The meeting sight distance is double that offered in Table 12.5 for gravel roads and is similar to the rationale offered in Chapter 5 with respect to barrier sight distance for paved roads where the deceleration rate is taken as 3.0 m/s^2. It may be useful to allow a modest margin of, say, 10 metres to ensure that there will be a gap between the vehicles after stopping.

12.5.4 Passing sight distance

It is not possible at this stage to offer any useful guidance on passing sight distances on gravel roads. As described in Chapter 5, passing sight distance requires a very elaborate model with numerous assumptions. It is possible that the passing sight distances offered in Chapter 5 could apply also to gravel roads except that, because of the loss of visibility that drivers experience on driving into the dust cloud generated by the leading vehicle, the passing manoeuvre may be initiated at a greater following distance than would be applied on a surfaced road. A further difference could be that following drivers would overtake at high speed differentials simply to ensure the shortest possible time spent in effect in driving blind.

Although this problem could be an interesting research project, there would be little point in undertaking it. As pointed out previously, there are so many variables in the model that, in spite of significant research into the passing model, any answer like the overtaking attributes of the 85th percentile driver will probably always have a question mark hanging over it. In the case of the gravel road, variables defining the surface of the road would have to be added to these, making any answer emerging from the research highly dubious. Probably the best approach would simply be to accept the values of passing sight distance offered in Chapter 5 and then to increase them by some or other arbitrary factor like, say, a further 20 per cent.

12.6 HORIZONTAL AND VERTICAL ALIGNMENT

As described in Chapter 6, the horizontal alignment comprises a series of tangents and circular curves. The two approaches to the determination of the horizontal alignment can be either by location of the points of intersection between successive tangents followed by fitting curves to the deviations between them or by locating the curves in the first instance and then joining them up by the insertion of tangents between them.

The latter is referred to as the curvilinear approach to alignment and can be done only on a survey plan in a drawing office. This is because, although the tangential approach is essentially a linear exercise, with the location of the points of intersection of two tangents being followed by selection of the radius of the curve between them, the curvilinear is very much a trial-and-error system with feedback loops. Locating the points of intersection between successive curves can be done in the field in a design-by-eye technique which, often, is applied to the design of low-volume roads because of the absence of adequate survey plans.

Selection of the length of tangents refers specifically to the maximum and minimum lengths, neither of which, preferably, should be exceeded. Very long tangents become boring and drivers could simply fall asleep while driving along them. At night, long tangents cause the dazzle created by approaching headlights to become a nuisance and, when the vehicles are getting close to each other, dazzle can become dangerous. This is because drivers then have a problem in seeing people at the roadside or cyclists who do not have reflectors on their bicycles. In Third World countries, the road reserve is often unfenced, partially for financial reasons but also so as not to impede the path of migrating wild life. At night, animals in the middle of the road, even if large, may not be seen in time to avoid them. Obviously, a rhinoceros or a hippopotamus possesses considerable stopping power and hitting one would probably have fatal consequences.

Short lengths of tangent make it difficult to achieve a smooth transition of the superelevation from a curve in one direction to a following curve in the opposite direction. Furthermore, curves in quick succession would, as a general rule, reduce the percentage of road on which passing opportunities exist. On the other hand, if the curves are short and the tangents long, the percentage of passing opportunities would increase.

It should be clear that the optimum horizontal alignment can be achieved only by a series of well-considered trade-offs between lengths of curves and tangents.

Consistency of design as discussed in Chapter 4 relates to the selection of radii of successive horizontal curves and applies equally to low-volume roads.

The application of superelevation to horizontal curves makes it possible to use shorter radius curves to achieve desired changes of direction. Values of superelevation of paved roads can range from crossfall, that is, when the camber of the outside lane is rotated to form a straight slope from shoulder to shoulder, to a prescribed maximum slope. Maximum slopes are usually 4, 6, 8 or 10 per cent depending on circumstances.

The lower rates of superelevation are normally applied to urban roads. This is because of spatial constraints attaching to the distance required to develop superelevation and also that it is difficult to provide direct property access across steeply superelevated cross-sections. Low values of superelevation are also required in areas prone to heavy rain, snow or icing conditions because vehicles travelling slowly around a steeply superelevated curve could slide down towards the inside shoulder.

The steeper values of superelevation apply to rural roads, and the highest values are normally associated with the highest design speeds. The limit of 10 per cent makes provision for vehicles that, for whatever reason, are not travelling at the design speed in negotiating the curve. The driver of a truck at crawl speed on a short radius curve would have to steer outwards to a significant extent to ensure that the curve is followed. The off-tracking of the rear wheels of a semitrailer on a short radius curve with a superelevation of 12 per cent is considerable. Trucks with high loads such as bales of lucerne have, at crawl speeds on a superelevation of 12 per cent, been observed to lose their loads.

On gravel roads, the maximum value of superelevation should be limited to 6 per cent or less. Water falling on the road surface would drain towards the lower shoulder and, at 6 per

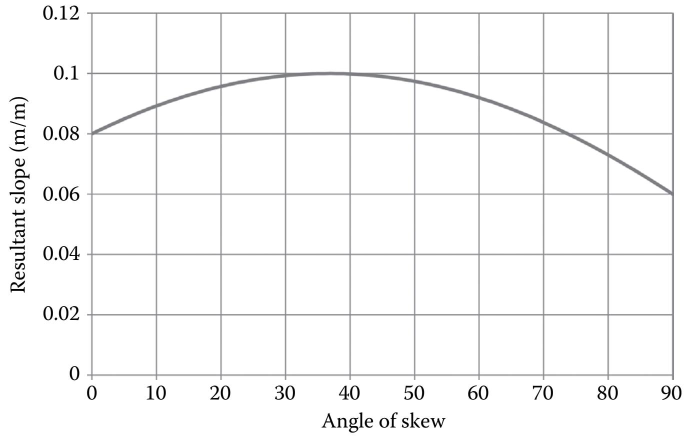


Figure 12.5 An example of magnitude of slope.

cent or more, there is a strong possibility of scouring of the road surface. Six per cent would thus be limited to areas where the road is substantially flat. As discussed in Section 12.8, the presence of a longitudinal gradient on a horizontal curve requires the solution of a vector triangle to establish the resultant of the vectors being the superelevation and the longitudinal gradient. Figure 12.5 illustrates that a superelevation of 6 per cent in combination with a longitudinal gradient of 8 per cent produces a resultant of 10 per cent at an angle of about 37° to the longitudinal direction of the road.

The required procedure for the design of curves on gravel roads is to select the value of e_{max} that is to be applied. In the interests of regional uniformity, this would normally be prescribed by the responsible Department of Transportation. The gradient of the road where the curve is to be located is then established. The relationship between crossfall and gradient shown in Section 12.8 is recast to be

$$S = \frac{1}{100}(c \sin\theta + g \cos\theta)$$

where

c = value of superelevation (%)
g = longitudinal gradient (%)
S = resultant slope (m/m)
θ = angle of resultant vector to road edge

The value of S is calculated for the range of θ from 0° to 90° similar to the typical case illustrated in Figure 12.5. The values of c shown on this curve can, for a range of skew angles, all result in the resultant vector having a slope of the selected value at various angles of skew. Reference to the tables of superelevation for above minimum radii of curvature will show the choice of radius available to the designer.

12.7 CROSS-SECTION DESIGN

Much of what was discussed in Chapter 8 applies equally to low-volume surfaced and unsurfaced roads and is not repeated in this chapter, where discussion is limited to the differences

Table 12.6 Passenger car units of vehicles likely to use the shoulders

Design vehicle	*Passenger car equivalents*
Pedestrian	0.15
Bicycle	0.2
Motor cycle	0.25
Bicycle with trailer	0.35
Motorcycle taxi	0.4
Motorcycle with trailer	0.45
Animal-drawn cart	0.7
Bullock cart	2.0

between conventional and low-volume roads. The cross-section of paved roads draws a distinction between the travelled way and the shoulders in specifying cross-sectional widths. The passenger car units (PCUs) in Table 12.6 reflect nonmotorised and other vehicles likely to be found on the shoulders of a low-volume road. It should be noted that pedestrians have dimensions and performance rates and they also have a personal space that should not be invaded by others without permission. Pedestrians are, in all senses of the phrase, 'design vehicles' and thus have a PCU value.

The shoulder widths of paved low-volume roads are shown in Table 12.7.

In the case of unsurfaced roads, this distinction is inapplicable as there is no clear demarcation between the travelled way and the shoulders. However, the outer portion of the gravelled cross-section also serves the same function as the shoulder on a surfaced road. If it is decided that a gravel road should have a width of, say, 5 metres what is constructed should actually have an overall width of possibly 6 metres or a little more. This applies only when there isn't much traffic using the shoulder. In effect, it is a basic 'shoulder' width for unsurfaced roads. Increased shoulder widths for unpaved shoulders in the presence of significant volumes of traffic are provided in Table 12.8.

Table 12.7 Shoulder widths for paved low-volume roads

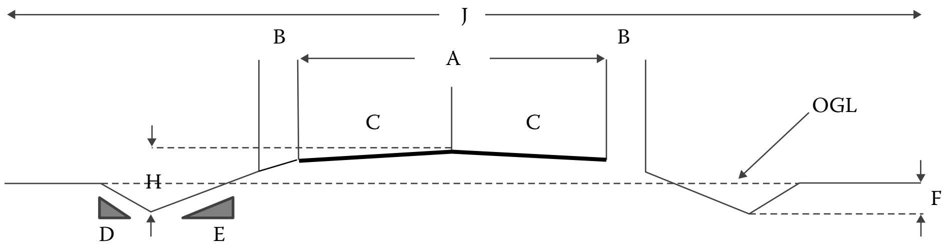
	Shoulder width for various ADTs (vehicles per day)			
Terrain	*150–400*	*75–150*	*25–75*	*<25*
Flat	1.5	1.25	1.0	Paving would not be applied to roads with ADTs of fewer than 25 vehicles per day.
Rolling	1.5	1.25	1.0	
Mountainous	1.0	0.5	0.5	
Escarpment	1.0	0.5	0.5	
Large population	3.5	3.5	2.5	
High PCUs	+2.0	+1.5	+1.25	

Note: ADT, average daily traffic; PCU, passenger car unit.

Table 12.8 Increased 'shoulder' widths for unpaved roads

Average daily traffic	*High population*	*High passenger car units*
150–400	+3.5	+2.0
75–150	+3.5	+1.5
25–75	+2.5	+1.3
<25	+2.0	N/A

Table 12.9 Typical cross-section for an unpaved road in flat terrain



		Value for ADTs (vehicles per day) of			
Label	*Element*	*<25*	*25–75*	*75–150*	*150–400*
A	Carriageway width (m)	3.30	5.00	5.50	6.00
B	Shoulder width (m)	0.50	0.50	0.75	0.75
C	Minimum crossfall/camber (%)	4	4	4	4
D	Backslope of ditch	Varies according to *in situ* material			
E	Sideslope of ditch	1*V*:4*H*			
F	Maximum depth of side drain (m)	0.50			
H	Depth below centreline (m)	0.35	0.35	0.50	0.50
J	Clear zone (m)	15.00	20.00	20.00	20.00

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note: Drains maintained by a grader are V-shaped. Drains maintained by labour-based methods are trapezoidal in cross-section with a minimum invert width of 400 millimetre and sides with a slope of 1*V*:2*H*. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

Tables 12.9 to 12.18 were taken from the South Sudan *Low-Volume Design Manual* and modified in light of previous experience.

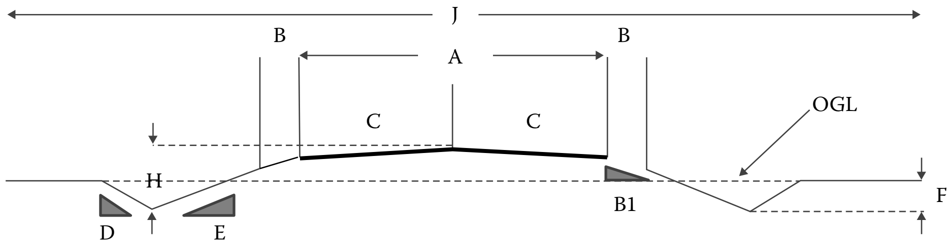
12.8 SINGLE-LANE ROADS

A single-lane road would normally have a width of 3 to 3.5 metres, plus the allowance for an equivalent shoulder discussed previously. If the intention is that a road should have single-lane operation, it should not be wider than 4 to 4.5 metres because the allowance for the shoulders would increase the overall width to 5.5 metres or more. There is a risk that motorists may perceive this as being a two-lane road and drive accordingly. Apart from the safety hazard that this presents, driving on what is intended to function as a shoulder could result in it being grossly deformed, placing the integrity of the travelled portion of the cross-section at risk.

Single-lane roads require the provision of passing bays. These should be spaced at between 300 and 500 metres apart and built at the most economic places in preference to the adoption of a fixed spacing.

Most countries specify a bus width of 2.6 metres, suggesting that passing bays should have an absolute minimum width of 6.0 metres. This would allow only about 300 millimetres between the outside edges of each bus and the edge of the road and also between buses. A width of 7.0 metres is considered to be the desirable minimum width of road at a passing bay. Ideally, passing bays should be intervisible. Considering the lengths of design vehicles as shown in Chapter 5, an overall length of 25 metres would be adequate for passing bays. This would be sufficient to accommodate a WB22 vehicle or anything up to four passenger cars.

Table 12.10 Typical cross-section for a paved road in flat terrain



Label	Element	Value (vehicles per day) for ADT of 25–75	75–150	150–400
A	Carriageway width (m)	3.30	6.00	6.5
B	Shoulder width (m)	1.00	1.25	1.50
B1	Shoulder crossfall (%)	6.00		
C	Minimum crossfall/camber (%)	3.00	3.00	3.00
D	Backslope of ditch	Varies according to *in situ* material		
E	Sideslope of ditch	1*V*:4.0*H*		
F	Depth of side ditch (m)	0.35		
H	Depth below centreline (m)	0.50		
J	Cleared zone (m)	20.00		

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note: Drains maintained by a grader are V-shaped. Drains maintained by labour-based methods are trapezoidal in cross-section with a minimum invert width of 400 millimetres and sides with a slope of 1*V*:2*H*. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

12.9 DRAINAGE

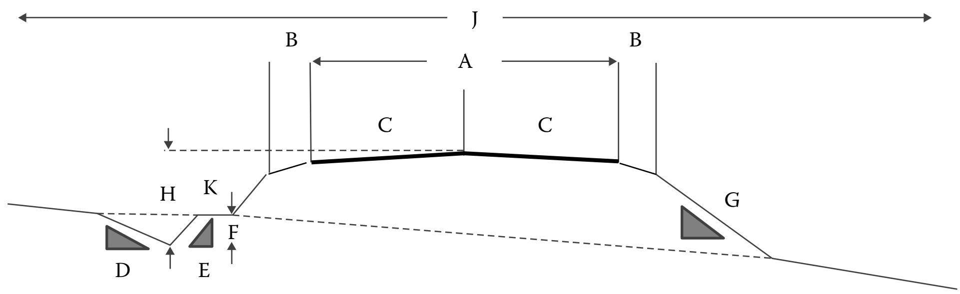
Designers versed in the design of surfaced road layouts have a tendency to consider their designs a snapshot as in a photograph, with little consideration for wear and tear. Designing gravel roads must take into account that the roads can change dramatically even after one rainstorm. One pass by a grader can substantially change whatever superelevation had been applied to the gravel surface.

As a consequence the environmental factors and traffic regime must be a very important consideration. Avoiding rutting by increasing the crossfall to slopes that would not be considered for surfaced pavements is one possibility. Although this is strongly counterintuitive, swiftly removing the water from the road means that there isn't much time for it to penetrate and soften the wearing course, after which scouring is inevitable. Frequent mitre drains and cross-drains help with runoff control, as will the selection of base materials that are cohesive.

Raising the level of the road intermittently to make provision for culverts in the form of nominal drainage, that is, where defined water courses do not exist and flow is principally overland, is not really practical because water could simply flow around the culverts. Where watercourses do exist, a problem could arise in the case of shallow watercourses. Local excavation in the immediate vicinity of a culvert would soon silt up. Furthermore, the geometry of the road over the culverts would have to be generous where speeds are high to prevent vehicles becoming airborne.

First World storm water drainage is based on the prevention of flooding of the road. This is achieved by the provision of an adequate height of earthworks and suitably sized drainage structures. In short, storm water is kept away from the road surface, that is, upstream

Table 12.11 Typical cross-section for an unpaved road in rolling terrain



Label	Element	Value for ADT (vehicles per day) of <25	25–75	75–150	150–400
A	Carriageway width (m)	3.30	5.00	5.50	6.00
B	Shoulder width (m)	0.50	0.60	0.75	0.75
C	Minimum crossfall/camber (%)	4.00			
D	Backslope of ditch	Varies according to *in situ* material			
E	Sideslope of ditch	1*V*:4.0*H*			
F	Maximum depth of side drain (m)	0.50			
G	Sideslope	1.00*V*:1.50*H*			
H	Depth below centreline (m)	0.50			
J	Clear zone (m)	15.00	20.00	20.00	20.00
K	Sideslope	1.00*V*:1.50*H*			

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note: Drains maintained by a grader are V-shaped. Drains maintained by labour-based methods are trapezoidal in cross-section with a minimum invert width of 400 millimetres and sides with a slope of 1*V*:2*H*. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

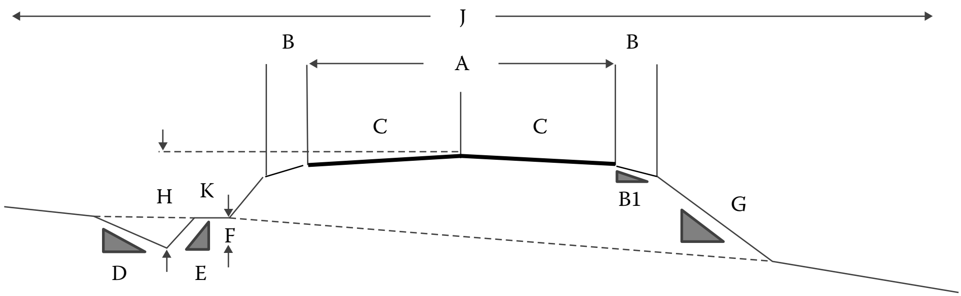
protection of the cross-section. Third World storm water drainage must, because of the absence of height of the gradeline, be designed to allow water to flow across the road in a sheet and to provide protection against cutting of the embankment by water flowing down the fill slope where this exists, that is, downstream protection of the cross-section. Protection is normally in the form of stone pitching or very low gabions.

The crossing of perennial streams is normally by low-level causeways, known in Britain as 'Irish bridges'. A multi-barrel causeway is illustrated in Figure 12.6.

Where the flow is more than the perennial flow as a result of rain upstream of the river crossing, the water simply flows across the top of the causeway. It is customary to place markers on either end of the causeway to indicate the depth of flow. A passenger car will float in standing water that is 600 millimetres deep. If the water is flowing, a depth of 300 millimetres is enough to move an average sized passenger car and 150 millimetres of fast flowing water is a sufficient depth to knock a pedestrian off his or her feet.

It has been said that the three D's of design are 'drainage, drainage, drainage'. This applies to surfaced roads mainly as a road safety measure where the surface should not include ponding water, which could result in vehicles hydroplaning out of control. On gravel roads, in addition to the avoidance of ponding, it is necessary to ensure that the surface does not scour to total destruction, as could result from one heavy downpour. A fine line must be drawn between having a gradient so flat that silt deposition occurs and one so steep that

Table 12.12 Typical cross-section for a paved road in rolling terrain



Label	Element	Value for ADT (vehicles per day) of		
		25–75	*75–150*	*150–400*
A	Carriageway width (m)	3.30	6.00	6.50
B	Shoulder width (m)	1.00	1.25	1.50
B1	Shoulder crossfall (%)		6.00	
C	Crossfall/camber (%)		3.00	
D	Backslope of ditch		Varies according to *in situ* material	
E	Sideslope of ditch		1.00*V*:4.00*H*	
F	Depth of side ditch (m)		0.35	
G	Sideslope		1.00*V*:1.50*H*	
H	Depth below centreline (m)		0.50	
J	Cleared zone (m)		20.00	
K	Sideslope		1.00*V*:1.50*H*	

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual.* Juba, 2013; modified in light of previous experience.

Note: Drains maintained by a grader are V-shaped. Drains maintained by labour-based methods are trapezoidal in cross-section with a minimum invert width of 400 millimetres and sides with a slope of 1*V*:2*H*. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

the surface can scour. Silt deposition could occur at flow speeds less than about 0.6 m/s and scour at flow speeds faster than as shown in Table 12.19.

Flow speed is dependent on the slope across which flow is taking place. The slope is the resultant of two vectors – the camber, crossfall or superelevation and the longitudinal gradient of the road. The resultant slope is given by

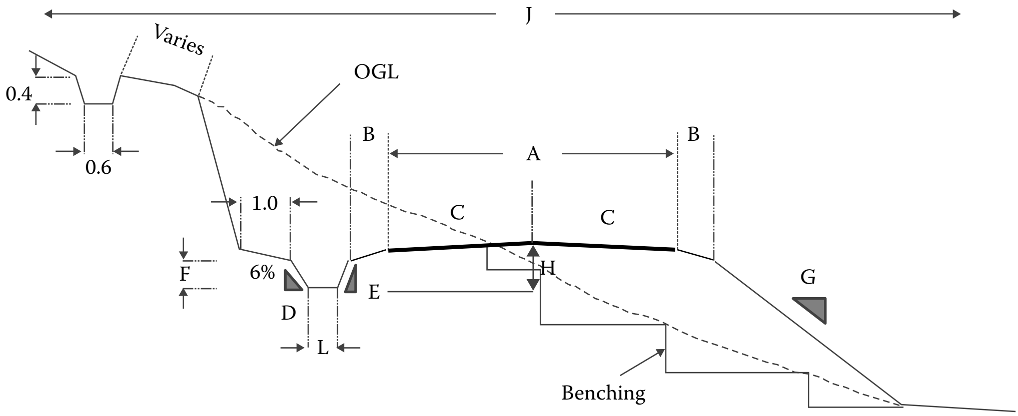
$$S = \sqrt{n_1^2 + n_2^2}$$

where

S = the resultant slope in m/m
n_1 = crossfall or superelevation (%)
n_2 = longitudinal gradient (%)

Figure 12.7 illustrates the case where the superelevation on a curve is 6 per cent and the gradient of the road is 8 per cent. If the direction of the slope is 0°, the magnitude of the slope would obviously be the same as the gradient of the road, in this case 8 per cent. If the direction is 90°, the magnitude of the slope would the same as the superelevation, in this case 6 per cent. Between these two values, the resultant of the two vectors would be higher than either of them, peaking at a value of 10 per cent when the angle of skew is at about 37°.

Table 12.13 Typical cross-section for an unpaved road in mountainous terrain

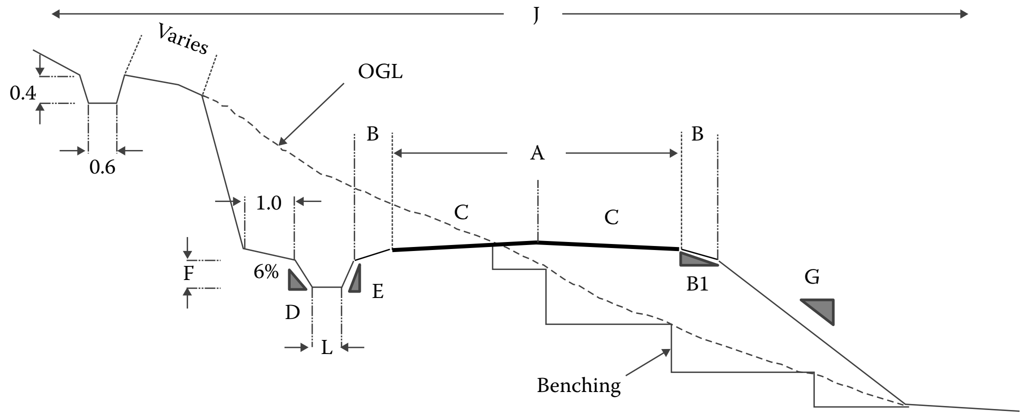


Label	Element	Value for ADT (vehicles per day) of <25	25–75	75–150	150–400
A	Carriageway width (m)	3.30	5.00	5.50	6.00
B	Shoulder width (m)	0.60			
C	Minimum crossfall/camber (%)	4.00			
D	Backslope of ditch	Varies according to *in situ* material			
E	Sideslope of ditch	1.00*V*:4.00*H*			
F	Depth of side drain (m)	0.35			
G	Sideslope	1.00*V*:1.50*H*			
H	Depth below centreline (m)	0.50			
J	Clear zone (m)	15.00	20.00	20.00	20.00
L	Side drain width (m)	0.50 minimum			

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note: Drains maintained by a grader are V-shaped. Drains maintained by labour-based methods are trapezoidal in cross-section with a minimum invert width of 400 millimetres and sides with a slope of 1*V*:2*H*. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

Table 12.14 Typical cross-section for a paved road in mountainous terrain

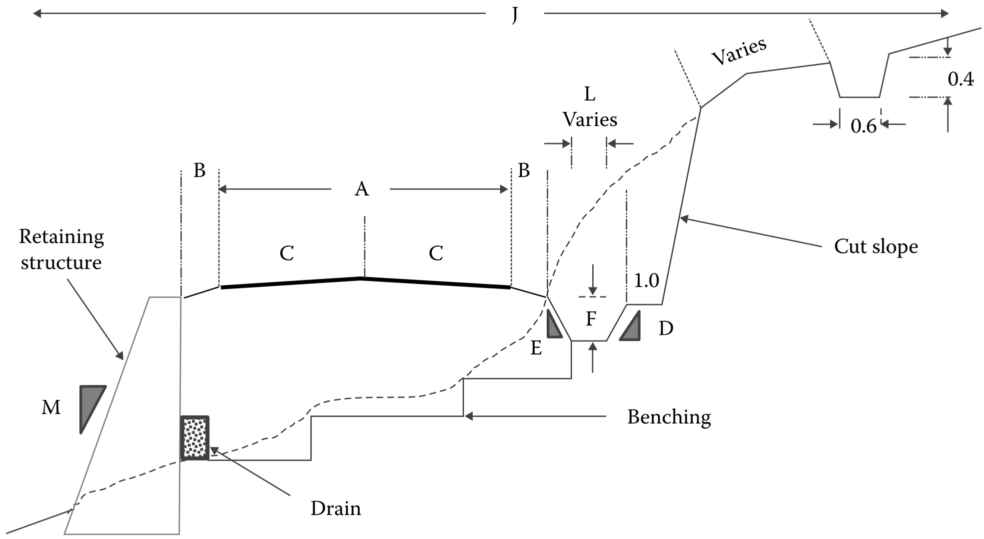


Label	*Element*	*Value for ADT (vehicles per day) of* 25–75	75–150	150–400
A	Carriageway width (m)	3.30	6.00	6.50
B	Shoulder width (m)		1.00	
B1	Shoulder crossfall (%)		3.00	
C	Crossfall/camber (%)		3.00	
D	Backslope of ditch		Varies according to *in situ* material	
E	Sideslope of ditch		1*V*:4*H*	
F	Maximum depth of side ditch (m)		0.50	
G	Sideslope		1*V*:1.5*H* (steepest)	
J	Cleared zone (m)		20.00	
L	Minimum side drain width (m)		0.50	

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual.* Juba, 2013; modified in light of previous experience.

Note: Drains maintained by a grader are V-shaped. Drains maintained by labour-based methods are trapezoidal in cross-section with a minimum invert width of 400 millimetres and sides with a slope of 1*V*:2*H*. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

Table 12.15 Typical cross-section for an unpaved road in escarpment terrain

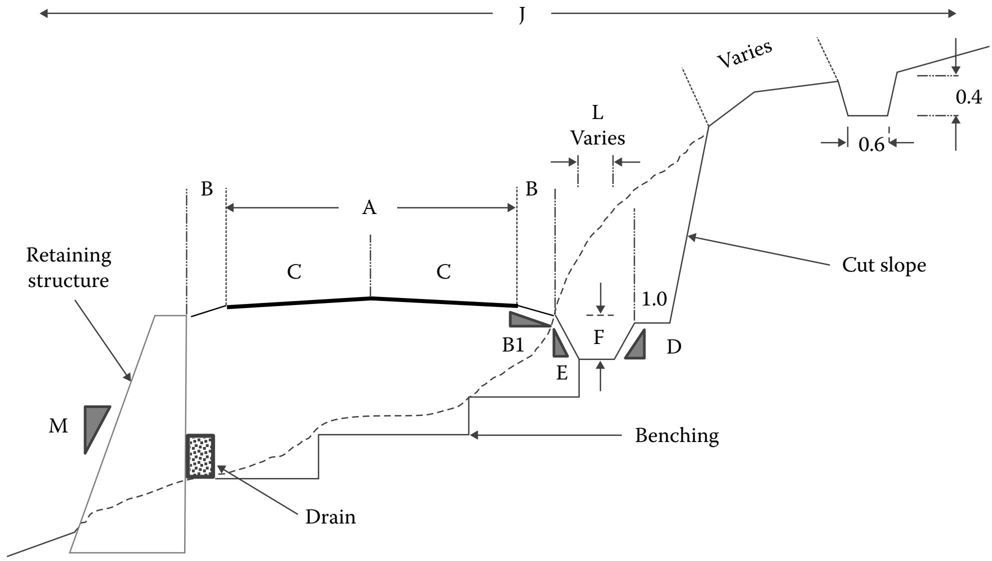


Label	*Element*	*Value for ADT (vehicles per day) of*			
		<25	*25–75*	*75–150*	*150–400*
A	Carriageway width (m)	3.30	5.00	5.50	6.00
B	Shoulder width (m)	0.60			
C	Minimum crossfall/camber (%)	4.00			
D	Backslope of ditch	Varies according to *in situ* material			
E	Sideslope of ditch	1.00*V*:4.00*H*			
F	Depth of side drain (m)	0.35			
J	Clear zone (m)	15.00	20.00	20.00	20.00
L	Minimum side drain width (m)	0.50			
M	Slope of retaining structure	Dependent on structural design			

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note: The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

Table 12.16 Typical cross-section for a paved road in escarpment terrain

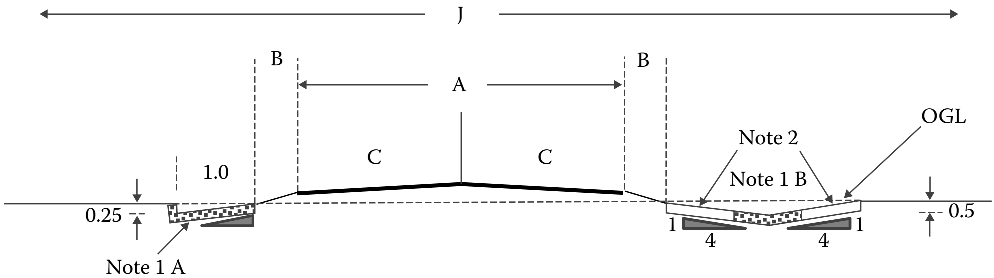


Label	*Element*	*Value for ADT (vehicles per day) of*		
		25–75	*75–150*	*150–400*
A	Carriageway width (m)	3.3	5.5	6.5
B	Shoulder width (m)	0.50	0.50	1.00
B1	Shoulder crossfall (%)		3.00	
C	Minimum crossfall/camber (%)		3.00	
D	Backslope of ditch		Depends on *in situ* materials	
E	Sideslope of ditch		1.00*V*:4.00*H*	
F	Depth of side ditch (m)		0.50	
J	Cleared zone (m)		20.00	
L	Side drain width		Minimum 0.50	
M	Slope of retaining structure		Dependent on structural design	

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note: Drains maintained by a grader are V-shaped. Drains maintained by labour-based methods are trapezoidal in cross-section with a minimum invert width of 400 millimetres and sides with a slope of 1*V*:2*H*. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

Table 12.17 Typical cross-section for an unpaved road in populated areas



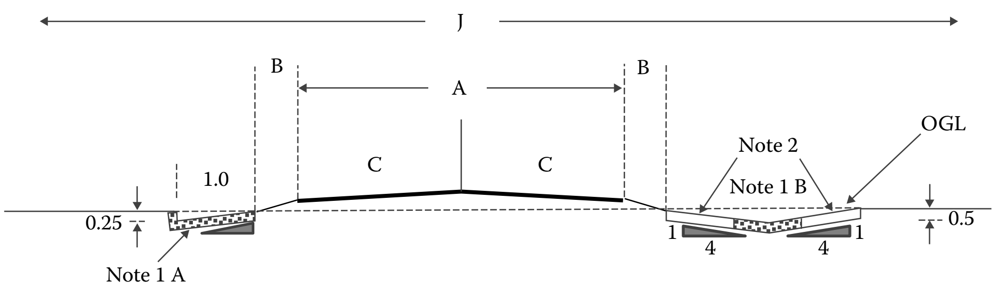
Label	Element	Value for ADT (vehicles per day) of			
		<25	*25–75*	*75–150*	*150–400*
A	Carriageway width (m)	3.30	5.00	5.50	6.00
B	Shoulder width (m)	0.50	0.60	0.75	0.75
C	Minimum crossfall/camber (%)	4.00 Dependent on structural design			
J	Clear zone (m)	15.00	20.00	20.00	20.00

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note:

(1) Open channel type A: 25 cm thick mortared stone pitching. Open channel type B: 25 cm thick concrete. (2) Wearing course. The choice of open channel is dependent on local conditions. Lined channels should be provided only where the maintenance of the road surface and camber at original levels is guaranteed. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

Table 12.18 Typical cross-section for a paved road in populated areas



Label	Element	Value for ADT (vehicles per day) of		
		25–75	*75–150*	*150–400*
A	Carriageway width (m)	5.00	5.50	6.00
B	Shoulder width (m)	0.60	0.75	0.75
C	Minimum crossfall/camber (%)		4.00	
J	Clear zone (m)		20.00	

Source: Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual*. Juba, 2013; modified in light of previous experience.

Note:

(1) Open channel type A: 25 cm thick mortared stone pitching. Open channel type B: 25 cm thick concrete. (2) The choice of open channel is dependent on local conditions. Surfacing of the shoulder is recommended. The clear zone can be and often is narrower than the total width of the road reserve. ADT, average daily traffic.

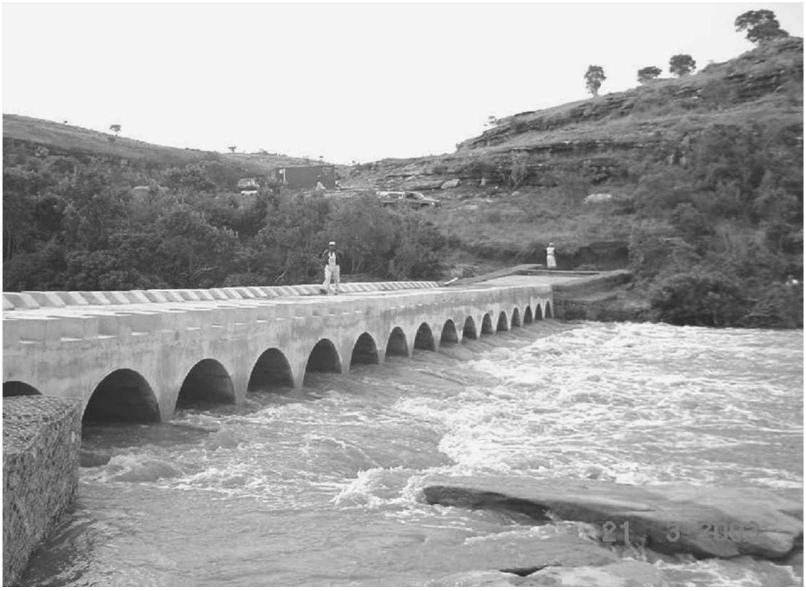


Figure 12.6 Multi-barrel causeway. (Photo: RC Burrell 2003.)

Table 12.19 Scour velocities for various materials

Material	*Maximum permissible flow speed (m/s)*
Fine sand	0.6
Loam	0.9
Clay	1.2
Gravel	1.5
Soft shale	1.8
Hard shale	2.4
Dolerite	4.5

Source: Burrell RC, Mitchell MF and Wolhuter KM. *Geometric design guidelines*. South African National Road Agency Limited (SANRAL), Pretoria, 2002.

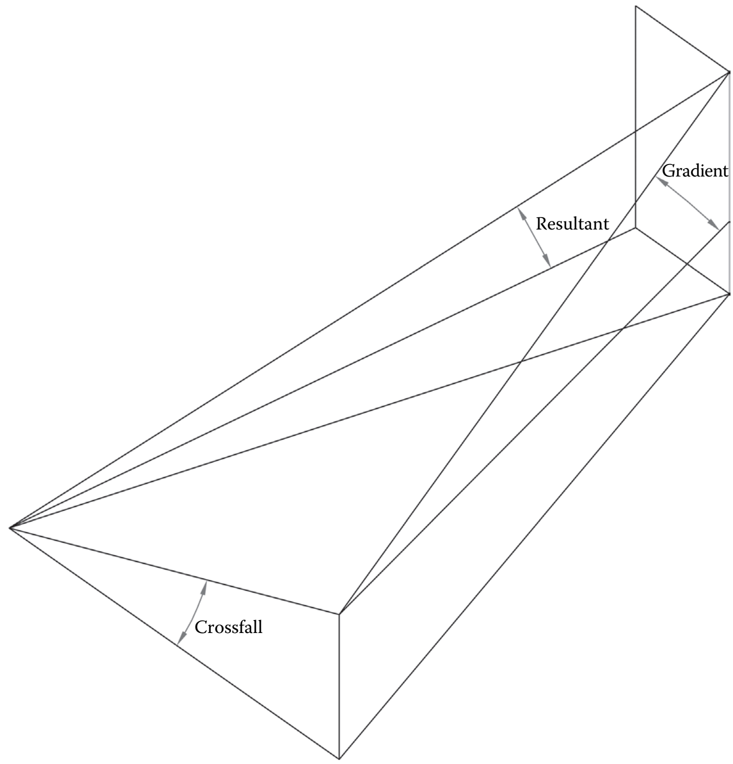


Figure 12.7 Slope vectors.

Chapter 13

Road safety

CONTENTS

13.1 INTRODUCTION

The ‘Safe System’ approach has generally been adopted worldwide. It is based on the Swedish ‘Vision Zero’ and the Dutch ‘Sustainable Safety’ and is built on the ethical basis that no one should be seriously injured or killed in his or her use of the road. Road authorities should thus strive to eliminate fatal and serious injury crashes by adopting a holistic systems approach. This system should recognise that people can and do make mistakes and that the human body is frail and can absorb only a limited amount of energy before it is seriously hurt.

The Safe System has four cornerstones (Department of Infrastructure Energy and Resources, 2006):

- Safe vehicles
- Safe road user behaviour on the road
- Safe speeds
- Safe roads and roadsides

Features that may not directly constitute geometric design but nevertheless impact on it are discussed in this chapter. In spite of efforts to make a road a safe as possible, vehicles will still leave the road in an unintended fashion. This could be the result of a collision with another vehicle or of driver error. The latter includes falling asleep at the wheel, driving too fast for the circumstances or substance abuse before driving. The last-mentioned features in far too many crashes or run-off-the-road (ROR) incidents. Safety-related organisations often quote percentages in the high 70s for cases in which substance abuse is the proximate cause of a crash. For whatever reason, vehicles could roll down fill slopes or hit immovable objects or cut slopes, other vehicles or people in the road reserve.

Geometric designers have a duty of care for all users of the roads they design. Figure 13.1 is a Venn diagram suggesting that human factors are a contributory factor to crashes to the extent of 93 per cent compared to the 34 per cent of road environmental factors. This does not mean that designers can shrug their shoulders and 'blame it all on the driver'.

As stated previously, the crash black spot is a well-known phenomenon. Crashes are generally randomly distributed along a road but there are always locations where a greater than the average percentage of crashes occurs along the length of the road. These are places either where drivers seemingly misread the information the geometric designer has provided or where the information provided is misleading or incorrectly placed. This could be because of information overload such as at a busy intersection or an optical illusion whereby a road on a curve actually appears to be straight on the basis of an unfortunate alignment of utility poles or chevrons. Foreshortening can cause a crest or sag curve to make a horizontal curve

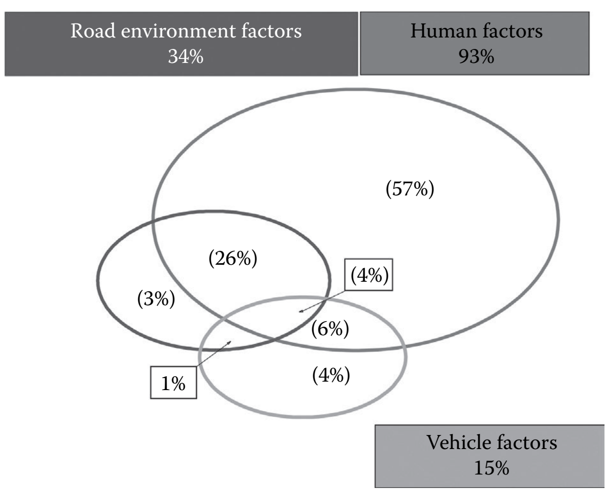


Figure 13.1 Factors contributing to crashes. (From Road Traffic Management Corporation. *South African road safety audit manual.* Pretoria, 2012.)

appear to have a larger or smaller radius than it actually does have. In consequence, a driver may enter a curve travelling too fast. Alternatively, because a curve radius appears smaller than it really is, a driver may slow down, perhaps sharply, surprising a following driver by the unexpected manoeuvre.

It is necessary to ensure that black spots, which are probably largely responsible for the 26 per cent overlap between human factors and road environment factors, do not occur. Furthermore, designers should seek to reduce the severity of the consequences of crashes that do occur. In the exercise of the designers' duty of care, it is necessary to evaluate the entire length of the project to determine where crash sites are likely to occur. This evaluation is referred to as a road safety audit and is the starting point in any safety-related exercise.

With possible crash sites identified, it is possible to plan, design and build remedial measures aimed at minimising the likelihood of such occurrences. Minimising their consequences where they do occur is also dealt with. The safety audit is a means of analysing the likelihood of crashes along the length of a road.

The application of remedial measures then follows the identification of the problem sites. Where there is a risk of vehicles leaving the road, three possibilities can be explored. First, cut or fill slopes, which are sufficiently gentle to ensure that they can be negotiated in safety, can be provided. Second, clear zones free of obstacles such as trees and structures that can cause either sharp deceleration or vehicles becoming airborne need to be created. Third, vehicles can be prevented from leaving the road altogether by the provision of safety barriers. These preventative measures are discussed.

Furthermore, the quality of life of the communities living adjacent to the road is also addressed. Commuters often find that their logical route between home and work is unacceptably congested. They then seek alternative routes usually along local streets broadly parallel to the arterial intended for commuter travel. This 'rat-running' is inimical to the quality of life of local residents. Noise and air pollution are known health hazards. Children playing in their local streets would be at risk from passenger cars travelling at high speeds on streets not intended for them. Chance meetings on the street by friends and neighbours are an important component of the quality of local life. This would obviously not take place on a street having the characteristics of a race track.

Traffic calming can make potential rat running routes less attractive than the intended route. It is actually a retrofit application of a whole panoply of measures to streets that had been laid out on the traditional grid pattern. Traffic calming is a two-pronged approach to the problem. It seeks to reduce the volume of traffic passing through a residential area and also to reduce the speed of traffic in the residential area. These measures are discussed in this chapter.

In the case of new developments, traffic calming can be built in to ensure that no road is a shortcut to anywhere. Town planning measures such as cell development, where an area is a complete microcosm with residential land uses being underpinned by local shops, schools and job opportunities and completely surrounded by higher order roads, can eliminate rat-running.

13.2 ROAD SAFETY AUDITS AND ASSESSMENTS

13.2.1 Introduction

The philosophy behind road safety audits (RSAs) originated by the Kent County Council in the United Kingdom during the 1980s as an independent checking of roads with the goal of improving the operational safety of projects. They were introduced in Australia and New Zealand in 1990 and have since become standard practice in many countries around the world (Schermers and Kraay, 1999). The United Kingdom, Australia, and New Zealand have since become leaders in refining and advancing the state of the practice.

The RSA approach departs from the traditional or reactive remedial treatment of high-accident locations by proactively identifying safety deficiencies during the design and building processes. They can also be conducted on existing roads when they are sometimes referred to as road safety assessments. Where the audit is applied to an entire road network, it is referred to as a road safety engineering assessment.

The RSA is defined as a *formal* examination of a future road or traffic project or any other project where interaction of the road with road users takes place, by an *independent*, *qualified* and *expert* examination team reporting on the crash potential and safety performance of the project. The keywords are that the examination team should be fully independent, qualified and expert in the conducting of safety audits.

The road safety engineering assessment establishes the road safety status of an entire network using a limited set of predefined key indicators. The function of the engineering assessment is to provide a list of prioritised locations that should be further investigated.

RSAs were introduced in the United States in 1996 as a result of a Federal Highway Administration (FHWA)-sponsored scanning tour of Australia and New Zealand. In 2004, RSAs were still considered to be in their infancy in the United States (Wilson and Lipinski, 2004). The suggestion was offered that, to advance and expand the application of the concept and to enhance safety benefits, training programmes should be launched to introduce more state Department of Transportation personnel to road safety audits practices and how they can be applied. Furthermore, it was suggested by Wilson and Lipinski that 'a compendium of best practices could be developed and disseminated to state Departments of Transportation, cities, and local road agencies'.

In general, to be safe, a road should (Road Traffic Management Corporation, 2012).

- Warn road users of any possible hazards, that is, there should be no surprises
- Inform road users of the type of unexpected conditions that are likely to be encountered
- Guide road users through sections with sometimes unexpected conditions
- Control and guide road users through conflict points or areas of conflict
- Forgive errant vehicles and the behaviour of road users involved

The broad intention of a road safety audit is to identify areas along a road where these requirements are not met. They thus seek, before the road is opened to traffic, to enhance road safety by reducing the likelihood of crashes occurring.

13.2.2 The benefits derived from RSAs

With regard to the implementation of recommendations for amendments during the design stage, it was found (Road Traffic Management Corporation, 2012) that

- The B/C (benefit/cost) ratio for individual audits ranged from 3:1 to 242:1.
- The B/C ratio for individual recommendations within a single audit ranged from 0.06:1 to 2600:1.
- More than 90 per cent of all implemented recommendations within the design stage audits had B/C ratios greater that 1.0:1.
- About 75 per cent of all implemented recommendations had B/C ratios greater than 10:1.

- The majority of design audit recommendations required responses costing less than \$1000 and, of these, the B/C ratio was greater than 10:1 in 85 per cent of the cases.

Implementation of the recommendations for reconstruction of existing hazardous areas could be expensive. However, in the case of audits on existing roads it was found (Road Traffic Management Corporation, 2012) that

- Implementing the proposed action on a range of road safety audits had a B/C ratio of between 2.4:1 and 84:1.
- The B/C ratio of individual proposed actions within road safety audits ranged between 0.003:1 and 460:1.
- More than 78 per cent of all proposed actions had B/C ratios greater than 1.0:1 and approximately 47 per cent had B/C ratios greater than 5.0:1.
- Approximately 95 per cent of all proposed actions with a cost of less than \$1000 had B/C ratios greater than 1.00:1.

European studies suggested that the cost of a road safety audit was typically less than 1 per cent of the construction cost of the project. The Surrey County Council compared audited schemes with similar nonaudited schemes and suggested that an audit could save at least one casualty per year. Denmark indicated that the first year rate of return ranged between 149 per cent and 600 per cent, implying that the cost of implementation of the remedial measures would be recovered in significantly less than 1 year (Road Traffic Management Corporation, 2012).

It must be stressed that the numbers quoted in the preceding paragraph are speculative. However, they illustrate that all roads agencies should include road safety audits as a normal part of their design portfolios. Furthermore, it should also be understood that the recommendations referred to earlier would not normally emanate from the road safety audit team. This is not their function. Their brief is purely to identify potential trouble spots. If they were to do any more than this they would tacitly become involved in the design process and thus lose their independence.

13.2.3 The stages of an RSA

As stated previously, an RSA is conducted prior to the opening of a road to traffic. It can thus happen at any time during the design process, as illustrated in Figure 13.2. In general, the earlier the RSA is conducted the greater are the rewards for conducting it. This arises from the extent of the cost of the abortive expenditure incurred prior to execution and implementation of the road safety audit. Clearly, if the audit is carried out after completion of construction but prior to opening the road to traffic, considerable cost could be incurred in execution of the recommended remedial works.

Road design normally proceeds in discrete stages and the appropriate time to carry out an audit is between stages. These stages are

- Basic planning
- Draft design
- Detail design
- Work zone traffic management
- Preopening

As audits proceed through the various stages, the level of detail moves from the broad brush towards the microscopic.

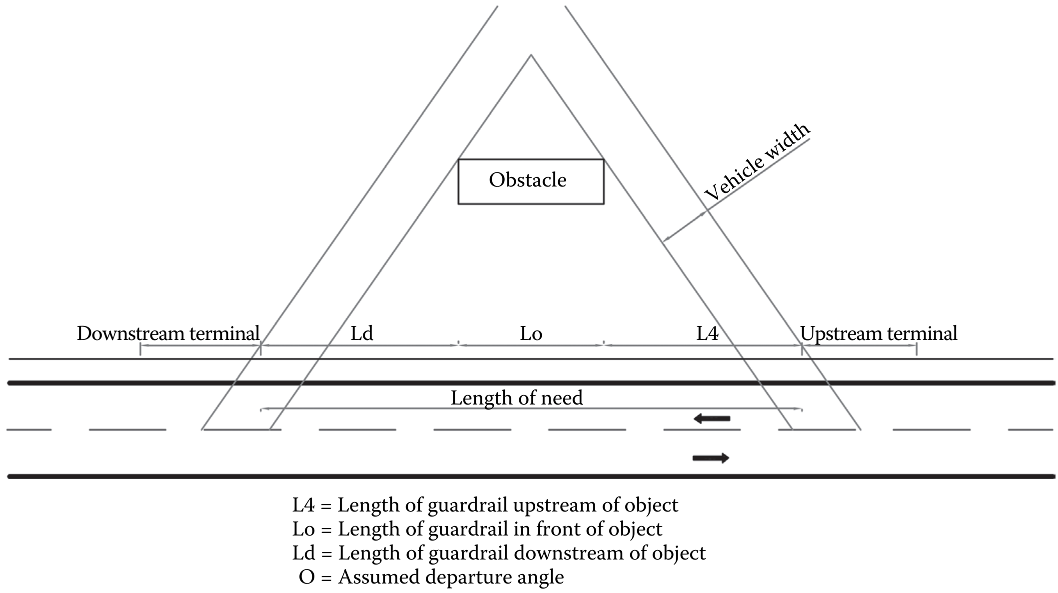


Figure 13.2 Length of need.

13.2.3.1 Basic planning

Basic planning is focussed on comparison of various corridors, typically 500 to 1000 metres wide and usually to a scale of 1:50,000. Comparison is on a basis of economic analysis and utility analysis as described in Chapter 16 with due regard for environmental issues and the needs of local communities affected by provision of the road – in short, a feasibility analysis of competing routes. As a general rule, public consultation would be carried out as part of the basic planning process with another round of public consultation carried out at the end of the detailed planning process. The latter is largely intended to reassure the public that its requirements as identified in the first round have, to some greater or lesser extent, in fact been addressed.

Detailed survey to a scale of, say, 1:10,000, often carried out by photogrammetric means, is then applied to the selected corridor. Route location is carried out within the selected corridor on the basis of the detailed survey and is preceded by selection of the design speed and limiting values of design guidelines such as minimum radius of curvature and *K*-values of vertical curves appropriate to the selected design speed.

At this stage, a road safety audit is aimed at identifying safety problems regarding the scope of the project. The choice of route including issues of route continuity and the appropriateness of the selected design speed and associated guideline values would be assessed. The impact of the selected route on the immediate environment and adjacent road network including the provision of accesses and intersections or interchanges is also assessed, as are various alternative design features of the project.

13.2.3.2 Draft design

A safety audit carried out at the end of the draft design phase would address the actual values of the design guidelines adopted along the length of the road for the design. It would focus on issues such as the horizontal and vertical alignment and their impact on available sight distances with regard to stopping, passing, intersections, and so forth. The appropriateness of the cross-section, including the selection of the various elements of the cross-section, their location in the cross-section and their dimensions would also be reviewed.

Where the designer had requested the application of a waiver with respect to a design exception or variance as discussed in Chapter 4, the impact of the exception or variance and the precise terms of the issued waiver on the safety of the project would need to be assessed.

As the majority of crashes occur at intersections, the access management applied to the project with regard to direct property access from adjacent properties as well as the layout of intersections and the configuration of interchanges would be closely scrutinised.

If the draft design audit is the first one to be carried out, it should include consideration of the issues that would have been dealt with in a stage 1 audit.

13.2.3.3 Detail design

The next phase of design after completion of the detailed design phase is concerned with the preparation of contract documents. The end of the detailed design phase is thus the last opportunity to influence the design before initiation of construction and is a final review of all the drawings that comprise the extent of the project. It would thus check that the points raised in previous audits have been satisfactorily dealt with or, if such audits were not carried out, their content would now all be included in the detailed design phase audit.

This audit would be concerned with the extent to which the road is self-explaining, that is, traffic arrangements and the transfer of information to the various user groups. Attention would be on

- Road traffic signs and markings
- Lighting
- Roadside safety as discussed in Section 13.3
- The needs of freight vehicles and public transport
- The requirements of vulnerable road users such as cyclists and pedestrians
- Design for access management including the detail of intersection and interchange layout

Drainage as discussed in Chapter 24 would be addressed in depth in this safety audit, as inadequate provision for drainage could result in the premature destruction of the road.

Furthermore, the designer should have made adequate provision for the accommodation of traffic during construction of the project. Arrangements for traffic management and their definition on the contract drawings would be assessed in this safety audit.

Information should have been provided to the contractor regarding the need for and extent of site trimming required. This includes the tidying of spoil heaps and borrow pits, dressing of cut and fill slopes to the specified batters and side slopes and general landscaping insofar as it has been damaged during the construction process. The extent and clarity of this information, whether by drawing or text or both, need to be audited.

If stage 1 and 2 audits had not been carried out, the matters discussed under those headings above would need to be incorporated in this stage 3 audit.

13.2.3.4 Work zone traffic management

Work zones comprise a highly hazardous mix of passing traffic, construction plant and labourers and other workers on foot. It follows that, simultaneously with the execution of the detailed design phase, a traffic management plan should have been developed. This would include the location of detours, with their geometry in terms of design speed and cross-section and signing. Traffic control measures by means of road signs and markings, signalisation and/or flagmen would need to be specified. The provision of safe access to adjacent properties would also need to be addressed.

Once this plan has been audited for adequacy and added to the contract drawings, the contractor would be required to prepare detailed traffic management plans for each proposed detour for scrutiny and approval by the engineer. These would be based on the initial contract drawings, taking cognisance of changed conditions in the field.

13.2.3.5 Preopening

By the time this stage has been reached, opportunities to modify the design at lowest cost have been exhausted. Design errors are now set in concrete. But this is the last opportunity for the identification of safety issues as they would be experienced by road users. The site should therefore by visited by the audit team both during daylight hours and after dark. The latter is necessary to ensure that the situation where drivers are constrained in their perception of the road can still be safe.

Attention would be paid to intersections and the manner in which the project is tied to the existing road network.

The ability to make changes to the design to resolve outstanding safety issues is limited at this stage and it is possible that mitigating measures may also be limited.

This audit is aimed at ensuring the following:

- Sufficient provision has been made for the various types of road users of the project.
- Potential roadside hazards have been adequately addressed.
- There are no side-effects emanating from differences between the design and the as-built drawings.
- All road signs and markings and lighting have been provided as defined in the detailed design drawings.
- All temporary works, markings and construction equipment that may constitute a hazard or cause confusion have been removed.

If any issues raised in previous audits have not been adequately addressed, they should be reiterated in the preopening audit report.

13.2.3.6 Checklists

Many authorities have developed check lists as an aide memoire to auditors. The Austroads manual, *Guide to Road Safety* Part 6: *Road Safety Audit* (2009b) and the Road Traffic Management Corporation document, *South African Road Safety Audit Manual* (2012), are cases in point. The Federal Highway Administration *Road Safety Audit Manual* (2006) includes ‘prompt lists’ and discussion of when and how to use them in Part C: *Road Safety Audit Tools*. Many American state transportation departments have also issued audit manuals that include checklists for use by road safety auditors (California Department of Transport [Caltrans], 2012).

The general intention with checklists is that they should be consulted only after completion of the various audits. This is because they should not be perceived as being comprehensive even though every effort has been made to make them so. Furthermore, slavish adherence to checklists may result in some or other design flaw being overlooked. There is, however, no reason why the designers could not use these checklists as guides to their designs. In the same way that writers cannot edit their own language usage, geometric designers would not be able to assess the safety of their proposed designs impartially. It is also assumed that they would not be overly familiar with the safety audit process. However, application of the checklists could be a convenient introduction to considerations of safety issues in design as well as a learning process.

13.3 ROADSIDE BARRIERS

13.3.1 Introduction

In spite of the best efforts of geometric designers, some vehicles will go out of control and leave the road in an unplanned fashion for any of a number of reasons, including

- Driver incapacitation as a consequence of substance abuse
- Tiredness
- Poor judgment of speeds on curves
- Skidding or sliding out of control because of water, snow or ice on the road surface
- Being hit by other vehicles or stray animals on the road

Whatever the possible reasons for ROR crashes, designers must attempt to ensure that they do not occur but, when they do occur, to ensure that the consequences are minimised.

13.3.2 The assessment of need

Barriers are, of themselves, hazards and the guiding principle is that they are the last resort after all other ameliorative measures have been considered and, for whatever reason, rejected. The sequence of the decision-taking process regarding roadside obstacles is

1. Removal
2. Relocation to reduce the chance of them being hit
3. Redesign so that they can be safely traversed
4. Redesign to be frangible or break away, or to otherwise reduce severity
5. Only when all possibilities have been exhausted, to shield with a safety barrier or crash attenuator

In this process, the economic and environmental consequences of the 'Do nothing' alternative have to be borne in mind. In short, what is required is a risk analysis approach to the decision of whether to provide a barrier or not. The basic philosophy of risk analysis is that it seeks to

- Control potential losses by determination of the costs associated with a loss-making situation
- Determine the probability of the loss-making situation occurring
- Relate potential losses with the cost of controlling or removing the loss-making situation

Risk management uses the outcome of the risk analysis process to determine whether

- The risk of a crash at a particular site is such that an action is required to mitigate this risk
- Even though some action could be warranted, there are other sites where the need for action is greater

- No remedial action is justified (the 'Do nothing' alternative)
- The priority of such action is so low that only very minor treatment needs be provided

The risk analysis and management processes are required because no road authority has unlimited financial and physical resources and, in applying its duty of care, the best return on the application of funding and resource has to be sought.

The assessment of need incudes a determination of the distance over which the barrier is required. This is generally referred to as the length of need and is illustrated in Figure 13.2. The variables that need to be considered are illustrated in Figures 13.3 and 13.4.

The distance upstream of the hazard necessary to shield vehicles from impacting the hazard is calculated as

$$X = \frac{L_A + \frac{b}{a} L_1 - L_2}{\frac{b}{a} + \frac{L_A}{L_R}}$$

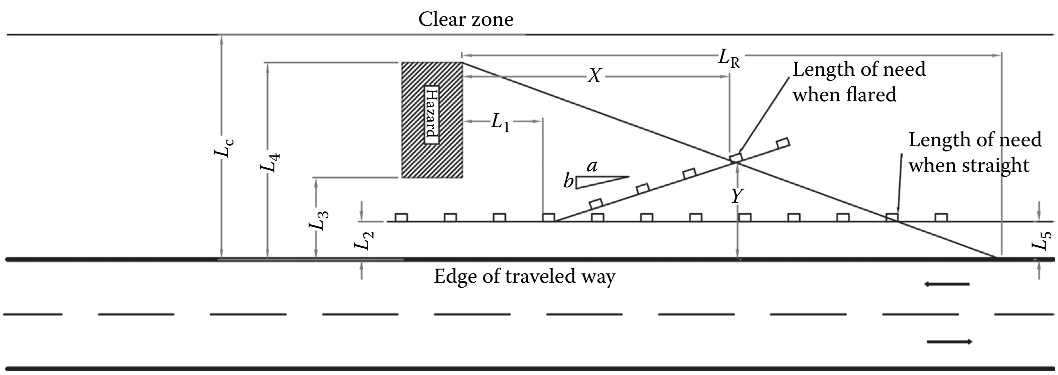


Figure 13.3 Upstream barrier design variables. (From Stephens LB. *Barrier guide for low volume and low speed roads*. Federal Highway Administration, Washington, DC, 2005.)

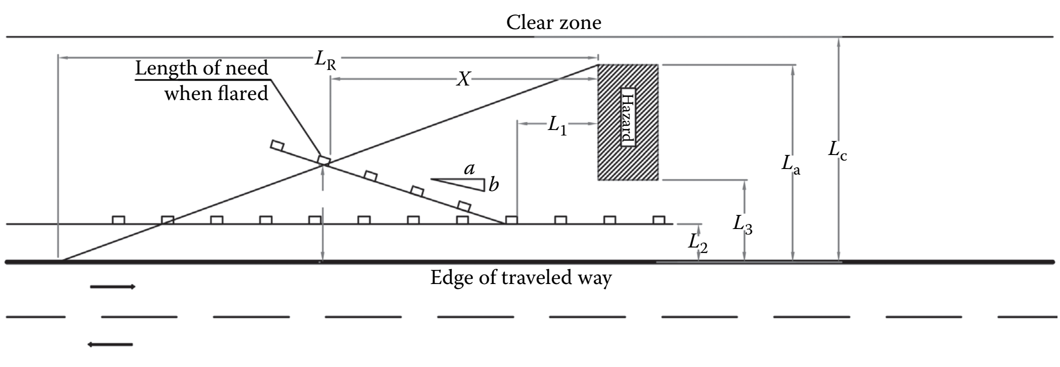


Figure 13.4 Downstream barrier design variables. (From Stephens LB. *Barrier guide for low volume and low speed roads*. Federal Highway Administration, Washington, DC, 2005.)

where

b/a = taper rate

L_1 = tangent length of barrier and defines the beginning of the flare

L_2 = distance from the edge of the travelled way to the front face of the barrier

L_A = distance from the edge of the travelled way to the back of the hazard

L_R = stopping distance between the edge of the travelled way and the back of the hazard Values are suggested in the AASHTO *Roadside Design Guide* (AASHTO, 2005)

L_S = shy line offset. Drivers tend to move away from rigid objects such as barriers or other vehicles. It is preferable to locate barriers beyond the shy line offset, which is generally of the order of about 1 metre

The downstream barrier design variables refer to the direction of travel on the far side of the road, that is, the opposing traffic length of need, and are illustrated in Figure 13.4. The edge of pavement is, in this case, the centreline of the road.

It should be noted that the terminal treatments both up- and downstream of the obstacle are outside the calculated length of need. Furthermore, the farther away from the road that the obstacle is located the greater is the length of the required guardrail until the point is reached where the obstacle is so far from the travelled way that the guardrail is no longer required. This is generally at a distance of about 9 metres from the edge of the travelled way.

13.3.3 Performance requirements

As stated previously, the primary purpose of roadside barriers is to prevent a vehicle from striking an obstacle or terrain feature that is less forgiving than itself. Assessing the effectiveness of a barrier is difficult because of the complex dynamics involved in a crash. The most effective method for testing the effectiveness of a barrier is through full-scale testing. In the United States, the document *NCHRP Report 350: Recommended Procedures for the Safety Performance Evaluation of Highway Features* (Ross et al., 1993) prescribed a total of six scenarios that could be employed as full-scale tests. This document was adopted in southern Africa, Australia and New Zealand.

In Europe, the standards adopted are similar in concept to those described in NCHRP 350 and are captured in the document EN-1317 published in 1998. Europe is concerned about the safety of motorcyclists and Technical Specification 1317-8 addressing this issue is currently in draft form.

NCHRP 350 was recently superseded by the *Manual for Assessing Safety Hardware* (AASHTO, 2009). The test scenarios have been slightly modified, generally by adopting heavier vehicle weights and modest changes to the angles of skew of impact between the test vehicle and the barrier. The current test scenarios are shown in Table 13.1.

Figure 13.5 illustrates a warrant for the application of a barrier for various heights of fill (metres). This figure is purely illustrative and should not be used without similar studies being carried by the highway agency concerned. Factors other than those implicit in the figure – the construction and crash costs – may also have to be taken into account in the determination of a comprehensive warrant. For example, there may be additional costs in respect of the cost of relocation of utilities or expropriation costs where a wider road reserve is required.

Table 13.1 Manual for Assessing Safety Hardware (MASH) test matrix for longitudinal barriers

		Test conditions		
Test level	*MASH test vehicle designation and type*	*Vehicle mass (kg)*	*Speed (km/h)*	*Angle (degrees)*
1	1100C (Passenger car)	1100	50	25
	2270P (Pickup truck)	2270		
2	1100C (Passenger car)	1100	70	25
	2270P (Pickup truck)	2270		
3	1100C (Passenger car)	1100	100	25
	2270P (Pickup truck)	2270		
4	1100C (Passenger car)	1100	100	25
	2270P (Pickup truck)	2270	100	25
	10,000S (Single unit truck)	10,000	90	15
5	1100C (Passenger car)	1100	100	25
	2270P (Pickup truck)	2270	100	25
	36,000V (Tractor-van trailer)	36,000	80	15
6	1100C (Passenger car)	1100	100	25
	2270P (Pickup truck)	2270	100	25
	10,000S (Single unit truck)	10,000	90	15

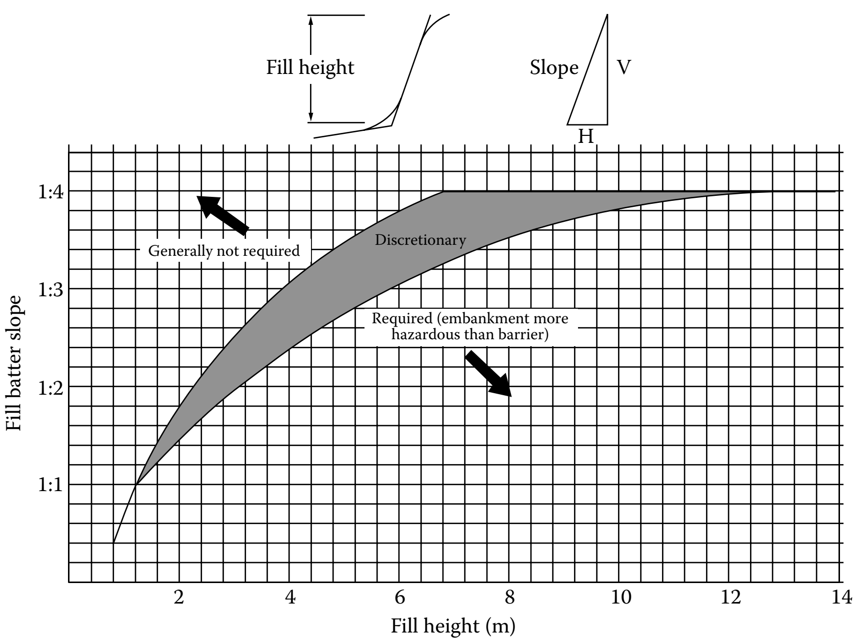


Figure 13.5 Example of a warrant for cost-effective barriers for fills. (From Burrell RC et al., *Geometric design guidelines*. South African National Roads Agency, Pretoria, 2002.)

13.3.4 Types of barriers

Types of barriers can be classified into three groups:

- Flexible barriers
- Semirigid barriers
- Rigid barriers

These are discussed further in the sections that follow.

13.3.4.1 Flexible barriers

Flexible barriers take the form of cables supported by posts as illustrated in Figures 13.6 and 13.7. Reference is often made to weak post and strong post systems. The difference between them is that the cross-section of the weak post is smaller than that of the strong post. In the case of timber posts, the strong post typically has a diameter of 175 to 200 millimetres, with the weak post having a diameter of 150 millimetres or less. The spacing of the posts in a weak post system is double that of the strong post system.

Two basic variations of cable barriers exist: low-tension and high-tension systems. Whereas low-tension systems are generic in the sense that they are not unique to any one supplier, high-tension systems are all patented by specific manufacturers.

Low-tension systems show a lateral deflection of anything up to about 4 metres whereas the lateral deflection in the case of high-tension systems is typically 3 metres or less. In consequence, the deceleration experienced by the occupants of a vehicle is low and the physical trauma they are likely to suffer is thus also low.

During the 1980s and 1990s low-tension systems were specified almost exclusively but have the disadvantage that, once the barrier has been damaged, the cable in the damaged section would fall to the ground and the entire length of barrier between terminals rendered useless. In spite of this weakness, they remain widely used worldwide, presumably because of their ease of erection and maintenance. The length of barrier between the terminals of a low-tension system is limited to between 500 and 600 metres. High-tension

Figure 13.6 Four-cable braided cable system.

(a)

(b)

Figure 13.7 Other cable barrier configurations. (a) Three-cable barrier and (b) two-cable median barrier on 2 + 1 cross-section. (From Federal Highway Administration. *Manual on uniform traffic devices*. Washington, DC, 2009.)

systems, on the other hand, retain their tension even though various supports may have been destroyed or knocked out of the sockets in which they were originally placed. There is thus no theoretical limit to the length of a high-tension system. These systems comprise three or four prestretched cables, with the prestressing tension being in the range of 10 to 40 kilonewtons.

Cable barriers tended to be used more in medians than at the edge of outer shoulders. This is largely because they can provide protection to vehicles on both sides of the median and are easy to repair when struck. Furthermore, if used as a roadside barrier and because of the large deflection they undergo when being struck, there is always the possibility that a vehicle, for example, a small passenger car, could go under the barrier and thereafter down the fill slope.

In theory, the width of the median, comprising the central island and the inner shoulders, should be of the order of 6 metres where cable barriers are used. This allows for a vehicle leaving the carriageway to be stopped by the cable barrier before encroaching on the opposing carriageway. Given that impact between two opposing vehicles could be at relative speeds of 250 km/h or more, the consequences would usually be fatal. Furthermore, the further trajectories of the struck vehicles could be such that numerous other vehicles could be involved in secondary crashes.

In spite of this caveat, cable barriers are widely used, and apparently with great success, on 2 + 1 cross-sections where the median would be only 1 metre wide or less. It is possible that this is because, apart from its obvious safety aspect, the cable barrier eliminates the possibility of drivers of vehicles on the single-lane side of the cross-section being tempted to enter the opposing lanes to overtake slower vehicles.

13.3.4.2 Semirigid barriers

Semirigid barriers include steel box beam guard rails, W-profile guard rails and thrie-beam guard rails mounted on steel or timber posts. Thrie-beam guardrails are similar to W-profile rail, but have three ridges instead of two.

Semirigid barriers have a lateral deflection of 1 to 2 metres compared to the 3 metres of the cable barrier and the zero deflection of rigid barriers. On being hit by an errant vehicle, the guardrail deforms to form a 'pocket' containing the vehicle, with the pocket extending as the vehicle slides along the guardrail before ultimately being brought to a halt.

The mechanism whereby the energy of the impact is dissipated is by deformation of the guard rail and the vehicle bodywork. Further dissipation of energy is by displacement of the post in the ground. Steel posts are also deformed on impact whereas timber posts absorb the impact by breaking. Slowing of the vehicle is by friction between the guardrail and the vehicle. Because of their rigidity, box beam systems spread the impact force over a number of posts.

The typical dimensions of a W-beam guardrail are shown in Figure 13.8. The spacing of the posts centre-to-centre in the case of the Armco W-beam guardrail is 1.6 metres for the strong post system and 3.2 metres for the weak post systems. The rail is blocked out from the post to ensure that vehicle wheels don't get snagged under the guardrail.

Various other profiles have been developed to suit different conditions. These include

- Midwest Guard Rail system
- The Gregory mini-spacer system
- The Trinity T31 system
- The Nu-Guard 31 system

All of these are described in the AASHTO *Roadside Design Guide* (AASHTO, 2005). Departures from the standard W-beam guardrail are relatively slight.

13.3.4.3 Rigid barriers

Rigid barriers are constructed of concrete and, historically, the first of these was the New Jersey concrete barrier. The currently preferred profile is the F-shape illustrated in Figure 13.9.

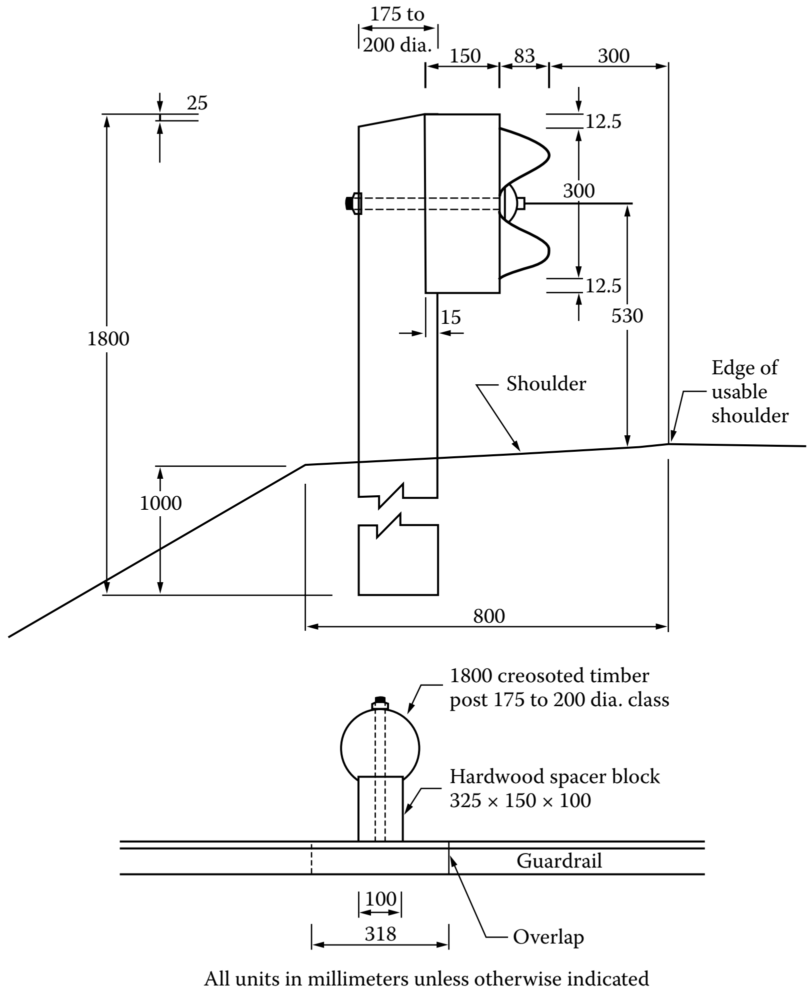


Figure 13.8 Dimensions of a W-beam guardrail. (From Wolhuter KM. *TRH 17: Geometric design of rural roads*. Committee of State Road Authorities, Pretoria, 1988.)

The lower of the two slopes of this profile provides redirection of vehicles where they impact the barrier at small angles of skew and low speeds. At higher speeds and more severe angles of skew, vehicles track up onto the higher slope. This has the effect of dissipating energy by lifting the vehicle off the pavement. Lifting the vehicle too high may result in its being capsized altogether and the F-profile reduces the extent of lift below that resulting from the New Jersey profile. In crash tests, the F-profile has been shown to be more effective than the New Jersey profiles insofar as it limits the extent of damage to vehicles and prevents rollover.

13.3.4.4 Barrier terminals

The terminals are the most dangerous portion of the guardrail because of the possibility of vehicles hitting them end on and being speared by the guardrail as illustrated in Figure 13.10.

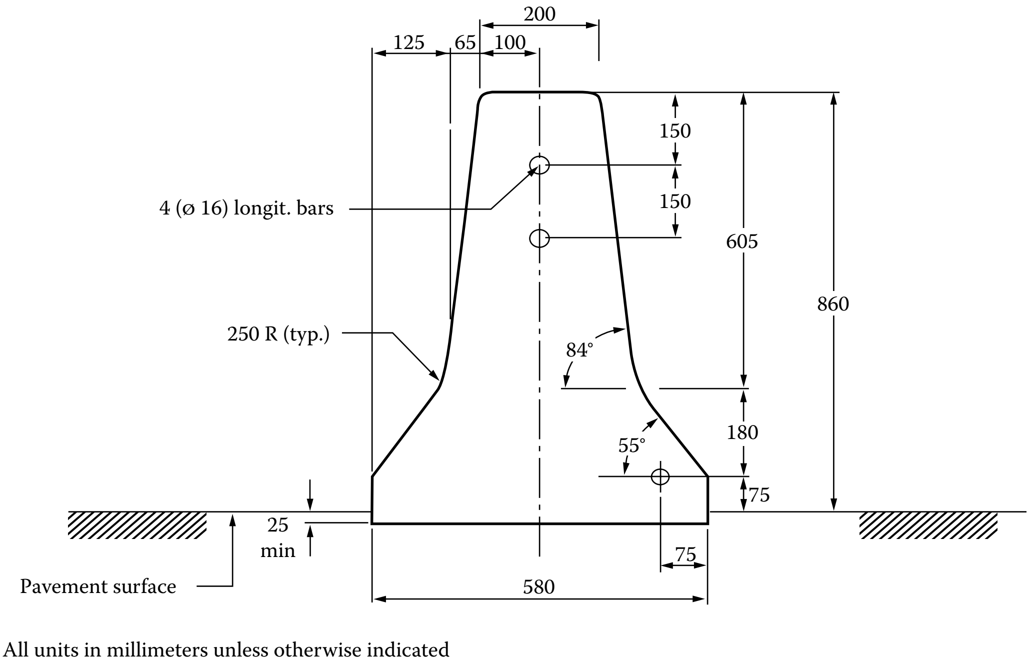


Figure 13.9 F-profile concrete barrier. (From Queensland Department of Transport and Main Roads. *Road planning and design manual*, Chapter 8, 2005.)

Figure 13.10 Vehicle impaled by guardrail terminal.

Before the inherent danger of their use was realised, flared end wings, also known as lobster tail end wings and illustrated in Figure 13.11, were widely used and, surprisingly, they are still on the market. These end wings, when applied at the upstream end of a guardrail, function like a chisel and can either impale a vehicle or cause the engine block to invade the passenger compartment. They are still in use but their application is limited to the downstream end of a semirigid barrier.

End treatments currently in vogue include the bull nose and the flared end treatment both of which are illustrated in Figure 13.12. The bull nose end treatment, if slightly splayed and modified, is intended to shear off its posts and roll up like a tin can. The flared end treatment usually follows a parabolic path and its end often is also buried. For some time, the end treatment was buried without being flared but this has largely been discontinued because vehicles could slide up the terminal and become airborne.

Much research has been done over the years and various nonproprietary designs intended to address the weaknesses of the standard terminal treatments have been developed. Reference is made to the breakaway cable terminal (BCT) which, although in common use, has been superseded by the modified eccentric loader terminal (MELT) illustrated in Figure 13.13 and the slotted breakaway cable terminal (SBCT). The MELT uses breakaway timber or steel posts and the SBCT is a weak post system incorporating a slotted W-beam that breaks and swings back behind the terminal.

There are also numerous propriety terminals available and designers should consult the manufacturers of these terminals for information regarding their failure mechanisms and appropriate applications.

The terminals of flexible barriers are relatively safe because the cable terminals are usually buried concrete blocks. While vehicles would tend to slide up the top cable, this cable

Figure 13.11 Flared end wing.

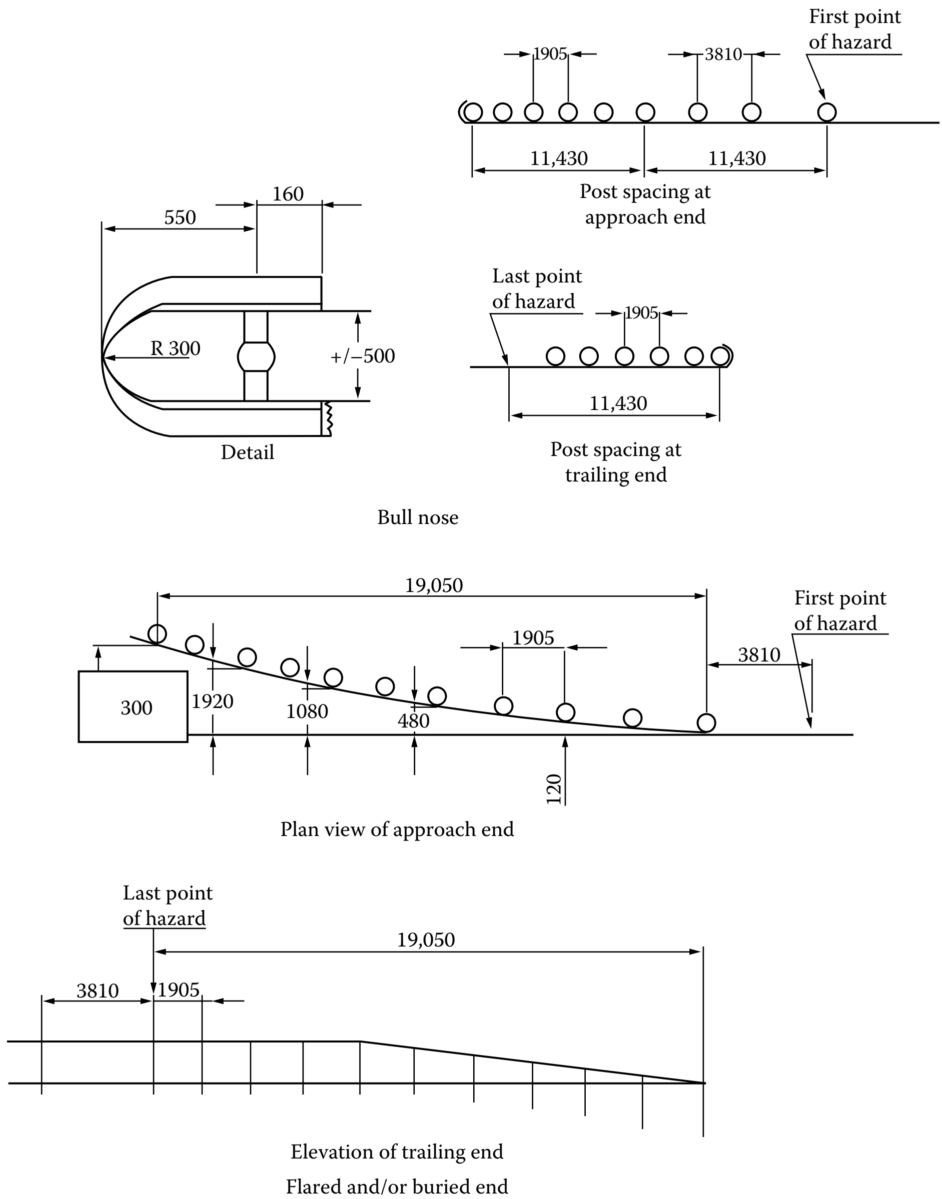


Figure 13.12 Some typical semirigid terminals.

could bend under the load and the barrier posts thus serve as impact attenuators. The barrier posts are usually inserted into short steel pipes and can be either dislodged or simply bent.

The terminals of rigid barriers may comprise a sloping section commencing at ground level and sloping up to the height of the barrier, as illustrated in Figure 13.14.

These are not safe as terminals where high speeds are expected such as on freeways and urban arterials. Vehicles impacting them would slide up the terminal and become airborne. However, at speeds of less than 60 km/h they are acceptable. At speeds higher than this it is

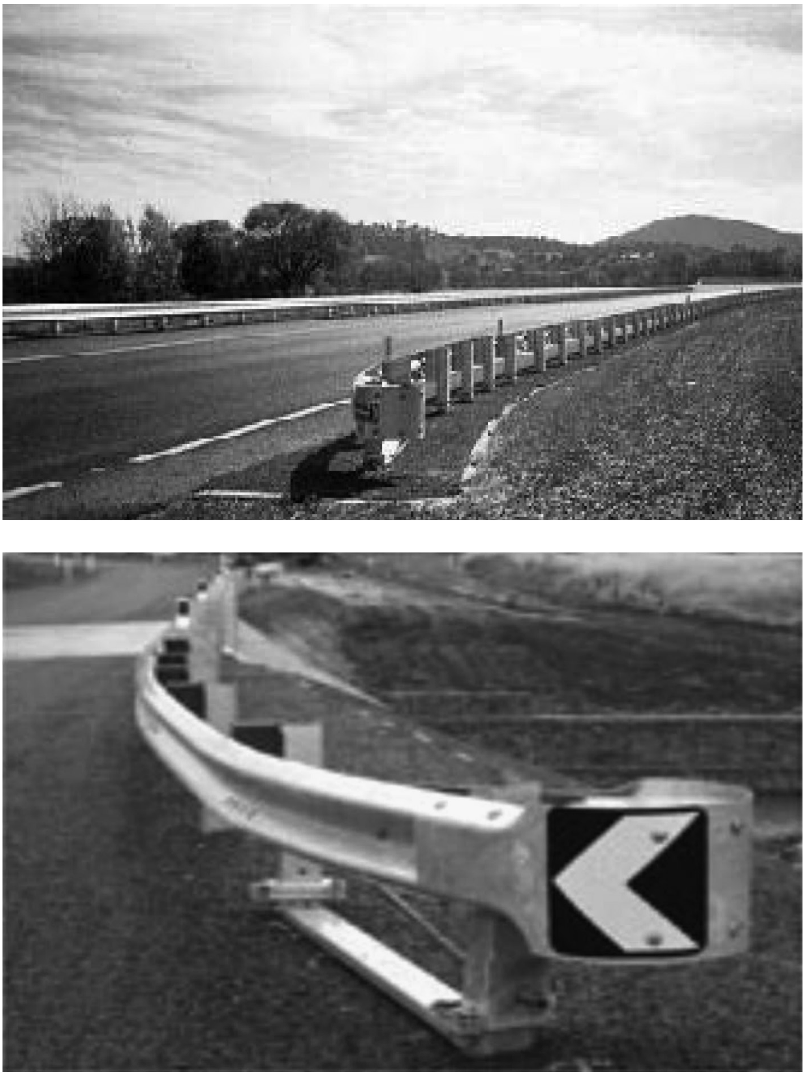


Figure 13.13 Modified eccentric loader terminal (MELT). (From Fitzgerald WJ. *W-beam guardrail repair: A guide for highway and street maintenance personnel.* Federal Highway Administration, Washington, DC, 2008.)

necessary to consider impact attenuation as the mechanism for bringing vehicles to a standstill. Impact attenuation is discussed in Section 13.4.

13.3.5 Median barriers

Median barriers are designed to reduce the risk of an errant vehicle

- Colliding with a vehicle traveling in the opposite direction
- Deflecting a vehicle back into the traffic stream traveling in the same direction
- Decelerating beyond tolerable occupant limits

Figure 13.14 Low-speed terminal for rigid barriers. (From Bligh R et al., *Median barrier guidelines for Texas*. Texas Department of Transportation, Austin, 2006.)

It has been suggested previously that out-of-control vehicles would, as a general rule, not travel further than 9 metres laterally from the edge of the travelled way. If the median, which includes the central island and the median shoulders, were to be narrower than this, there is always the possibility that a vehicle could cross it and end up in the opposing carriageway. Traffic volumes on dual carriageways, particularly on freeways, are such that a head-on collision is almost inevitable. The closing speed of two vehicles on reciprocal headings would probably be of the order of 250 km/h and fatalities would be sure to result. The impact would send both vehicles on unpredictable trajectories, resulting in a multi-vehicle pile-up and more injuries and fatalities. In short, where medians are narrower than nine metres, median barriers would be essential. If there is a significant height difference between the two carriageways, it may be necessary to consider the installation of a barrier on a median wider than this.

It must be noted that median barriers also, as stated previously, constitute a hazard. If a barrier is provided and centrally located in the median, it follows that the distance available for recovery is halved. Where a driver could have recovered control of the vehicle if the full width of the median had been available, it is possible that a crash involving the median barrier could occur because of the lesser width (Bligh et al., 2006).

Caltrans recommends that consideration be given to the provision of median barriers if a location has had either of the following:

- Three or more cross-median crashes and a total cross-median crash rate of at least 0.5 crashes per mile (0.3 crashes per kilometre)
- Three or more fatal collisions and a crash rate of at least 0.12 crashes (0.07 crashes) per mile per year in a 5-year period

Caltrans, however, suggests that the probability of a cross-median crash is low if the annual average density traffic (AADT) is less than 20,000 or the median width is more than 22 metres.

Median barriers are normally flexible or cable barriers, which cost only about one quarter of the cost of a concrete barrier to install. Experience has shown that, where cable barriers have been installed, fatal crashes no longer occur. Median barriers may also comprise back-to-back guardrails mounted on single posts. Where medians are very narrow, it may be necessary to consider rigid, or concrete, barriers.

The use of safety barriers tacitly assumes that there is sufficient space behind the barrier to allow for barrier deflection. If this is, in fact, the case, it would be useful to consider the use of a gating terminal. Gating terminals allow a vehicle to pass through the leading end of the barrier and come to rest in a runout area. According to Australian practice, the MELT terminal referred to in Section 13.3.4.4 and illustrated in Figure 13.13 is a gating terminal that requires a runout area with a width of about 6.0 metres and a length of 22.5 metres. The runout area has to be clear of all obstacles and be drivable.

13.4 THE CLEAR ZONE CONCEPT

In the 1960s, Jack Leisch was propounding the concept of the Forgiving Highway. The philosophy was that, regardless of the mistake made by a driver, the consequences of that mistake should be, at most, trivial. Hitting a roadside barrier could not be construed as being trivial and, furthermore, it is plainly uneconomic to provide barriers for every metre of the road. The alternative is to provide a fill or cut batter sufficiently flat to ensure that an out-of-control vehicle does not become airborne, roll or dig in and somersault. In short, the driver has to be afforded an opportunity to recover control of the vehicle. This does not necessarily mean that it would be possible to return to the travelled way because this would imply a change of direction against a significantly adverse superelevation. All that the driver could expect to do was to maintain the direction of travel or to turn to travel down the fill slope or up the cut slope at right angles to the contours of the slope.

The concept of the Forgiving Highway meant that this could be accomplished without the risk of hitting any obstacle. As stated earlier, it has been found that, in 80 to 90 per cent of the occurrences of running off the road, a vehicle would move no further than about 9 to 10 metres away from the travelled way. This is the area that has to be cleared of all obstacles, be they trees, culverts, road signs or street lighting and signalisation. In practice, this is not always achievable and the Forgiving Highway, as worthy a goal as it might be, sometimes has to be replaced by a compromise, the Clear Zone concept.

This concept suggests that every obstacle likely to be encountered by an errant vehicle be scrutinised and subjected to a range of possible alternative actions. These, in descending order of desirability are

- Total removal of the obstacle
- Relocation of the obstacle
- In the case of a manmade obstacle, modification of the obstacle by providing it with a frangible or slip base
- Shielding the obstacle with a barrier
- Doing nothing

Doing nothing would be legitimate if traffic volumes were so low that the risk of a vehicle hitting the obstacle is negligible. Doing nothing could be somewhat of a misnomer because it could actually include ‘doing something’, for example, treatment of the fill or cut

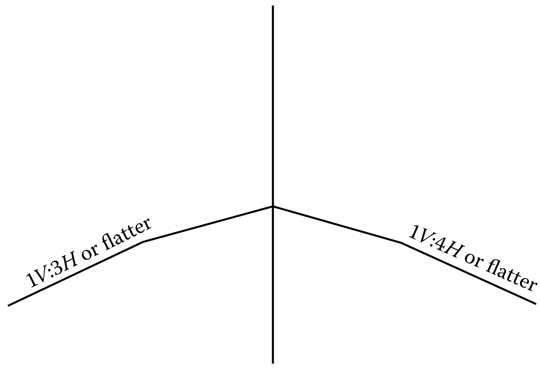


Figure 13.15 A Mansard cross-section.

slope. This treatment refers to the entire longitudinal length of the slope to be addressed and not simply in the immediate vicinity of the obstacle. In general, a slope of 1*V*:3*H* or steeper is considered to be nontraversable, that is, a vehicle on it and being out-of-control will remain so, and with a high propensity to capsize, until the bottom of the slope has been reached after which a further clear space is required for the vehicle to be brought to a stop. If such a slope has to be provided, a barrier at the top of the slope would have to be considered.

As an alternative, a Mansard roof cross-section, sometimes called a 'barn roof' cross-section as illustrated in Figure 13.15, could be considered.

As shown, a recoverable slope of 1*V*:4*H* or flatter is located adjacent to the shoulder. A nonrecoverable slope (not steeper than 1*V*:3*H*) follows. If space constraints force a slope steeper than this, it may be necessary to revert to the use of a barrier at the top of this slope depending on the height difference still remaining to be traversed.

The bottom of a cut section usually includes a side drain. The sides of the drain should be at least 1*V*:3*H* for the foreslope. In this case, the back slope could not be steeper than 1*V*:10*H* if the vehicle is not to dig in and then somersault. As an alternative, the foreslope could be at a slope of 1*V*:10*H* and the backslope at 1*V*:3*H*.

13.5 IMPACT ATTENUATION

Nongating terminals are designed to arrest the errant vehicle when impacted and to redirect the vehicle along its length, without allowing it to pass through the terminal. This type of terminal is appropriate if a hazard exists behind the terminal, where vehicle penetration is not acceptable, and it is not possible to extend the barrier to accommodate a gating terminal. The nongating terminal is intended to be hit end on and then to collapse while bringing the impacting vehicle to a safe halt. These energy-absorbing terminals are often described as 'crash cushions'.

Impact attenuators were previously constructed of drums filled either with sand or water. This form of impact attenuator was invented by John Fitch, a racing driver who claimed that the thought came to him because of the sand-filled drums he used to protect his tent from strafing Messerschmitts in North Africa during World War II. Their layout is illustrated in Figure 13.16.

The Fitch barrel form of attenuator works by momentum transfer. Sand or water is forced out of the drums so that the momentum of the vehicle is transferred to the moving mass of

Figure 13.16 Fitch barrels.

sand or water. The depth of water or sand increases from the front row of drums toward the back thus absorbing more momentum as the moving vehicle gets closer to the protected object. The typical layout of drum impact attenuators for various widths of obstacle is illustrated in Figure 13.17.

A similar idea is embodied in the REACT 350 system. It comprises so-called 'smart' cylinders made of high molecular weight, high density polyethylene (HMW/HDPE) plastic. The cylinders deform under load and thereafter regain up to 90 per cent of their original shape and capacity without maintenance or repair. The system includes units designed for impact at design speeds from 72 km/h (45 mph) to 100 km/h (62 mph).

Typical impact attenuators are illustrated in Figure 13.18. These are nongating attenuators and function by deforming in a linear or concertina fashion. They are specifically designed to match the requirements of NCHRP 350 and focus on ease of replacement and maintenance.

13.6 TRAFFIC CALMING

Traffic calming is focussed on the quality of residential life in urban areas. It is a retrofitting exercise aimed resolving the problems caused by the traditional grid layout of many older urban areas. In Chapter 4 it is pointed out that residential streets, being mainly collectors and local streets, have functions other than the smooth flow of traffic at high speeds. The basis for traffic calming is to broaden traffic engineering to a context-sensitive approach including designing for these other functions that are concerned with the social aspects of the street network. Social aspects include a sense of community awareness with people meeting casually in a neutral area without having to go to the extent of formal invitations to visit. They also include children playing team games in the street, riding bicycles and generally enjoying a safe and peaceful environment.

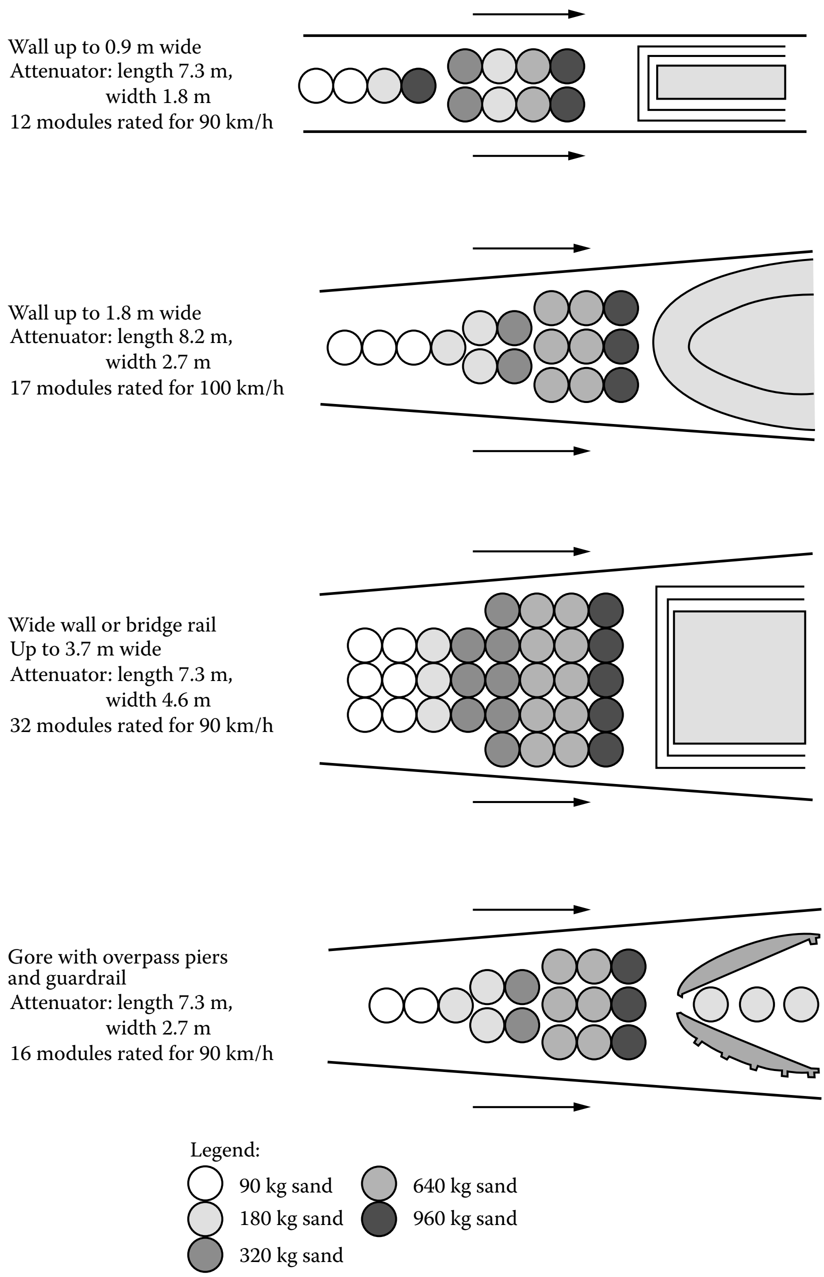


Figure 13.17 Typical layout of drum impact attenuators. (From Burrell RC et al., *Geometric design guidelines*. South African National Roads Agency, Pretoria, 2002.)

The need for traffic calming stems from the fact that commuter routes between residential areas and places of work tend with time to become congested. Drivers then seek alternative routes that are less stressful than arterials which in peak hour traffic are frequently operating in the start and stop mode of level of service F. These alternative routes are usually through residential areas and reference is made to rat-running. The quality of residential life

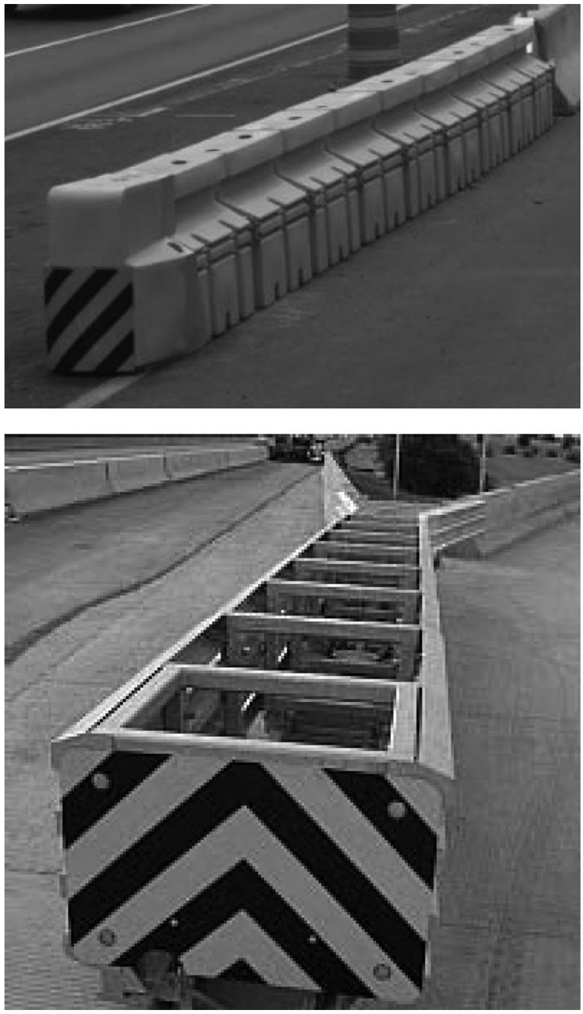


Figure 13.18 Typical proprietary impact attenuators. (From Washington State Department of Transportation. *Design manual.* Olympia, 2014.)

is downgraded by the noise levels and air pollution generated by the passing traffic, and the local street as a playground for children is definitely not an option.

Residential areas are served by collectors and local streets. Pedestrian volumes are low and children playing in streets should not be unexpected. Traffic calming comprises reduction in traffic volumes as well as reduction in traffic speeds and should be brought to bear on the design of residential streets.

Traffic volumes can be kept low by ensuring that residential streets do not offer a convenient alternative to congested arterials. This is achieved by selective street closures replacing four-legged intersections by three-legged intersections using diagonal diverters as illustrated in Figure 13.19.

Traffic speeds are generally kept low through the application of speed humps and chicanes. Speed humps are typically about 4 metres long and span the full width of the road. They are typically parabolic or sinusoidal in profile with a height in the range of 75 to

Figure 13.19 Diagonal diverter.

100 millimetres. The application of speed humps is limited to residential streets. They are totally inappropriate on higher order streets.

Speed humps are requested invariably by local residents and criticised by the drivers they are intended to slow down. Court actions for damages suffered by vehicles traversing speed humps at too high a speed are usually unsuccessful provided the speed humps are visibly marked and preceded by warning road signs. A legitimate criticism is their effect on emergency vehicles. An ambulance with a patient on board has to add about 10 seconds of travel time for every speed hump encountered and fire engines about 3 to 5 seconds. The alternative is to replace speed humps with speed cushions. These are speed humps with level cut-throughs spaced at the width of the wheelbase of trucks. These also have the effect of providing smooth passage for buses.

While speed humps reduce speeds by the use of uncomfortable vertical accelerations, chicanes reduce speeds by means of short radius horizontal reverse curves or deflections as illustrated in Figure 13.20.

13.7 ARRESTOR BEDS

Arrestor beds, also known as truck ramps, runaway truck lane, escape lane or emergency escape ramp are traffic devices that enable vehicles that have suffered brake fade to stop safely. They are typically long, sand or gravel-filled lanes adjacent to roads with steep downgrades, and are designed to accommodate large trucks. The loose material allows the trucks' momentum to be dissipated in a controlled fashion. An example of an arrestor bed is shown in Figure 13.21.

Arrestor beds are normally used adjacent to long steep downgrades. Drivers, particularly those who are less than expert, are inclined to rely on their brakes to control the speed of their vehicles. In consequence, the brakes overheat and fail. More experienced drivers change to lower gears before essaying a long down grade using, as a general rule of thumb, that the gear engaged on the downgrade is the same as that needed in the opposite direction. Many authorities provide warning signs to this effect, and the W7 series of signs illustrated in Figure 2C-2 of the *Manual on Uniform Traffic Control Devices* (MUTCD; Federal Highway Administration, 2009) indicates the format of signs appropriate to arrestor beds.

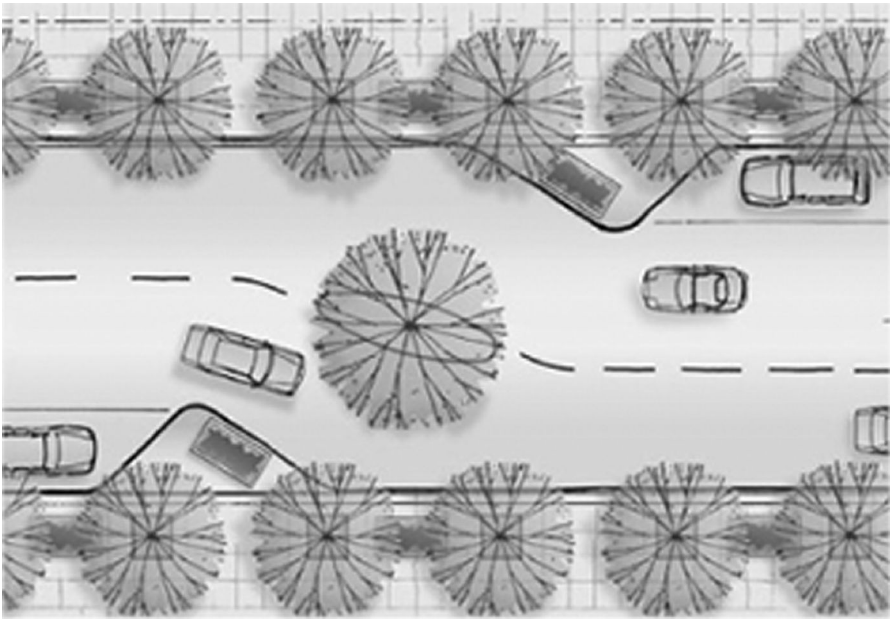


Figure 13.20 Schematic of a chicane.

It is also recommended that compulsory truck stops be provided at the commencement of a long downgrade. This will ensure that the truck's brakes are cool at least at the start of the downgrade. Furthermore, it sometimes happens that, in changing gears when the truck's speed is already out-of-control, it is simply not possible to engage the lower gear. The compulsory stop will make it easy for the driver to pull away from the stop while selecting the appropriate gear in the process.

Figure 13.21 Arrestor bed on A5 in Germany.

Figure 13.22 Safety net arrestor bed.

Various alternative types of arrestor beds exist. These include the upgrade ramp, which is normally very long because it does not use a gravel or a sand bed or the loose sand pile. The sand pile is also not recommended, as rainwater can cause the sand pile to compact and, under winter conditions, the sand pile can freeze. Both of these conditions result in the sand pile not performing as it should and also make it possible for trucks to overturn on hitting the sand pile. The layout of a typical arrestor bed is shown in Figure 13.22.

A proprietary safety net system comprising a stainless steel mesh can be used if space is too restricted for the provision of a conventional arrestor bed. This is illustrated in Figure 13.23.

The Fitch barrels at the end of the ramp should be noted.

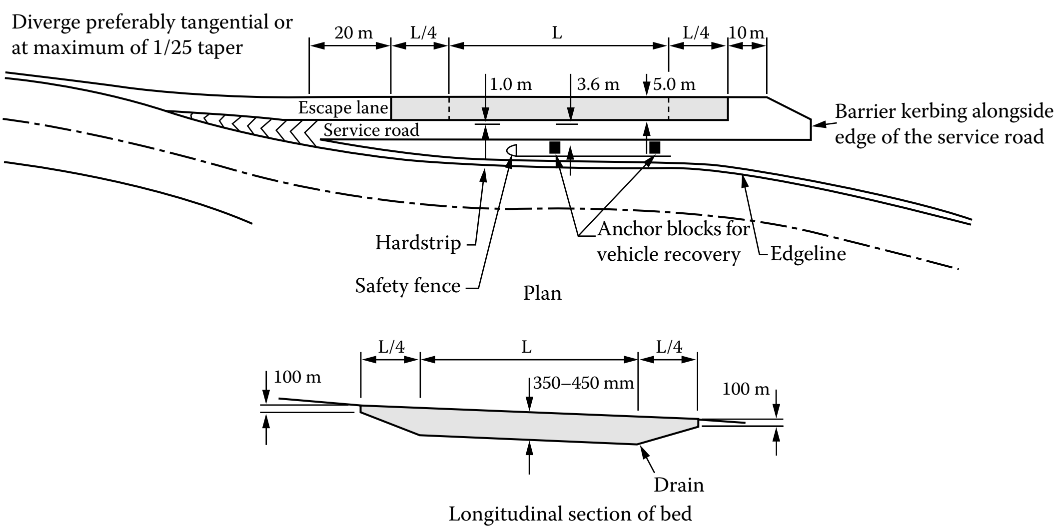


Figure 13.23 Typical layout of an arrestor bed. (From Burrell RC et al., *Geometric design guidelines*. South African National Roads Agency, Pretoria, 2002.)

13.8 ELDERLY DRIVERS

As individuals age, their faculties diminish. Vision becomes less acute, hearing starts to fail and cognitive ability is both reduced and slowed. In addition, physical suppleness declines, reducing, for example, the ability to turn the head to observe features outside the normal range of vision of a person looking straight ahead. Finally, physical frailty causes the elderly to be more likely to suffer serious injury in an incident that, in a younger person, may merely result in bruising.

Possibly because of advances in medical science and definitely because of an increased focus on healthy living, people are living longer. In consequence, the proportion of the driving population that can be considered elderly is increasing. First World populations are aging fairly rapidly. In America, the population growth rate of those at the normal retirement age of 65 or older is currently three times that of the total population. In 1998, the average life expectancy in the United States was 76.1 years whereas 200 years previously it was less than 40 years and was still below 50 years in 1902. In 2000, the elderly accounted for 12.4 per cent of the US population and it is estimated that this group will comprise 20 per cent of the total population in 2030.

In Europe, 15 per cent of the population was older than 65 in 2000 and current estimates are that this percentage will double by 2050. The negative population growth rates of First World countries are the main reason for the rapid increase in the percentage of the elderly in the population.

As shown in Table 13.2, observations made on site indicate that wrong-way driving and failing to yield or to stop are the predominant behaviour patterns of elderly drivers, with ROR crashes also being very frequent. Elderly drivers also frequently failed to signal their intentions. These behaviour patterns are particularly significant in the case of intersections.

The numbers reflected in the table show that the incident either triggered a crash or that a violation occurred and a citation was issued or that the incidents were merely observed but without any crash or the issue of a citation occurring.

Various interventions could be considered to ensure that the elderly do not prematurely lose the freedom of movement available with the use of a passenger car. For example:

Table 13.2 Numbers of incidents attributed to elderly drivers

Incident	*Crash*	*Violation*	*Observation*	*Total*
Wrong-way driving	29	149	13	191
Failing to yield or to stop	74	114	3	191
Off road	176	8	1	185
Turning across oncoming traffic	46	43	0	89
Driving abnormally slow	0	56	9	65
Rear end crash	49	0	1	50
Backing into other vehicles or obstacles	32	1	1	34
Crossing lane markings	5	25	25	30
Failure to yield to pedestrians/ cyclists	16	5	5	24
Miscellaneous	58	43	43	140

- Elderly drivers could be retrained to accommodate the diminution of abilities that normally accompanies aging.
- Elderly drivers could be required to take driving tests periodically and failing the tests could automatically result in cancellation of the driver's licence.
- Design policies and standards could be revised to be more generous to elderly drivers although it is not economically feasible to endlessly increase standards, and a cut-off point would be reached where drivers would have to give up driving either voluntarily or involuntarily.
- Designers take cognisance of the ordered list of driving errors shown in Table 13.2 and seek to simplify the driving task as far as possible to minimise the problems experienced by elderly drivers.

13.9 PROVISION FOR MOTORCYCLISTS

As stated previously, in European countries there is a perception that cable barrier systems are particularly hazardous for motorcyclists, with references being made to cable barriers as 'cheese cutters'. Political pressure in Norway has resulted in cable barriers being banned in that country. No reason for this perception has been found. Given that motorcyclists falling off their bikes are essentially unprotected except for whatever safety clothing or crash helmets they may be wearing, the nature of the barrier they hit is irrelevant (Daniello and Gabler, 2011). With all types of barrier, be they flexible, semirigid or rigid, motorcyclist injuries are likely to be serious, if not fatal.

The Norwegian ban on cable barriers is probably an ill-considered knee jerk reaction to lobbying by active motorcyclist-based organisations. In terms of traffic composition, motorcycles form a small percentage of the total traffic stream. Benefitting them at the expense of the rest of the traffic stream just does not make sense. Cable barriers have been found to be the most effective of all types because their flexibility results in significant reductions in the level of damage suffered by vehicles and their occupants compared to the other forms of barrier.

Research on which dummies are propelled into barriers is generally based on the dummies striking the barrier head first whereas the trajectory of the dummies could take the form of sliding with any part of the body leading the slide, rolling or somersaulting. More research is thus needed to take account of these alternative forms of translation. In the case of a flexible or a semirigid barrier, a sliding motorcyclist could actually pass underneath the barrier and continue the slide down the earthworks fill slope until finally stopping, dazed but possibly relatively unhurt. Adding a further barrier virtually at ground level is thus a contravention of the philosophy of the barrier being a lesser danger than the object it shields.

The jury is thus out on the form of barrier most suitable to the accommodation of an unseated motorcyclist.

Chapter 14

Vulnerable road users

CONTENTS

14.1 INTRODUCTION

Historically, designers focussed on the rapid and efficient movement of motorised vehicles as transportation has always been defined as the safe, efficient, convenient and affordable movement of people and goods with minimum side effects. The last mentioned focussed on the environment and even that was biased by what was visible to the occupants of a moving vehicle. Obviously, pedestrians and others had no business being in the road reserve. Sidewalks were provided in urban areas but only as an afterthought.

Fortunately, sanity has prevailed. With the notable exception of the freeway, where the presence of pedestrians is actively discouraged albeit for their own safety, it is now realised that everyone in the road reserve has a perfect right to be there. Pedestrians moving between origin and destination, window shoppers and people sitting in sidewalk cafes or meeting friends by chance contribute to the richness of the urban environment, as do children playing in the streets. It follows that they all have to be catered to.

Vulnerable road users are all those who are not in vehicles. Because of the difference in speed of animal-drawn vehicles relative to that of motorised vehicles, vulnerable road users also include those sitting in animal-drawn vehicles.

There is a growing belief that motorcyclists should also be considered to be vulnerable road users. While they are on their motorcycles and moving with the stream of vehicles there isn't a problem. However, when they are either knocked off or fall off their motorcycles, the only protection they have is a safety helmet or leather clothes (generally referred to as 'racing leathers'). Injuries suffered by motorcyclists are generally severe, if not fatal. They are not, however, discussed in this chapter because the principal injury from a geometric design point of view is limited to the effect of striking a guardrail in a sliding, rolling or somersaulting mode. Motorcyclists are considered in Chapter 13.

In this chapter the nature and needs of the various vulnerable users are discussed, as are the design features necessary to accommodate them.

14.2 PEDESTRIANS

14.2.1 Introduction

The traditional approach to design is to define a distribution of whatever quality is often of interest, be it sight distance, reaction time or any other among the host of parameters of design. A level of the distribution is then selected as the basis of design, and this could be the 85th percentile at the high end of the range or the 15th percentile at the low end. The argument usually offered is that it is not economically effective to address the entire range of values of the parameter of interest.

A new approach to design is, however, now emerging and this has to do with the right to access. It is based on the concept that no one should be discriminated against or denied access to public facilities or services. This includes all those with any manner of disability such as impaired sight or hearing or difficulty in walking. In certain instances protection of these rights is enshrined in the Americans with Disabilities Act of 1990. The argument offered in addressing this issue is that roads should be designed for all users, including those with disabilities. Increasing accessibility for users with disabilities obviously increases the accessibility for all users.

In this section an attempt is made to identify design issues and the manner in which they can be resolved to the benefit of all.

14.2.2 The pedestrian as a design vehicle

Insofar pedestrians individually occupy a physical space and move at speeds that vary according to preferences and capabilities, they are as much a 'design vehicle' as any passenger car, truck or bus and correspondingly in need of definition.

14.2.2.1 Dimensions

The body ellipse of the 50th percentile pedestrian has a major axis of 0.6 metre and a minor axis of 0.5 metre. This body ellipse does not provide for the additional space required by people on crutches or in wheelchairs. Visually impaired people using a cane swing it from side to side for a distance approximately double the width of the major axis. If they rely on seeing-eye dogs, they would occupy a space approximately 1 metre square. People pushing prams or baby walkers require a longitudinal space of about 2 metres. The dimensions of pedestrians, specifically the amount of space they need to not feel hemmed in by others, is shown in Figure 14.1.

This personal space varies according to circumstances. In a crowded situation, for example in an elevator, people would tolerate others standing so close to them that they are almost touching. When two people are in conversation, a distance of 1.0 to 1.5 metres apart would be more acceptable.

Vertical clearances have to make provision for visually impaired pedestrians. A protrusion with its bottom edge about 0.7 metre above ground level would not be detected by a person using a cane. Overhead signs should be mounted at a minimum height of 2.5 metres above ground level to ensure that a visually impaired pedestrian does not suffer head injuries.

14.2.2.2 Wheelchairs

Wheelchairs may be manually powered or battery operated. In addition to conventional wheelchairs, there are wheelchairs designed for sporting activities or, with balloon or other wider wheels, even for access to beaches.

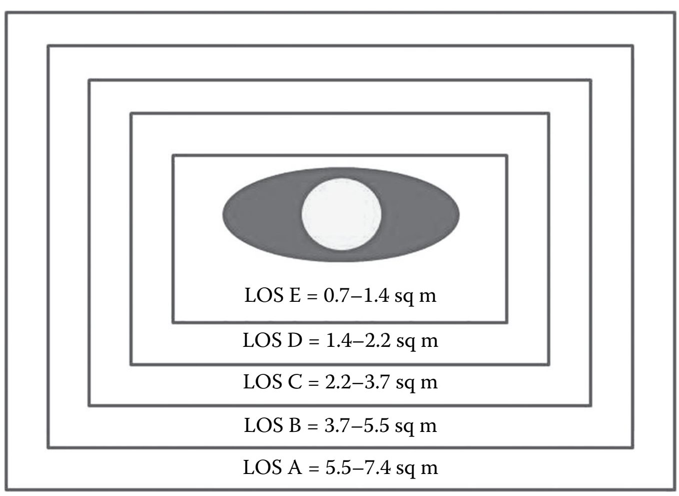


Figure 14.1 Spatial requirements of pedestrians.

Table 14.1 Typical dimensions of relevance for a wheelchair user

Characteristic	*Dimension (m)*
Eye level of wheelchair user	1.245
Seated height of wheelchair user	1.300–1.385
Width of wheelchair user allowing for elbows	0.9
Turning circle in a manually powered chair	See Figure 14.2
Turning circle diameter in an electric wheelchair	2.4000–4.350

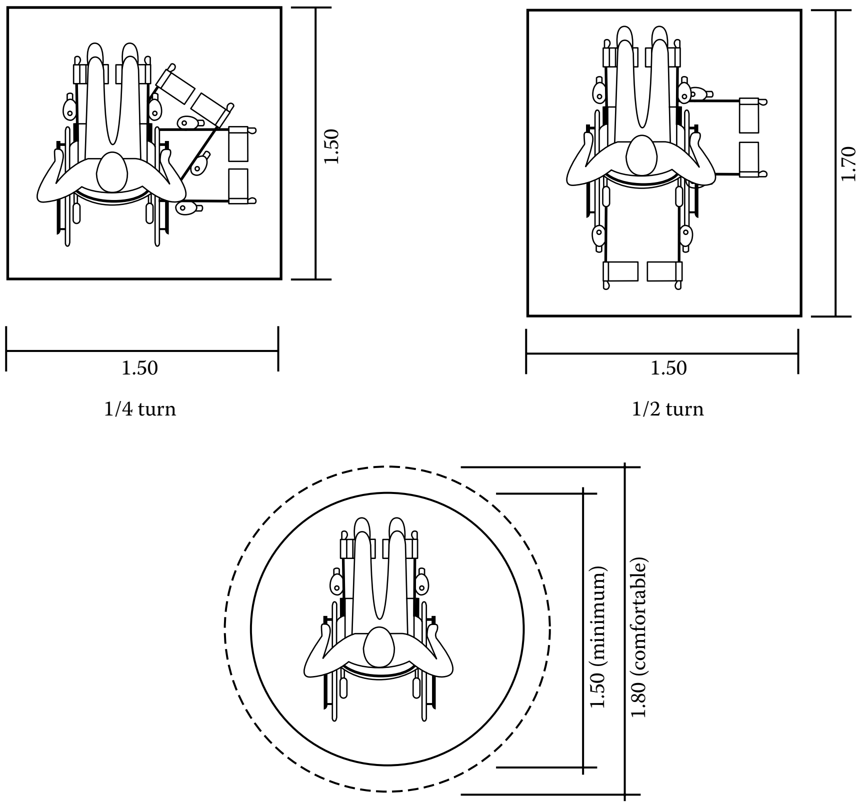


Figure 14.2 Required turning space for a wheelchair. (From South African Department of Transport. *Pedestrian and bicycle facility guidelines Draft 1.0*. Pretoria, 2002.)

Wheelchairs occupy a space of 1.2 metres by 0.75 metre while stationary. Any opening to be traversed by a wheelchair should thus not be narrower than at least 0.8 metre. An opening of any length, such as a passage, should have a minimum width of 0.9 metre. Two wheelchairs could pass each other only with difficulty if the total width of the opening or passage is less than 1.8 metres. The typical dimensions of relevance for a wheelchair user are illustrated in Table 14.1.

14.2.3 Performance of pedestrians

Except for getting exercise, Americans do not seem to walk much. This is attributed to urban planning being oriented towards travel by passenger car. The average of the top 30

American cities in terms of percentages of trips made by walking amounts to only a little more than 5 per cent. In Europe, walking and cycling comprise 20 to 50 per cent of all modes of transport. In Switzerland, 45 per cent of trips are made by walking and 5 per cent by bicycle. The averages for Spain, Germany and Sweden are 23 per cent by walking and 9 per cent by cycling. In the United Kingdom the percentages are 24 per cent by walking and 3 per cent by cycling (Kuzmyak and Dill, 2012). Given the extent to which walking features as a mode of transport, it follows that it merits close analysis.

Apart from dictating the length of a trip, speed of walking also determines the length of time that a pedestrian will need to cross a street. If gaps of this duration or longer are few and far between, pedestrians may become impatient and start taking chances on crossing streets one lane at a time. It is suggested that a lane marking does not constitute an adequate pedestrian refuge. In the interests of road safety, a form of control, such as providing for compulsory stopping of motorised vehicles to allow pedestrians to cross, is necessary. This is discussed in Chapter 17. Designing on the basis of an 85th percentile walking speed implies that people who experience difficulty in walking could end up being stranded in the middle of the street once the design time has elapsed.

Walking speeds for an older person may range from 1.25 to 1.30 m/s and, for a younger person, about 1.5 m/s. In imperial measure a brisk walking pace is about 4 miles per hour (which is convenient for people involved in route location – every 15 minutes elapsed means that you are 1 mile further from your starting point so you always know exactly where you are on your 1:50,000 survey plan). For design purposes, 1.25 m/s can be adopted for calculating the length of time it takes to cross a road. However, if the pedestrians being designed for include a high percentage of the elderly, a walking speed of 1 m/s should be adopted.

14.2.4 Geometric planning for pedestrians

Sidewalks are provided for travel parallel to the centreline of the road or street. The characteristics of a well-designed sidewalk are (Boodlal, 2003) that it includes

- Adequate width
- Safety and security of use
- Continuity and connectivity
- Landscaping
- Social space
- Accessibility to ALL users regardless of impairments or handicaps

Adequate width is determined by the number of pedestrians to be accommodated and the extent of the other uses to which sidewalks may be put. For determination of the width of sidewalk required by pedestrians, readers are referred to Chapter 23 of the *Highway Capacity Manual* (Transportation Research Board, 2010).

A distinction is drawn between safety and security of use. Safety implies that pedestrians should not feel threatened by adjacent traffic and security refers to an absence of threat from criminal elements. The former is achieved by providing an adequate separation between pedestrians and vehicular traffic and the latter by ensuring that there aren't any convenient hiding places from which pedestrians could be attacked. The separation between vehicular and pedestrian traffic is achieved by the provision of landscaping as a buffer and also space for the planting of shade trees and flower beds and the location of benches.

Social space makes provision for participation in public life such as at sidewalk cafes, roadside vendors of foodstuffs or browsing through the wares on offer at flea markets. It is in fact the provision of social space that adds to the richness and character of the urban scene.

Accessibility as understood by the Americans with Disabilities Act is aimed not at some arbitrary percentage such as the 85th percentile pedestrians but at the 100th percentile pedestrians.

Sidewalks have to comprise distinct areas or zones, each of which is aimed at one or more of these objectives. The required width of the sidewalk is dictated by the totality of these characteristics. Four zones are required:

- The kerb zone, which is only about 0.1 to 0.2 metre wide, is required to accommodate drainage of the adjacent street. It keeps water from the roadway from flooding the sidewalk and also serves as a clear demarcation of the spaces provided for use by vehicles and pedestrians. Kerbs can take the forms of barriers intended to restrict the access of vehicles to the sidewalk area or may be mountable to allow for parking in the sidewalk area such as may be acceptable in residential areas. The kerb is a useful guide to pedestrians who are visually impaired.
- The landscaped zone has a minimum width of 1.8 metres on major streets and 1.2 metres on local streets and collectors. It provides space for plantings such as shade trees and flower beds and benches to ensure that there are no obstructions to walking as a mode of transport. Street furniture such as poles for street lighting and signage, fire hydrants, drinking fountains and telephone kiosks would be located in the landscaped zone. The proliferation of the cell phone may, however, cause the telephone kiosk to become an object of historical interest only. On bus routes, the landscape zone would have to include space for bus stops and shelters, typically 1.5 metres wide, although, as an intermittent object, space for these could be taken from the other objects located in the landscaped zone. In regions subject to heavy accumulations of snow, the width of the landscaped zone would be indicated by the need for storage of snow to keep streets and sidewalks snow-free.
- The pedestrian travel zone, with its width determined by the volume of moving pedestrians to be accommodated, should be free of all obstructions and have a minimum clearance profile of 2.5 metres height and 1.8 metres width. The minimum width is dictated by the space necessary for two wheelchairs to pass each other and the minimum height ensures that visually impaired pedestrians will not suffer head injuries while moving along the street.
- The verge or frontage zone is located between the pedestrian travel zone and the road reserve boundary. The minimum width of the verge is dictated by the fact that pedestrians cannot feel comfortable if walking with their shoulders continually brushing against a wall or shop window. A shy distance of at least 0.6 metre should be provided. The verge zone is often used by visually impaired pedestrians who use building faces as a guide when using a long cane or rely on the sound reflected from building faces as a navigational guide. If this use is at all frequent, the verge zone should be 0.6 to 1.2 metres wide. For fear of injury to passers-by, doors should not swing open into the verge zone.

The verge zone fulfils a valuable social function, as it would be the location of sidewalk cafes, flea markets and vendors of foodstuffs and other products, and would also include marketing of wares by shops forming the boundary of the verge zone.

The areas used by moving vehicles and pedestrians and the kerb zone are purely functional and sterile. The verge zone and the landscaping zone, on the other hand, provide the character of the entire street including determining whether it could be classified as a boulevard or not.

As discussed in Chapter 8, services such as water reticulation, sewerage, power supply, and so forth are located preferably in the verge.

14.2.5 Geometric design for pedestrians

Geometric design relates to sidewalk gradients and crossfalls. It also addresses the issues of the location, traffic control and detail of street crossings. In large measure, design is aimed at the accommodation of handicapped pedestrians. These include

- Elderly pedestrians
- Wheelchair users
- Visually impaired pedestrians
- Pedestrians with impaired hearing
- Children, whose only impairment is their youth

This focus ensures that if all handicapped individuals are accommodated, everyone else will also be accommodated.

14.2.5.1 Gradients

The gradients of sidewalks generally follow those of the adjacent street pavement. Provided the streets are not too steep, this is generally not a problem but, ideally, sidewalk gradients should not exceed 5 per cent. They could, however, be as steep as about 8 per cent but only for a distance of less than 9 metres. After this distance, a landing must be installed.

The landing should have a minimum dimension of 1.5 square metres and be located either in the landscaped zone or the verge zone to be clear of the pedestrian travel zone. If there are shade trees in the landscaped zone, this would be the preferred option but, otherwise, location in the verge zone would have the benefit of keeping resting pedestrians away from the exhaust fumes of passing vehicles.

14.2.5.2 Crossfalls

The Americans with Disabilities Act requires that crossfalls may not exceed 2 per cent. Anyone who has ever tried to maintain a reasonably straight path pushing an airport trolley loaded with heavy luggage across even as flat a slope as this will know just how difficult it is. The trolley tends to turn down the slope, and travel at right angles to the slope requires significant effort to maintain the desired direction of travel. Pedestrians in wheelchairs will find that severe crossfalls will cause them to veer towards the kerb and into the street.

The effect of severe crossfall can be compounded by coinciding with a steep longitudinal gradient. The slope that then has to be counteracted is the resultant of the two vectors: the cross-slope and the gradient.

14.2.5.3 Street crossings

Street crossings comprise kerb ramps and street markings defining the width of the crossing. The street markings include the basic zebra crossing of alternating black and white painted stripes.

If the street cross-section includes a median island, it is normal practice to provide a cut-through of the island at the level of the street surface. This is for the benefit of wheelchair users, people pushing prams and cyclists. Median islands generally have a nose, as illustrated in Chapter 8 and the cut-through is some distance behind the nose. This distance is selected to ensure that the length of the cut-through is 1.5 metres or longer and its width is equal to that of the crossing.

14.2.5.4 Driveway crossings

The purpose of driveway crossings is primarily to ensure safe passage of vehicles across a sidewalk to an adjacent property. Because of the very low speed of travel, vertical curves are not necessary and changes of gradient can be instantaneous. If the change of gradient is negative, as in a crest curve, the vehicle can drag its underside along the road surface whereas with a positive change of gradient, the front and rear bumpers (fenders in the United States) could hit the surface and be damaged. These problems can be solved by the application of simple geometry and are not dealt with further here.

Whatever the gradient is along the driveway, provision should be made for a length of about 2 metres which is at a gradient of not more than 2 per cent. This would, for pedestrians, be a cross slope and should, for preference, be located at the road reserve boundary. This is to allow for a vehicle coming off the street and stopping prior to the area where pedestrians may be encountered.

14.2.5.5 Ramps

In some cases, sidewalks could be so steep that they would have to be replaced by flights of stairs. Individual steps should have risers of not more than 190 millimetres and tread depths of not less than 250 millimetres. In this case, every 16th step – roughly a rise of 3 metres – would have to comprise a full width landing to allow for a rest area. Clearly, with a mean gradient of the order of 76 per cent, such sidewalks would not be adjacent to a street. They would be limited to exclusively pedestrian precincts.

Ramps may have a maximum gradient of 8 per cent, with 5 per cent being preferred. They would have to have a length of landing extending across the full width of the sidewalk after a horizontal distance of 37.5 metres had been traversed to achieve a height gain of 3 metres but with 30 metres being preferred regardless of the height of rise achieved. The adjacent flight of steps would require only 4.0 metres to climb through a height of 3.0 metres. The ramp and the stairs would thus have to be completely independent structures.

A height difference of 7.0 metres above the centreline of a road or street is necessary to assure the clearance of 5.7 metres recommended in Chapter 7. A ramp starting at the same level as the shoulder of the road would thus have to clear about 7.2 metres. At an average gradient of 8 per cent, the ramp would have a length of 90 metres and include two landings at least 1.5 metres in length located at the 30- and 60-metre distances up the ramp, that is, an overall length of at least 93 metres and possibly more if a significant volume of wheelchair traffic is expected to require a greater length of landings.

Space constraints normally require ramps to have a spiral configuration. In this case, the diameter of the inner balustrade of the ramp should be approximately 30 metres if the spiral is to encompass only one complete revolution. This is to match the requirements of maximum gradient and rest landings and suggests that the spiral ramp would almost certainly require a local widening of the road reserve.

14.2.5.6 Pedestrian crossings

A well-designed pedestrian crossing includes

- Visible and possibly raised crosswalks of an adequate width
- Kerb ramps with detectable warnings
- Information in the form of
  - Signs and markings
  - Traffic control signals where appropriate on the basis of vehicular and pedestrian traffic volumes
- Islands:
  - Median
  - Refuge
  - Corner islands
- Crossing times that are adequate in terms of both duration and frequency
- Kerb extensions (also known as bulb-outs)
- Adequate sight distance both for pedestrians and drivers

Raised crosswalks are usually in the form of speed humps with a flat top and are used as a traffic calming device. If raised not more than about 25 millimetres, they do not reduce vehicle speeds but provide visually handicapped pedestrians with a useful guide in crossing the street.

There is little or no uniformity worldwide on the form of traffic control that should be applied to a pedestrian crossing. The well-known zebra crossing comprising alternating black and white stripes requires motorists in some countries, but not all, to stop in the presence of pedestrians. Traffic signals may have pedestrian signage of the form of 'Walk/Don't Walk' or the symbols of a standing and a walking man. The walking man is shown in green and the standing man in red.

The United Kingdom has a host of rather fancifully named traffic control devices. These include

- **Pelican** (**Pe**destrian **Li**ght **Con**trolled) crossings with red/yellow/green signals facing drivers and red man/green man signal heads on the opposite side of the road facing the pedestrians waiting to cross
- **Puffin** (**P**edestrian **U**ser-**F**riendly **In**telligent) crossings. where crossing time is established on a cycle-by-cycle basis by on-crossing detectors
- Toucan crossings, designed for both pedestrians and cyclists, hence 'two can', typically used adjacent to a cycle path
- Pegasus crossings, which are similar to Toucan crossings but have a red/green horse symbol replacing the cyclist and, in addition to the normal mounting height of push button, also have higher mounted push buttons for horse-riders

In the United States, Richard Nassi invented the High Intensity Activated Crosswalk, or HAWK beacon (Fitzpatrick and Eun Sug Park, 2010). This form of control, referred to in the *Manual on Uniform Traffic Control Devices* (MUTCD; Federal Highway Administration, 2009) as the pedestrian hybrid beacon, is used at intersections with Stop control on the minor legs which ensures that unwanted traffic isn't attracted to the major legs. Control of vehicles is by means of a signal configuration of two red lenses over a yellow lens as shown in Figure 14.3.

Traffic control of pedestrian crossings is usually incorporated into urban traffic control systems where these have been implemented. Control can also be by means of push buttons activated by waiting pedestrians. Push buttons are normally located as closely as possible

Figure 14.3 The HAWK traffic control system. (From Fitzpatrick K and Eun Sug Park. *Safety effectiveness of the HAWK pedestrian crossing treatment*. Federal Highway Administration, Washington, DC, 2010.)

but no closer than 750 millimetres to the kerb ramp at a maximum height of 1 metre from the ground for the benefit of wheelchair and no further than 1.5 metres from the crosswalk.

14.2.5.7 Kerb ramps

To cross a street, wheelchair users will require kerb ramps to be provided. These ramps must also be such that they do not confuse visually handicapped pedestrians. Various types of ramps can be considered, including

- Perpendicular
- Diagonal
- Parallel
- Combination

A fifth type has the entire corner depressed to match the street level. This is not recommended, as visually handicapped pedestrians will have no guidance regarding where the crossing point actually is and orientation to a right angle crossing of either of the streets would be impossible to determine.

The basic components of a kerb ramp that is wheelchair friendly and also offers guidance to visually handicapped pedestrians are shown in Figure 14.4. The blocks at the bottom of the ramp serve as a detectable warning for visually impaired pedestrians. They comprise truncated domes with

- A base diameter of 23 to 26 mm
- A top diameter about half that of the base diameter
- A height of 5 millimetres
- A spacing centre to centre of 40 to 60 millimetres

The domes span the full width of the ramp across a lateral distance of about 600 millimetres.

The four basic kerb ramp types listed in the preceding are illustrated in Figure 14.5. There are advantages and disadvantages attached to each of these ramp types and these are listed in Table 14.2.

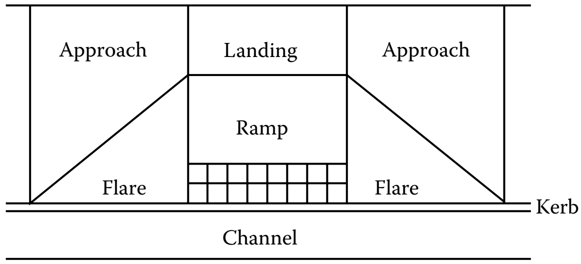


Figure 14.4 Components of a typical kerb ramp. (From Boodlal L. *Accessible sidewalks and street crossings*. Federal Highway Administration, Washington, DC, 2003.)

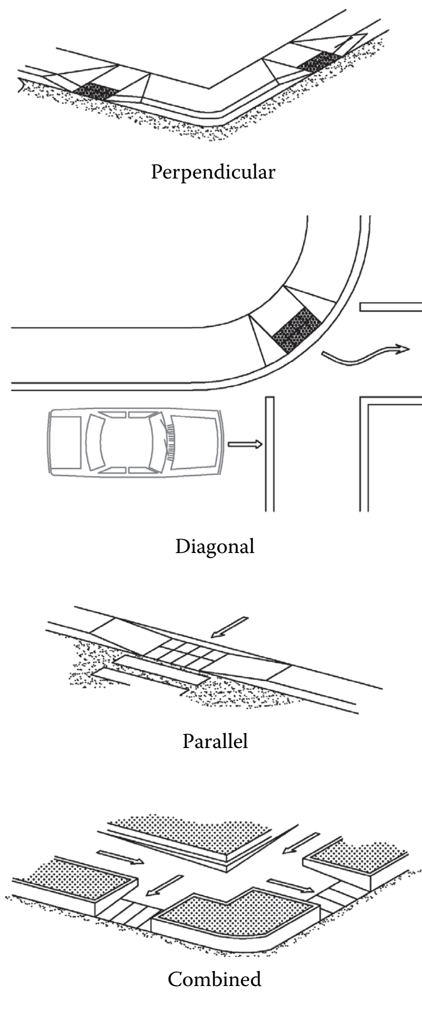


Figure 14.5 Basic kerb ramp types. (From Boodlal L. *Accessible sidewalks and street crossings*. Federal Highway Administration, Washington, DC, 2003.)

Table 14.2 Comparison of kerb ramp types

Ramp type	*Advantages to pedestrians*	*Disadvantages to pedestrians*
Perpendicular	Ramp aligned with crosswalk. Straight path on tight radius. Two ramps per corner.	May not provide a straight path of travel on large radius corners.
Diagonal	Not recommended.	Visually impaired pedestrians may mistake a diagonal ramp for a perpendicular ramp and unintentionally travel into the intersection area. May conflict with motorists travelling straight or turning if corner radius is small.
Parallel	Requires minimum road reserve width. Provides an area to align with crosswalk. Allows the ramp to be extended to reduce ramp gradient. Provides edges on the side of the ramp that are clearly defined for visually impaired pedestrians.	Pedestrians need to negotiate two ramp gradients, which is difficult for wheelchair users. Improper design may result in water and/or debris accumulating on the bottom landing of the ramp.
Combination	Does not require turning or manoeuvring on ramps. Ramp aligned perpendicular to the crosswalk. Level manoeuvring areas at the top and bottom of ramps.	Visually impaired pedestrians need to negotiate sidewalk ramps.
Depressed corners	Eliminates the need for kerb ramps.	The illusion may be created that the sidewalk and street are a combined pedestrian space. Improper design may cause trucks to travel onto the sidewalk in making tight right turns. More difficult for the visually impaired to detect the boundary between the sidewalk and the street. Service dogs may not distinguish the boundary between the sidewalk and the street and continue walking. Motorists may be encouraged to turn faster by traveling onto the sidewalk.

Source: Boodlal L. *Accessible sidewalks and street crossings*. Federal Highway Administration, Washington, DC, 2003.

14.2.5.8 Sidewalk and street crossing surfaces

Sidewalks should, as a general rule, be paved. Gravel or earth sidewalks become muddy when rained upon, resulting in pedestrians preferring to walk on the travelled way if this is surfaced. Unfortunately, this places pedestrians in a hazardous situation, with the loss of safety including being drenched by spray kicked up by the tyres of passing vehicles. During rainstorms, pedestrians also become less visible to passing motorists.

The sidewalk surface should preferably be of bitumen or concrete. Other surfaces such as brick or paving stones, however, provide a very attractive appearance, enhancing the overall aesthetics of the road or street. The only requirement is that the gaps between the individual paving stones should not be sufficiently wide to trap the wheels of a wheelchair, pram or stroller. Racing bicycles have very thin tyres. If a tyre falls into a gap between stones, the rider could fall off the bicycle and be injured.

Also, for the safety of cyclists there should be no loose stones on the sidewalks. Hitting these could also cause riders to become unseated. A pedestrian stepping on a loose stone may suffer a twisted ankle and fall. In the case of elderly pedestrians, the possibility of a consequent fractured hip cannot be overlooked.

14.3 PROVISION FOR CYCLISTS

14.3.1 Introduction

In developing countries, for many people the bicycle is the only form of personal mechanised transport. For long distances, people are dependent on public transport. In First World countries, cycling is a useful form of exercise and can also be the answer to unacceptable levels of congestion. As previously suggested, Americans do not walk much. Neither do they ride bicycles. This is clearly reflected in statistics on body mass index (BMI).

The BMI is the body mass in kilograms of an individual divided by the square of the person's height in metres. BMI measurements are used to assess whether a person's body weight is considered to be within a normal, overweight or obese range, as shown in Table 14.3.

Applying this indicator to the populations of various countries indicates that Americans are the most obese of the various populations worldwide, closely followed by those of Great Britain, as indicated in Figure 14.6. It is suggested that cycling is an excellent form of exercise that can improve an individual's BMI.

14.3.2 The cyclist as a design vehicle

As is the case with pedestrians, cyclists also have to be defined in terms of the space they occupy. These are shown in Figure 14.7.

The longitudinal dimension applies to the width required of a median island to ensure that the bicycle does not protrude into the adjacent lanes, thus presenting a hazard to passing vehicles. Three values are offered for the lateral width of a cyclist:

- 0.6 metre Stationary
- 1.0 metre Minimum manoeuvring space
- 1.5 metres Comfortable moving space

The comfortable moving space is required for two cyclists to be able to pass each other or ride side by side, hence implying a path or lane width of 3 metres.

Table 14.3 Body mass index measurements and interpretation

Body mass index	*Interpretation*
<18.5	Underweight
18.5 to 24.9	Healthy weight
25–29.9	Overweight
30–40	Obese
<40	Severely obese

Source: National Heart, Lung and Blood Institute. *Body mass index*. Bethesda, MD, undated.

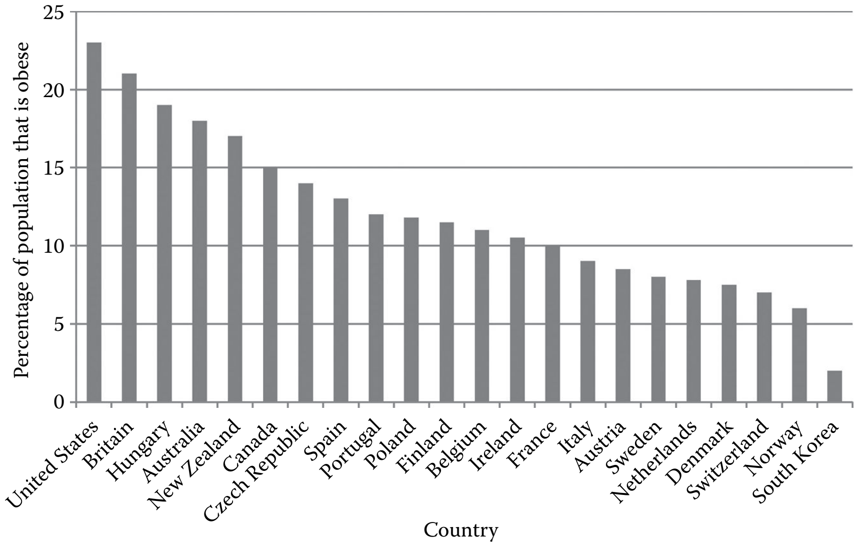


Figure 14.6 Worldwide percentages of the overweight population. (From Mauroit C, Ghaman R, Kamali F, Kurosaka T, Talens H, Gerlach J, Bezak B and Prat D. *Human powered transport*. World Road Association [PIARC], Paris, 2008.)

14.3.3 The performance of cyclists

Research shows that in Western countries one hour of travelling time to work is a generally accepted maximum. On the basis of assumed speeds of 2.5 km/h for walking and 20 km/h for cycling, it follows that the maximum distance for walking to work is 2.5 kilometres, and cycling to work has its maximum at 20 kilometres.

14.3.4 Geometric planning for cyclists

Town planning was historically aimed at single uses for defined areas. Employment opportunities were focussed on offices in the central business district (CBD) and in heavy and light industrial areas. None of these were allowed in residential areas. Given the ranges of distances suggested previously, mixed land uses have much to recommend them. Walking or cycling to work automatically generates the exercise needed for health reasons. Furthermore, taking people out of their motor cars reduces exhaust emissions, hence creating a healthier environment.

Planning for dedicated cycle lanes as part of the urban road cross-section and also for dedicated cycle paths removed from the lanes intended for motorised vehicles make cycling a more attractive and safer proposition. These automatically result in the benefits to health from getting exercise, improvement of the environment by reduction of exhaust emissions from motorised vehicles and significantly less vehicle clutter given the difference in size between a passenger car and a bicycle.

People may elect to cycle long distances as a means of getting exercise but prefer shorter distances, typically of the order of 5 kilometres or less, for commuter trips. Urban planning should therefore seek as far as possible to create trips between origins and destinations of this length. Cell development whereby an area surrounded by arterials or freeways provides a complete package of residential space, recreational space, schools and employment opportunities would create an environment in which people would wish to live, work and play. This approach to town planning thus has much recommend it.

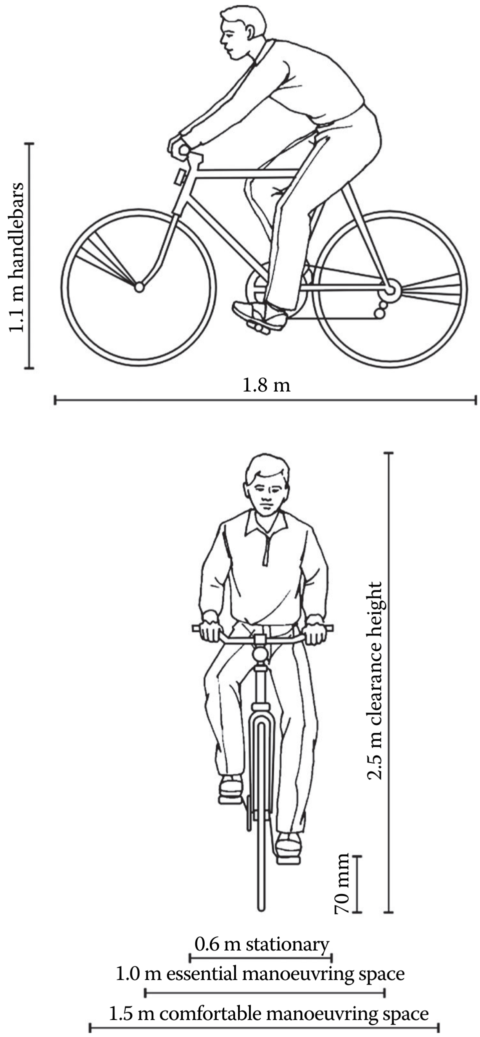


Figure 14.7 The dimensions of a cyclist. (From South African Department of Transport. *Pedestrian and bicycle facility guidelines Draft 1.0*. Pretoria, 2002.)

14.3.5 Geometric design for cyclists

The geometry of cyclists' lanes immediately adjacent to the traffic lanes is dictated by the geometry of the travelled way in terms of gradient and horizontal and vertical alignment. In the case of cycle trails and paths, that is, alignments independent of highways for motorised vehicles, geometric design requires consideration of

- Stopping sight distance required by cyclists to perceive a hazard and to stop safely
- Gap acceptance required to enter or to cross a street in gaps in opposing traffic

It is recommended that AASHTO's *Guide for the Development of Bicycle Facilities* (2012) should be acquired by designers who are designing these facilities. These and other matters are dealt with in far more detail in that document than can be given justice to in a few

Figure 14.8 The Segway.

paragraphs. It also introduces the Segway illustrated in Figure 14.8 in detail as a design vehicle.

14.3.5.1 Design speed

AASHTO's *Guide for the Development of Bicycle Facilities* specifies a minimum design speed of 30 km/h. The 85th percentile speed of cyclists is 22 km/h, which suggests that the 30 km/h design speed may be generous. Recumbent cyclists have an 85th percentile speed of 29 km/h and hand cyclists an 85th percentile speed of only 8 km/h.

14.3.5.2 Stopping sight distance

Stopping sight distance is based on a range of design speeds, reaction time of 2.5 seconds and an acceleration rate of –2.5 m/s^2 being applied to the relationship

$$\mathrm{SSD} = v \quad 0.278t + 0.039\frac{v}{a}$$

where

v = speed (km/h)

t = reaction time (seconds)

a = acceleration rate (m/s^2)

Table 14.4 Stopping sight distance for cyclists

Bicycle design speed (km/h)	*Stopping sight distance (metres) for gradients of (%)*						
	−15	*−10*	*−5*	*0*	*5*	*10*	*15*
20	30	25	25	20	20	20	30
30	55	45	40	35	35	35	30
40	90	70	60	55	50	50	45
50	130	100	85	75	70	65	60

Source: South African Department of Transport. *Pedestrian and bicycle facility guidelines Draft 1.0*. Pretoria, 2002.

On downgrades of 4 per cent or steeper a design speed of 50 km/h should be considered. If a cycle path is unpaved, the possibility of there being loose material on the surface would cause cyclists to proceed with caution and a design speed of 25 km/h may be adequate. Sight distances appropriate to various design speeds are shown in Table 14.4.

14.3.5.3 Horizontal alignment

Recumbent cyclists with their 85th percentile speed of 29 km/h can tolerate a minimum radius of horizontal curvature of 26.8 metres. It is suggested that horizontal alignment design thus be based on a minimum radius of 30 metres.

While wheelchair users accept only a low rate of lateral acceleration, their 85th percentile speed suggests that a radius of horizontal curvature of 12 metres would be adequate. The selection of this radius would, however, suggest that cyclists should be precluded from access to such sidewalks.

14.3.5.4 Vertical alignment

The parabola has been adopted internationally as the preferred shape of a vertical curve. The basic form of the parabola is expressed by the relationship

$$y = ax^2 + bx + c$$

So that the gradient, dy/dx, at any point, x, on the curve is given by

$$\frac{dy}{dx} = 2ax + b$$

and the rate of change of curvature is the constant

$$\frac{d^2y}{dx^2} = 2a$$

It is convenient to express the constant as

$$K = \frac{L}{A}$$

If the length of the vertical curve is greater than the sight distance ($L < S$), the minimum K-value of the crest curve is given by the expression

$$K = \frac{S^2}{100\left(\sqrt{2h_1} + \sqrt{2h_2}\right)}$$

In the second case, the relationship is

$$K = \frac{2S}{A} - \frac{200\left(h_1^{0.5} + h_2^{0.5}\right)^2}{A^2}$$

where
S = stopping sight distance for selected design speed and object height (m)
h_1 = cyclist eye height (m)
h_2 = object height (m)
A = algebraic difference in gradient between the approaching and departing grades

The first of the two relationships offered above is more convenient to calculate than the second, and using it also in the case of the sight distance being longer than the curve does not result in significant errors.

The values of K for the range of design speeds and associated stopping sight distances offered in Table 14.4 are shown in Table 14.5.

14.3.5.5 Lane and path width

A lane forming part of the travelled way but reserved for cyclists may be as little as 1.2 metres wide if intended to function as a single one-way lane.

If a path separate from the travelled way is to be wide enough to permit either two cyclists riding side by side or cycle traffic in both directions, its minimum width should be 3.0 metres.

The Americans with Disabilities Act Accessibility Guidelines (ADAAG) requires 0.815 metre (32 inches) at a point and 0.915 metre (36 inches) continuously for passage by a single wheelchair. The minimum width of path to allow two wheelchairs to pass each other is, according to ADAAG, 1.525 metres (60 inches).

Inline skaters do not move in a straight line but rather in a series of sweeps from side to side. Their 85th percentile sweep width of 1.5 metres would only just fit a 3.0 metre width path.

14.3.5.6 Gap acceptance

Consideration of gap acceptance applies to intersection design. The left turn is arguably the most complex manoeuvre that a cyclist has to execute. In the presence of heavy vehicular

Table 14.5 K-value of crest curves for various design speeds

Design speed (km/h)	K-*value of crest curve*
20	2
30	7
40	14
50	32

traffic the wisest move would probably be to dismount at the intersection and cross the road as a pedestrian. Gap acceptance by cyclists is modified by traffic density (Plumert et al., 2007). At high densities they are prepared to accept smaller gaps than otherwise would be the case. The research indicated that children and adults are prepared to accept similar sized gaps. If the 'critical gap' is assumed to be the gap accepted 50 per cent of the time, it would be in the range of four seconds. Approximately 80 per cent of gaps of 5 seconds would be accepted. It is suggested that, for design purposes, a gap of 5 seconds or less would require traffic control by signalisation. Readers are referred to the discussion of the Poisson distribution in Chapter 17 for the required analysis.

Chapter 15

Public transport

CONTENTS

15.1 INTRODUCTION

Urban traffic congestion is a manifestation of an imbalance in the supply/demand equation, with demand typically exceeding supply. It can thus be addressed either by providing more infrastructure (supply) or by utilising modes more efficient as people movers than the passenger car (demand). There are limits to which provision of infrastructure can be employed as a solution to the problem and, in many cities across the world, these limits have already been reached. Geometric planning and design must provide an infrastructure supportive of public transport in order to offset the undoubted convenience of the passenger car.

15.2 HISTORICAL OVERVIEW

Public transport is its own worst enemy. It is public transport that made it possible for people to live in the suburbs and work in the city centres. The resulting low population density caused public transport to become uncompetitive and in need of subsidisation to function at all. All trips other than the 15 per cent at the peak hour that constitutes the home-to-work trips had to be undertaken by private passenger cars. Many of the home-to-work trips were also undertaken via passenger car.

Throughout history most people worked at home and many still do. With the coming of the Industrial Revolution, there was a separation between the home and factories or workshops but most trips were by foot and of short distance (Wachs, 2014). Dock labourers lived within walking distance of the ports and steel workers near the mills. Clothing was largely made by hand and workers walked to small local shops, often taking their work home with them. Trips, although short, were not pleasant. Although sea coal produced a remarkably smoke-free burn, it generated a fair amount of carbon monoxide as well as some of the sulphurs. The streets were full of garbage and manure from the horses pulling coaches and wagons.

A coach builder by the name of George Shillibeer was offered work in Paris, where he was commissioned to build some unusually large horse-drawn coaches of 'novel design'. The aim was to design a coach capable of transporting a whole group of people, perhaps two dozen, at a time. He came to the conclusion that similar vehicles in London, but for the fare-paying public with multiple stops, would be a paying enterprise, so he returned to his native city. His first London Omnibus took up service on 4 July 1829.

Only the rich could afford the fares and they, being professional people and merchants, moved their homes to the outskirts of the cities to secure the benefit of more space and cleaner air although their offices remained in the city centres. Rails were placed in the streets, hence reducing the number of horses required to move the buses and, in the 1830s, the horse-drawn carriages were replaced by steam buses. In 1882, electric trolley buses powered from overhead cables were introduced. The first internal combustion engine buses were used in 1895. This coincided more or less with the invention of passenger cars by Benz.

These innovations reduced the cost of travel and made it available to many more workers. The rush to the suburbs was on and a natural geographic division between home and workplace took place. Land use planning focused on single usage and job opportunities were excised from residential or 'dormitory' suburbs. Railway companies 'commuted' the fares of their passengers, enabling them to travel in both directions – home to work and back again – on a single monthly ticket, hence adding the word 'commuter' to describe those people travelling on a regular basis between home and work (Wachs, 2014).

Public transport had made it possible for people to live in pleasant if not actually idyllic surroundings while working in densely populated areas. The increasing availability of passenger cars has, however, resulted in more people travelling to work by passenger car. The outcome is the congestion that brings entire cities to a gridlocked standstill.

15.3 NETWORK CONSIDERATIONS

The pendulum has now swung back. Where urban planning previously actively pursued a policy of decentralisation, this resulted in public transport becoming uneconomic. The problem was that buses had to wind for long distances through low-density residential areas to acquire a full load of passengers whereas, to be economic, buses should be filled as quickly as possible and then run as fast as possible to the desired destinations of the majority of passengers. The planning emphasis now falls on densification by the creation of mixed usage areas. Each area should contain

- Residential accommodation
- Schools

- Places of worship
- Sporting facilities
- Open spaces in the form of parks, playgrounds and the like

These areas would thus be, in effect, miniature villages in the larger urban area and could result in the development of a strong community spirit.

Mixed usage makes it possible to meet more job-related needs by walking or cycling or public transport, hence lessening the need for travel over long distances to get to work.

This is not to say that what is intended is the creation of a fortress mentality. There will still be fairly large numbers of people commuting to job opportunities elsewhere. These are the people who should preferably be accommodated by the provision of public transport.

As stated in Chapter 14, research shows that in Western countries 1 hour travelling time to work is a generally accepted maximum. On the basis of assumed speeds of 2.5 km/h for walking and 20 km/h for cycling, it follows that the maximum distance for walking to work is 2.5 kilometres and cycling to work has its maximum at 20 kilometres. If the home-to-work trip includes more than one mode, it follows that the maximum walking distance has to be reduced. It is suggested that 30 minutes of walking combined with 30 minutes would probably be acceptable. If the home-based work trip has a longer duration than about 90 minutes, this would be a compelling reason for relocation closer to the work place. Alternatively, the home environment must be so attractive that commuters are prepared to accept the penalty of a long work trip.

Assuming a practical walking distance of, say, 1.25 kilometres to the nearest bus stop, this would, in the case of a conventional grid layout, have two components, one parallel to the bus route and one at right angles to it. If the plot widths fell in the range of 50 to 200 metres it follows that the walking distance as the crow flies is 206 metres or less.

In South Africa, the apartheid regime moved all black people away from the cities with a policy of 'White by Night'. This placed an intolerable burden on the poor, who were suddenly confronted by enormous costs of commuting in terms of both money and time. Many people complained that they had to travel for 9 hours a day to do an 8-hour job. The post-1994 government has not been able to solve the problem but what has been happening is that the more affluent black people are buying property in previously exclusively white areas. Working class black people are also drifting towards the urban areas but have to settle for squatting in shacks on the outskirts of the urban areas.

Where a third mode, be it travel by air or by train, enters the picture, this model of commuting distance is of no use. In most international airports, walking distances from a parked car to getting checked in and then a further distance to arrive at the boarding gate is substantial and there is, of course the obligatory wait of a half-hour or more before boarding the aircraft.

Reverting to a walking distance of 206 metres or less, it follows that a grid could be constructed comprising parallel collectors at a spacing of 350 metres. This would ensure a reasonable walking distance to the nearest bus stop. A similar spacing is suggested for the terminus.

The geometry of local access streets could include steep gradients, narrow cross-sections and very small kerb radii at intersections. Passage of a bus through a local street system could thus include its grinding its way up steep hills, swinging wide into opposing lanes to complete its turn and generally suffer a protracted journey time to pick up a full complement of passengers. It follows that bus routes should, as far as possible, be kept off local streets.

Arterials can be included in bus routes provided they don't include bus stops. As an alternative, the bus rapid transport (BRT) system could be applied. This would include bus routes that are effectively not part of the arterial system as such on arterials but with passenger access being provided from the crossing roads to stations within the median islands. BRT systems are discussed further in Section 15.6.

15.4 MODAL TRANSFER STATIONS AND TERMINI

15.4.1 Introduction

Various modes of transport exist. These include travel by air or on water and the various forms of land transport by rail or by road. Of these, road transport does not require support by any other mode whereas the other modes all depend on road transport to some greater or lesser extent. For air travel to be effective, airports need to be provided with links by rail or by road to the land uses to which they provide service. Similarly, sea ports and other inland water-based transport also require links by rail or by road to the communities being served. Finally, travel by rail requires the support of road-based transport at the various railway stations.

Modes of road-based transport can be walking, cycling, private passenger car or public transport for people or truck for the movement of freight. Implicit in the previous paragraph is the fact that no trip can be by means of a single mode. The simplest of all trips, the commute from home to workplace using the worker's own passenger car, starts with (presumably) a short walk from the home to the garage. The intermediate part of the trip by passenger car terminates at a basement or parking garage and is followed by walking or any other convenient mode between the parked car and the workplace.

It is clear that the importance of modal transfer is such that, if it is not properly designed and catered to, the entire transportation network could fail.

15.4.2 Transportation hubs

A transportation hub is a place where passengers or freight are exchanged between vehicles or between transport modes. A convergence of two or more modes of public transport is thus normally considered to be a transportation hub. For example, a bus station located at a railway station would be considered as a transportation hub. The design of transportation hubs is normally a complex exercise because it is necessary to accommodate two or more different types of vehicles, each with its own dimensions and operational requirements, in close proximity to one another and usually in a fairly restricted space.

As a general rule, it is necessary to separate the operations of the various classes of vehicles to ensure the efficient operation of a transportation hub.

15.4.3 Park-and-ride facilities

In the example of a park-and-ride facility at a bus station, there should be separate road networks for each class of vehicle and the entrances and exits to each should preferably be from different links in the road network. Buses should enter and exit the park-and-ride facility from an arterial, almost certainly with the assistance of vehicle-actuated signal control, whereas passenger cars should enter and exit their parking area from a street that is possibly lower in the network hierarchy.

Figure 15.1 illustrates a possible layout for a transportation hub involving a modal transfer between passenger cars and buses.

The layout shown is utilitarian in the extreme and would be vastly improved by landscaping and the planting of numerous shade trees in the passenger car parking area.

An operating park-and-ride facility is illustrated in Figure 15.2.

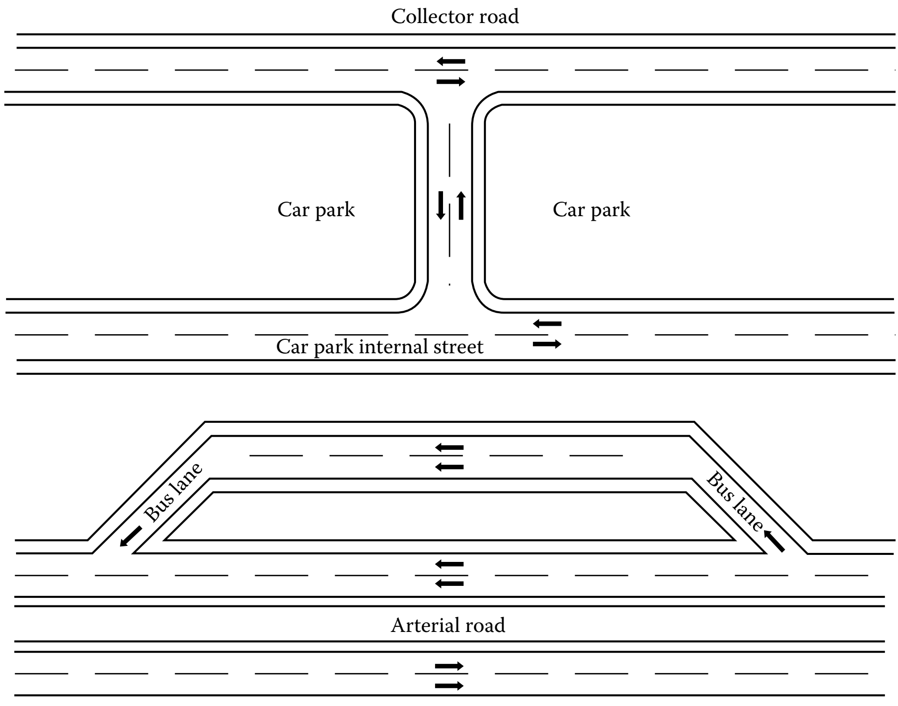


Figure 15.1 A typical layout of a park-and-ride transportation hub.

Figure 15.2 Pear Tree Park and Ride car park, Oxford, United Kingdom.

15.4.4 Kiss-and-ride or drop-off points

The principle of separation of vehicle types applies also in the case of kiss-and-ride hubs. A passenger in a private passenger car is brought to a bus stop and the dropping off vehicle then moves away, making space for another vehicle to drop off its passenger(s). Parking is literally for a matter of seconds only so that there is no need for large parking areas. The need here is for generously designed off-and-on ramps and relatively short areas where passenger cars can pause long enough for their passengers to disembark. The nomenclature presumably stems from a wife bringing her commuter husband to the bus stop and then returning home.

A typical kiss-and-ride hub is illustrated in Figure 15.3. Signage for kiss-and-ride facilities is illustrated in Figure 15.4.

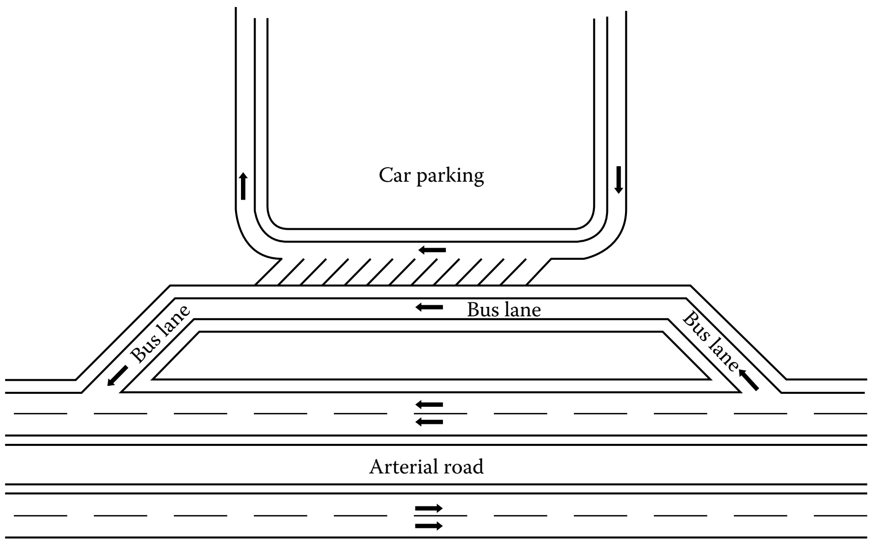


Figure 15.3 A typical kiss-and-ride or drop-off facility.

Figure 15.4 Signage for a kiss-and-ride facility.

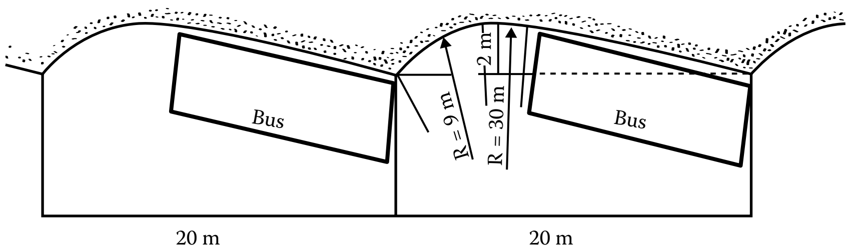


Figure 15.5 Sawtooth configuration for bus parking areas. (From American Association of State Highway and Transportation Officials [AASHTO]. *Roadside design guide*, 4th ed. Washington, DC, 2011b.)

15.4.5 Bus stations

Bus stations are the termini where various bus routes converge. These bus routes will almost certainly differ in the volume of passengers they have to accommodate. A reasonable frequency of service should be maintained on all of them to ensure the attractiveness of public transport to commuters, and this implies that different sizes of buses may need to be used. The alternative is to use a one size fits all approach and having large buses running virtually empty during off-peak times.

The design of the bus stations should thus be such that they can accommodate vehicles of different sizes and with different operating characteristics. However, buses can vary enormously in size, from the 13- and 16-seater minibuses, to the 25-seater midibuses, the conventional municipal 80-seater bus and the double decker bus carrying up to 95 passengers. Double decker buses are usually shorter than standard municipal buses so that there is relatively little difference in their carrying capacity. Articulated buses are usually 18 metres in length compared to the 11- to 14-metre length of the standard rigid chassis buses. The articulated bus can also take the form of a bi-articulated bus with two trailer sections as opposed to the one of the normal articulated bus. Bi-articulated buses have an overall length of about 25 metres and a carrying capacity of 200 passengers.

Very often, the parking bays of a bus terminus have a diagonal or sawtooth configuration as illustrated in Figure 15.5. This provides ease of manoeuvring and parking for the bus traffic and shorter walking distances between the booking hall and the selected bus for passengers. This configuration, however, results in the length of the various buses having an influence on the length of the bay when measured at right angles to the kerb line.

It follows that where different sizes of bus have to be accommodated, their parking areas should be separated to provide standard operating condition for all bus sizes.

The sawtooth configuration requires that buses have to reverse either into or out of the parking bays. The normal operation is reversing out of the parking bay. Reversing an articulated vehicle requires high skills levels. Bi-articulated vehicles are, to all intents and purposes, impossible to reverse in a controlled fashion and need to be towed out of their parking bays.

15.4.6 Bus stops and shelters

Bus stops are provided along a route at points where passengers would choose to alight from or board buses. Sometimes a bus stop would accommodate very few people but where about five or more passengers are likely to board a bus, it may be decided by the local authority that a shelter should also be provided.

In urban areas, bus stops would normally be located at intersections. A mid-block location would usually require would-be passengers to walk an unnecessary distance. Bus stops should normally be located immediately downstream of the intersection. This is because a bus parked at an upstream location would create a serious obstacle to sight distance to the right. The driver of a right-turning vehicle would be totally unsighted with regard to passengers who have just disembarked from the bus and would have to creep forwards to a point where it would be possible to see past the nose of the parked bus to check for the possibility of passengers crossing the cross-street or the street on which the driver is.

There is also the possibility that the bus could seek to merge with vehicles in the right-hand lane at the time when the driver of a following vehicle would be seeking to pass the bus with a view to executing a right turn.

The left turn with a bus parked in the bus stop does not present a problem to the drivers of any other vehicles. It does, however, become a challenge for a left-turning bus driver who would have to exit the bus stop and cross possibly several lanes of opposing vehicles without having much space within which to complete all these manoeuvres. A downstream bus bay would be the preferred option. This would also allow a bus driver to execute the necessary lane changes before the next intersection.

Bus bays are not normally constructed in 60 km/h zones. This is because, in urban areas, bus routes are normally located on collector streets and higher in the hierarchy of urban streets, where traffic volumes can be fairly high even outside peak times. A bus stopped in a bus bay could thus find itself trapped by opposing traffic and be in position of having to bulldoze its way back into the stream traffic. Stops are thus located in the travelled way. Following vehicles are thus required to stop if passing opportunities do not present themselves.

Bus bays may, however, be considered if the stop is very close to the departure side of a signalised intersection in a way that would severely impact intersection operation. They may also be considered where the stop is

- Used as a timing point, where buses may need to wait for several minutes if running early
- Used as a bus driver changeover point, requiring the bus to stop for longer periods
- A particularly high loading point, where the time taken to load passengers can regularly take minutes
- In a rural area as illustrated in Figure 15.6

An urban bus bay is shown in Figure 15.7 for an articulated bus. In this case, the overall length increases to 65 metres, with a straightening distance of 26.0 metres, entry taper of 19 metres and exit taper of 20 metres.

In urban areas with on-street parking, a 20-metre no-parking zone should be provided on both sides of the bus stop to enable the bus driver to clearly see the entrance to the bus stop and any waiting passengers and to manoeuvre easily in and out of the bus stop. If articulated buses are used on the route, the length of the no-parking zones should be increased to 30 metres.

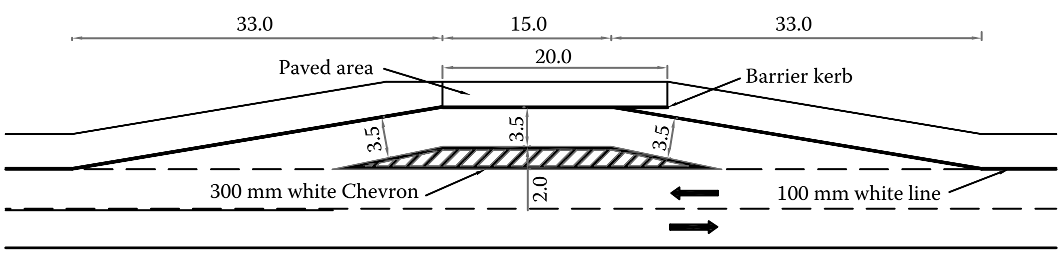


Figure 15.6 The geometry of a rural bus stop.

Figure 15.7 A typical urban bus bay.

On a surfaced road, the bus stop should also be surfaced, preferably with paving blocks of a colour different from that of the paved surface of the road. The contrasting colour highlights the difference in function between the bus stop and the through lane so that motorists would be less inclined to use the bus stop as a passing lane. Bitumen is dissolved by fuels and oil and paving blocks will guard against fuel and oil spill. Very often, a portion of the through lane is demarcated as a bus stop. In this case, the bus stop area should clearly be indicated as such by the use of a contrasting colour, which will also support protection of the road surface against fuel and oil spills. Road marking paints often have a low skid resistance, especially when wet, and it is suggested that the painted surface should be constructed with aggregate coarser than that of the surrounding road surface.

Ideally, the passenger waiting area should be paved to provide an area where waiting passengers can stand and embark the bus free of mud or dust. Kerbing would clearly demarcate the boundary between the waiting area and the bus stop and would also provide a modicum of protection from buses encroaching on the waiting area. The paved waiting area should have a slope of between 0.5 per cent and 3.0 per cent away from the road surface to prevent water standing in the area where passengers are attempting to board a bus.

Where bus shelters are provided, there should be a clear panel or opening on the approach side of the shelter. This will make it easier for passengers to see an approaching bus and also make it possible for the driver to see that there are passengers waiting to embark. Seating is normally provided for waiting passengers. Bus shelters are normally targets of vandalism and they should be constructed of vandal-proof materials such as concrete.

15.5 DEDICATED LANES

15.5.1 Introduction

There is a natural temptation for commuters to prefer to use their own vehicles. The passenger car is a true point-to-point vehicle that provides an unscheduled service as and when required. It provides space for carrying small packages and the seating is usually significantly more comfortable than on a bus. The passenger car is usually not required to stop to pick up other passengers and is thus faster between origin and destination than a bus. Unfortunately, the passenger car is not without its penalties. With an average occupancy of between one and two passengers compared to the number of commuters who can be transported by bus, the use of passenger cars directly results in congestion and, in the extreme case, gridlock.

What is required to wean the commuter from use of his or her own passenger car is to reduce the difference in convenience between the bus and the passenger car sufficiently for the commuter to change his or her commuting preference from passenger car to bus.

In this section, various strategies to increase the popularity of the bus are discussed. These include

- High-occupancy vehicle lanes
- Exclusive bus lanes
- Managed lanes

Ultimately, all of these, in combination with others relating to comfort levels and convenience, result in the relatively new concept of BRT.

15.5.2 High-occupancy vehicle lanes

High-occupancy vehicle (HOV) lanes are intended to increase the average occupancy of vehicles on the road. Their function, in fact, is to move people rather than vehicles. The use of theses lanes is limited to vehicles carrying more than a specified number of occupants, typically two or three including the driver. Some local authorities also permit usage of HOV lanes by

- Motorcycles
- Emergency and law enforcement vehicles
- Low-emission and other 'green' vehicles
- Single-occupancy vehicles on payment of a toll

The HOV concept has been widely applied in the United States but significantly less so in Europe. This is because the United States, with its widespread suburban sprawl, encourages the use of passenger cars whereas public transport is far more prevalent in Europe. It is noted that many urban dwellers in Europe do not, in fact, feel the need to possess cars at all.

HOV lanes can be created by building additional lanes on existing multi-lane roads or by the simple expedient of taking one existing lane in each direction and declaring it to be an HOV lane. This requires a minimum of three lanes in each direction. With only two lanes in each direction, the HOV lane would constantly be 'invaded' by drivers wishing to overtake slower vehicles in the general-purpose lane. The lane to be selected as an HOV lane is usually the one closest to the median. It is intended to function as a high-speed lane in as well as a high traffic volume lane whereas the other lanes would be subject to the turbulence of merging and diverging traffic and also have slow-moving trucks in the traffic stream.

To be effective, HOV lanes require ongoing law enforcement. Many people have tried to circumvent the provisions of law enforcement, for example, by using inflatable dolls as 'passengers' but the outlay in the purchase of the dolls and the size of the fine imposed when caught tend to make this not particularly effective.

The 'slug line' serves the same purpose but with real people. Slugging is an interesting form of symbiosis. It is a variant of hitchhiking. Students heading along the George Washington Parkway to the University of the Americas provide the number of passengers that a solo motorist requires to be able to enter the HOV lane. He or she benefits from the shorter drive time and the students get to the university at zero transport cost.

The criticism has been raised that the HOV lanes are underutilised and do not adequately offset the delays suffered by traffic in the other lanes. Some states are now offering a tolling facility on HOV lanes whereby a driver who does not otherwise qualify for access to the HOV lane can use it on payment of a fee.

15.5.3 Exclusive bus lanes

Exclusive bus lanes are lanes that are reserved at certain times of the day for use by buses only. They are usually located in the median island and are often reversible. Examples exist, however, of kerbside lanes being utilised as exclusive bus lanes. Exclusive lanes may be operate either in the same direction as the other lanes in the cross-section or may be contra-flow lanes. The creation of exclusive bus lanes is the first step towards the development of a BRT system.

Bus lanes can be located on conventional highways such as arterials and freeways, referred to as on-street service options, or as busways, referred to as off-street service options. On-street service options have the benefit of relatively low cost and ease of implementation. The options that exist are the placement of the bus lane, whether adjacent to the kerb or to the median, the direction of flow, normal or contraflow and the composition of the traffic flow in the bus lane – buses only or mixed with taxis or freight vehicles (Miller, 2009).

15.5.4 Managed lanes

It is not possible or even necessarily advisable to keep on adding to infrastructure to accommodate the increase in demand for transportation. The total number of vehicle miles travelled has increased by 70 per cent in the last 20 years whereas highway capacity has grown only by 0.3 per cent (Federal Highway Administration, 2008). Furthermore, many road agencies are being confronted by a serious transportation funding crisis. Some way must therefore be sought to make the existing network more efficient.

These agencies are turning more and more to ways to better manage the flow of traffic on the existing facilities. They are seeking to regulate demand and separate traffic streams to reduce turbulence and also to utilise any available and unused capacity. These operational policies are evolving into the notion of 'managed lanes'.

The managed lane concept is typically a 'freeway within a freeway' whereby a set of lanes within the cross-section is separated from the general purpose lanes. A high degree of operational flexibility is built into the facility so that operations can be managed to respond to changing needs, including growth. The principal management strategies include pricing, vehicle eligibility and access control.

A typical application of managed lanes is the process whereby the direction of flow can be varied in a lane to accommodate the difference between the peak hour flow and the flow in the reverse direction. Lanes that are located within the median can thus accommodate inbound flows during the morning peak and outbound flows during the evening peak. A further application is to allow drivers of vehicles that do not qualify for access to HOV lanes to have access to these lanes against the payment of a toll and hence not only generate much needed funding but also spread the traffic load across more lanes.

15.6 BUS RAPID TRANSIT SYSTEMS

15.6.1 Introduction

Bus rapid transit (BRT) is a bus-based form of mass transit sometimes referred to as a 'surface subway' because it contains many of the features normally associated with the operation of subways. To be considered a proper BRT system, buses should operate for a significant part of their journey within a fully dedicated right of way, in order to avoid traffic congestion.

The first BRT system was the Rede Integrada de Transporte ('Integrated Transportation Network') in Curitiba, Brazil, which entered service in 1974. There are currently 168 cities that have adopted BRT systems and these transport about 31 million passengers per day.

15.6.2 The unique features of BRT systems

Various features of a BRT system that do not appear in a conventional bus system are aimed at eliminating the major weakness of public transport, specifically journey time.

- BRT systems normally have their alignments in the centre of the road to avoid kerbside delays.
- In conventional bus systems, passengers buy their tickets from the bus driver. This slows down the boarding process considerably. BRT systems replace this with ticket sales at stations usually via the medium of slot machines such as those used to pay for parking at shopping centres.
- Where BRT buses have to go through intersections, signal control gives buses priority. This is achieved by the buses being equipped with transponders that are detected by the traffic signal controllers. The green time on the bus route can be extended or the opposing red time shortened to ensure that the bus can pass through the intersection without stopping.
- As discussed in the following section, the vehicles themselves are specifically designed to function in a BRT mode.

15.6.3 BRT vehicles

BRT vehicles normally have low floors to eliminate the need for passengers to climb up steps. The floor level is typically that of the bus station so that passengers in wheelchairs can easily board the buses.

There is an argument in favour of guided buses. The buses have to park very closely to the boarding platforms to ensure that wheelchair wheels are not trapped between the bus and the boarding platform. This can be achieved by the use of specially designed kerbs such as Kassell kerbs or by guide wheels mounted on the bus. These effectively take over the steering of the buses from the drivers. An example is shown in Figure 15.8.

Other forms of guidance include optical means whereby buses have precision-docking capabilities as efficient as those of light rail and reduces dwell times, making it possible to drive the vehicle to a precise point on a platform according to an accurate and reliable trajectory. The distance between the door steps and the platform is optimised to be less than 5 centimetres. Level boarding is possible, and there is no need to provide a ramp for people with mobility impairments.

Magnetic guidance uses magnets embedded in the road surface.

Figure 15.8 Kerb guidance of BRT vehicle.

Part II

Supporting disciplines

Chapter 16

Project analysis

CONTENTS

16.1 INTRODUCTION

The availability of funding is always a major constraint in the provision of infrastructure. The decision of whether a road is to be provided is thus dependent on its cost in relation to the benefits it will generate. This comparison is generally referred to as cost/benefit analysis (CBA). Benefits may be directly economic by reducing the costs incurred by the road user in using the road or they may be indirectly economic in the sense of supporting the development of a geographical region. Alternatively, a road may hold little or no economic benefit as such but is nevertheless useful to the community at large. It has, in fact, a value of utility. Examples are access to schools, hospitals, places of worship and recreational facilities or areas. In these cases, a form of CBA known as social CBA (Snell, 2011) or utility analysis is brought into play. Both forms of analysis are discussed in this chapter.

Two scenarios are also discussed. In the first instance, the situation is that of identifying the preferred solution to a problem amongst a number of competing alternative solutions. An illustration of this is the selection of the preferred route between two common points. The competition between the alternatives stems from the fact that, of all of them, only one will be built. The second instance is that of noncompeting alternatives. This is illustrated by a need to compare two different routes – the route between towns A and B compared to that between towns C and D. The point at issue is that selection of the one alternative does not preclude provision of the other and the only limitation is the budgetary restraint on the funding source. At a higher level than a line department, such as Transport, Health or Education, to wit the Treasury seeking to allocate funding between competing authorities, comparison may be between the building of a hospital and a school, that is, totally different schemes.

In the case of competing alternatives, economic analysis normally takes the form of benefit/cost studies or net present worth (NPW) whereas noncompeting alternatives are normally compared by application of the rate of return methodology. These forms of analysis are discussed in this chapter. Utility analysis, being a noneconomic form of CBA analysis, is also discussed.

16.2 FUNDAMENTAL PHILOSOPHIES OF ENGINEERING ECONOMICS

There are 10 basic principles that have to be borne in mind in conducting an economic evaluation of a project (Grant et al., 1982):

1. Decisions are among alternatives.

   There is no need for a decision unless two or more alternative courses of action are possible. Often, decisions are made by default simply because the decision maker fails to recognise the possibility of alternatives and considers only one possible solution to the problem to be addressed. The alternative most frequently overlooked is the 'Do nothing' alternative.

   Doing nothing is somewhat of a misnomer; it could be taken to mean that there isn't any expenditure even on necessary maintenance and the situation will continue to deteriorate. In the context of CBS it must be understood that, in the 'Do nothing' alternative, the existing situation and quality of service will be maintained and that an expenditure of time, money and effort will be involved to achieve an unchanged level of benefit.

   It does not necessarily follow that investment in infrastructure is a necessary prerequisite to achieving a desired goal. In Chapter 4 discussion focussed, inter alia, on context-sensitive design. This is not surprising, as the focus of this book is on design. However, reference is increasingly to context-sensitive solutions because it has been realised that context sensitivity can be achieved without the application of measures requiring design. Similarly, a change in operational procedures may result in achieving an increase in throughput in a factory without having to invest in major plant improvements. Improved maintenance practices could result in a reduction of road user costs, one of the major benefits normally considered in CBA.
2. Decisions should be based on the expected consequences of the various alternatives, all of which will occur in the future. Prediction carries with it a level of uncertainty and hence risk. The benefits arising from the provision of a road are achieved over its

lifetime, and it is common cause that a road continues to be useful long after the end of its design life. Benefits accruing in the far future are more difficult to predict than those in the near future.

3. It is essential to decide whose viewpoint should be adopted.

   In the case of business enterprises, analysis would be from the viewpoint of the owners or shareholders for whom a healthy balance sheet and profit and loss statement are of prime importance. In the case of government departments, the choice of viewpoint is more complicated. Historically, Roads Departments did not carry out comprehensive CBA studies and were usually satisfied with 'least cost' analyses if, in fact, any economic analysis was conducted at all. Initially, least cost related purely to minimum construction cost. With time and increasing sophistication, it was realised that cheap construction could generate an enormous price tag for maintenance. 'Least cost' then came to be understood to be the total of construction plus maintenance. The discounting of maintenance cost over the lifetime of the road was a first step towards proper CBA studies.

   At about the time that Roads Departments were transforming themselves into Departments of Transportation, it was finally realised that the concept of 'benefits to whomsoever they may accrue' first formalised in the United States Flood Control Act of 1932 applied also to roads and that the road user was the beneficiary of improvements to the road network. Furthermore the road user was, simultaneously, through fuel and other taxes, the principal source of funding for their provision. Road user benefits are now part of the analytical process and include time savings and reductions in running costs and accident costs in the spectrum of benefits. The costs are those incurred by the Road Authority in the planning, design, construction and maintenance of the road.

4. Alternatives can be compared only if the consequences of their implementation are commensurable, that is, that they expressed in numbers and the same units apply to all the numbers.

   Comparisons cannot be drawn between apples and pears. They can only be between apples and apples or pears and pears. It is not useful, for example, to say that a bridge will cost a certain number of dollars and will have the benefit of saving 10 minutes a day for each of 4000 people. Similarly, benefits and costs can be compared only if they are measured in some or other common unit of measurement.

   Only money units lend themselves to the need for commensurability. There is a caveat that applies to their use and that is that, apart from being an economic disaster, zero inflation does not normally exist except for very short periods of time. The purchasing power of whatever unit of currency, be it the dollar, the pound or the euro, is used in the analysis will change with time. The base year of the analysis must therefore be specified so that it is understood that, for example, comparisons are being drawn in terms of the 2015 dollar.

   All benefits realised and costs incurred, whether in the past, present or future, will be discounted to the value expressed in the currency of the base year. It should be clear that inflation does not feature in economic analyses. Discounting refers to the fact that a dollar now has a higher utility than a dollar sometime in the far future. Time does, in fact, have a cost.

5. Only the differences between alternatives are relevant in their comparison.

   A CBA determines the differences between competing alternatives. Points of similarity thus are superfluous.

6. Separable decisions should be made separately.

   Many projects go through a spectrum of development, typically the addition of lanes or auxiliary lanes with time. Each new development must be able to pay its own way in terms of additional benefits in relation to the additional costs incurred. Otherwise, the fact that proposed increments of investment may not be economically productive could be concealed by evaluation of the project as a whole.

7. It is essential to have at least one or, for preference, several, criteria for decision making.

   As suggested earlier, not all consequences of decisions about investment in infrastructure are reducible to monetary terms. Utility analysis as a subset of CBA then comes into play and weighting needs to be applied to such irreducible data. Weighting is typically based on engineering judgment and it follows that criteria have to be developed to ensure that these judgments are as unbiased as possible. Having more than one criterion would make it possible to refine the value attached to any particular benefit.
8. The primary criterion to be applied in a choice between alternative investments should be selected with the objective of making the best use of limited resources.

   In evaluating alternative proposed investments, the problem is usually one of deciding whether an investment will be productive enough in terms of benefit in relation to the expenditure incurred. In essence, this is a case of establishing whether the rate of return achieved is acceptable when compared to one or more alternatives. In most forms of CBA, the internal rate of return is a given and is thus the primary criterion in the choice between competing alternatives.
9. Secondary criteria that reflect the lack of certainty about the future are helpful because the best of monetary estimates of consequences of a particular course of action would, in all probability, not be accurate. The further in the future the estimates are projected, the greater is the likelihood of error.
10. There are always consequences of choice of alternatives that cannot be reduced to monetary terms. Utility analysis and the term 'irreducible data' have been discussed earlier.

16.3 EQUIVALENCE

Problems of economy invariably involve determining what represents the minimum outlay over an extended period of time. It is thus necessary to recognise the time value of money. Because of the existence of interest, a dollar now is worth more than the prospect of a dollar at some future time. Interest may be thought of as money paid for the use of borrowed money. This is the viewpoint of the borrower whereas the viewpoint of the lender is that interest is the return obtainable from the productive investment of capital. This rather broader view is the one typically applied in economics (Grant et al., 1982).

Equivalence can, however, be more easily grasped in terms of the concept of interest paid on money borrowed. A loan can be repaid in numerous different ways. Four options can be listed:

- Scheme I. A single lump sum payment at the end of the loan period, with the lump sum including both capital and interest
- Scheme II. Payment of interest only during the currency of the loan, with a single payment of the capital at the end of the loan period
- Systematic reduction of the principal of the loan, with this method subdividing into
- Scheme III. Uniform repayment of the principal, with diminishing interest
- Scheme IV. Uniform payments where the sum of interest charged plus principal redeemed remains constant

Using the formulae listed in Section 16.4, comparison of the total payments made with respect to each repayment scheme, assuming a loan of \$10,000 at 9% over 10 years would show that these are

- Scheme I. \$23,670
- Scheme II. \$19,000
- Scheme III. \$14,950
- Scheme IV. \$15,580

suggesting that Scheme I is the best if the viewpoint of the lender is adopted and Scheme III the best in the view of the borrower. In fact, neither conclusion is correct because the time value of money is ignored. Regardless of the particular series of repayments that is adopted, the fact of the matter is that they have one thing in common and that is that they all serve to secure a loan (the principal) of the same value. In short, their present worth is identical.

Stated more broadly, if a service is to be provided and various repayment schemes are on offer, these should not affect any analysis of the economics of providing the service. An intelligent decision can be made only if all schemes are converted to some common platform, which may be an equivalent present-day single payment (NPW) or an equivalent uniform series of payments (Scheme III). Engineers tend to favour NPW because of its convenience whereas economists tend to favour the equivalent uniform series on the grounds that NPW ignores the costs associated with the need to borrow the money in the first instance. In essence, interest paid, particularly if the period of repayment is long and the rate of interest high, has a present worth of its own. As any homeowner who is paying off a bond can testify, this amount is substantial whereas, in an engineering analysis, it is, in fact, zero. The point is that the engineering analysis is solely concerned with the relative merits of two competing schemes. And these compete in the sense that they represent alternative ways of providing the same service, for example, a transportation link between two nodes.

16.4 INTEREST FORMULAE

In the interests of brevity, the derivation of the various formulae is not given here. These can, however, be found in any economics text.

The symbols used in the formulae have the following meaning:

i = interest rate per interest period
n = number of interest periods
P = present sum of money
F = sum of money at the end of n periods from the present date that is equivalent to P with interest, i, or future worth
A = end-of-period payment (or receipt) in a uniform series continuing for the coming n periods, with the entire series being equivalent to P at interest rate i, usually referred as the annuity

Although the formulae offered apply to any interest period, be it a day, week, month or year, the interest normally pertains to a year and is referred to as the nominal rate of interest. The convention adopted throughout is that reference is to end-of-period payments.

The formulae are as follows (Wolhuter, 2003c):

1. Given P, to find F

$$F = P(1 + i)^n$$

2. Given P, to find A

$$A = P \frac{i(1+i)^n}{(1+i)^n - 1}$$

3. Given A, to find F

$$F = A \frac{(1+i)^n - 1}{i}$$

4. Given A, to find P

$$P = A \frac{(1+i)^n - 1}{i(1+i)^n}$$

5. Given F to find P

$$P = \frac{1}{(1+i)^n}$$

6. Given F to find A

$$A = F \frac{i}{(1+i)^n - 1}$$

Equations given under 4, 5 and 6 above are the reciprocals of those under 2, 1 and 3 respectively.

The expression $(1 + i)^n$ is referred to as the single payment compound amount factor, *F*/*P*, with the shorthand *F*/*P* having the sense of 'To find *F* given *P*'. Its reciprocal, *P*/*F*, meaning 'To find *P* given *F*' is called the single payment present worth factor.

A fund established to produce a desired amount at the end of a given time span by means of a series of end-of-period payments is referred to as a sinking fund so that the expression $i/[(1 + i)^n - 1]$ is known as the sinking fund factor, *A*/*F*. The reverse case involves a single present day investment, *P*, with a view to a series of uniform end-of-period payments, *A*, where the capital recovery factor, *A*/*P*, is given as $[i(1 + i)^n]/[(1 + i)^n - 1]$. The capital recovery factor is always equal to the sinking fund factor plus the interest rate and is used in the solution of many problems in engineering economy.

The reciprocal of the sinking fund factor is called the uniform series compound amount factor, usually abbreviated to series compound amount factor, *F*/*A*. The reciprocal of the capital recovery factor is known as the uniform series present worth factor, *P*/*A*.

Prior to the development of calculators that could handle exponentiation, interest tables, from which these factors could be read off, were useful and are still to be found as appendices to many texts on economics.

In the preceding formulae, the value of i, the rate of interest, is given. It is sometimes useful, though, to know what the rate of return for a project really is rather than whether or not the investment meets a given standard of excellence. Calculation of the rate of return is discussed in Section 16.6.5.

When reference is made to a nominal rate of interest, this is understood to mean that interest is compounded annually. Savings and loan accounts are often subject to compounding at shorter periods with reference to quarterly, monthly or daily balances.

A distinction is drawn between the nominal interest rate and the effective interest rate. These terms may be defined more precisely by considering the case of interest being compounded m times a year at an interest rate of r/m per compounding period. The nominal rate of interest is $m \cdot (r/m) = r$, whereas the effective rate is $(1 + r/m)m - 1$. To, once again, avoid efforts to compare apples and pears, engineering economic studies tend to favour dealing with effective rates rather than nominal rates. The compounding period can be made shorter and shorter until, in the limit, it can be said that interest is continuously computed and added to the principal. In practice, the result of continuous compounding is very close to that obtained using monthly compounding with a nominal rate, r.

16.5 PROBLEMS INVOLVING GRADIENT

Expenses and disbursements need not always remain constant and can increase or decrease by varying amounts. A case in point is that maintenance expenditure on a road can be expected to increase with its age. Ultimately the point is reached whereby where the maintenance bill becomes so high that the road has to be rehabilitated. If the increase in expenditure is the same for every successive year, reference is made to a uniform arithmetic gradient.

In this case, a payment of G at the end of the first year is followed by a payment of $2G$ in the following year, $3G$ in the third year and so on. One way of calculating the outcome is to treat each year's payment as the principal for a reducing number of years. Some manipulation will indicate that this is equivalent to a sinking fund so that the sum of the compound amounts is

$$F = \frac{G}{i} \; \frac{(1+i)^n - 1}{i} - \frac{nG}{i}$$

with the equivalent uniform annual figure found by multiplying this amount by the sinking fund factor for n years.

Another gradient form of problem involves escalation. In this case, a payment made in any given year is larger than that of the preceding year by a constant factor. This is identical to the problem of a final amount calculated from a single principal amount with the interest, i, replaced by the anticipated growth, g.

Although it is mathematically easy to combine interest and growth into a single formula, this is not recommended. The problem is that it can easily lead to misinterpretation. Economic analyses use the currency of the base year for purposes of comparison and inflation represents the strength (actually weakness) of the currency of some future year. Inflation thus has no place in economic analyses and should first be eliminated by recalculation of the magnitude of future expenditure from the currency of that year into the currency of the base year although actually paid over in the year in question. Thereafter it is possible to start comparing the merits of various proposals in terms of the interest formula given previously.

16.6 METHODS FOR COMPARING COMPETING PROPOSALS

16.6.1 Introduction

The introduction of the time value of money into economic studies reflects the requirement that capital be recovered with an acceptable return. Proposed investments in capital items or engineering projects such as a road are therefore unattractive unless a prospect for their being recovered with interest exists. The rate of interest should be the minimum rate of return deemed attractive under the particular circumstances and this rate of return is often referred to as *i** (pronounced eye-star).

Four basic methods of comparison of alternatives, each of which have differing patterns of disbursements and receipts, can be considered:

- NPW, with a stipulated minimum attractive rate of return, *i**, as an interest rate
- Benefit/cost ratio, with the stipulated minimum attractive rate of return, *i**, used as an interest rate
- Equivalent uniform annual cash flow with a stipulated minimum attractive rate of return, *i**, used as an interest rate
- Prospective, or internal, rate of return, with the calculated rate of return compared with the stipulated minimum attractive rate, *i**

With the exception of arithmetic error, the four methods lead to the same decision among competing alternatives. Each method has its advantages and disadvantages so that the selection of a method of comparison is not necessarily a matter of indifference.

It will also become clear that the selection of the minimum acceptable rate of return, *i**, is critical to the outcome of the exercise. An investment that may look attractive with *i** set at 5 per cent would cause horror if *i** was 15 per cent.

Two of the four methods of comparison of alternatives are illustrated by an example in which a contractor is considering replacing a materials handling process currently using hand labour to break down oversize material to crushing by mechanised methods. These serve to demonstrate that the outcome is unaltered by the selection of the method of comparison adopted.

Two options are to be considered:

- Use plant, which is anticipated to have a duration of 10 years, being custom-built for the operation.
- Invest in general-purpose equipment that is more expensive than that considered in the first alternative but with the possibility of having a substantial salvage value at the end of the 10-year period, because the life expectancy of the more expensive equipment is anticipated to be 20 years.

The labour cost currently involved in the operation is $92,000 per annum. The plant required by the first alternative will cost $150,000 but will reduce the annual labour cost to $33,000. Additional annual payments with respect to power and maintenance amount to $38,000. The second alternative requires a plant costing $25,000. This plant is estimated to have a salvage value of $50,000. In addition, because of a host of automatic features, the labour cost is $14,500 per annum. The requirements in terms of power and maintenance are estimated at $41,750.

The economist decides that these schemes, with the ‘Do nothing’ alternative as Scheme A and the alternatives as Scheme B and Scheme C respectively, should be compared on the basis of a rate of return, i^*, of 9 per cent.

16.6.2 Net present worth

Calculation of NPW is often called discounting so that the interest rate used in the calculation is referred to as the discount rate. An important use of NPW is in trial-and-error calculations to determine unknown rates of interest or return. The major application is, however, the comparison of alternative series of estimated receipts and disbursements on the basis of a single disbursement and a current value on prospective receipts.

The application of NPW is illustrated, using the schemes of materials handling outlined previously.

16.6.2.1 Scheme A (Do nothing alternative)

All that is required in this case is to convert the series payment of $92,000 per annum to a single principal (*P*/*A*, 9%, 10) and this is

$92,000 * (6.418) = $590,500

16.6.2.2 Scheme B

Capital acquisition is presumed to be done in year zero so that

Present worth of first cost = $150,000
Present worth of annual disbursements $64,000 * (6.418) = $408,000
Total NPW = $558,080

16.6.2.3 Scheme C

Present worth of capital acquisition = $250,000
Present worth of annual disbursements (*P*/*A*, 9%, 10) = $361,000
Total present worth = $611,000
Less: Present worth of salvage value (*P*/*F*, 9%, 10) $50,000 * 0.4224 = $21,100
Total NPW = $589,900

This comparison between the three schemes shows that there is little to choose between Schemes A and C, with Scheme B being the best of the three alternatives.

16.6.3 Equivalent uniform annual cash flow

Equivalent uniform annual cash flow (EUACF) is often referred to as the annual cost of the alternatives considered. It is possible to derive annual cost by applying the capital recovery factor, (*A*/*P*, 9%, 10), to the present worth. In the case where there are many disbursements or incomes most conveniently expressed as annual cash flows, it is more convenient to convert those expressed in terms of NPW to EUACF.

16.6.3.1 Scheme A

As the 'Do nothing' alternative, this is already known to cost $92,000 per annum.

16.6.3.2 Scheme B

An initial outlay of $150,000 is called for and this must therefore be converted to its annual equivalent (*A*/*P*, 9%, 10).

EUACF *A* = $150,000 * [0.09(1.09)10]/[1.0910 – 1]

= $150,000 * 0.1558

= $23,700

This must be added to the cost

Initial outlay *A* = $23,700

Labour = $33,000

Power, etc. = $31,000

Total EUACF = $87,700

16.6.3.3 Scheme C

In this case the initial outlay is $250,000 so that the equivalent annual cash flow (*A*/*P*, 9%, 10) is *A* = $250,000 * 0.1558 = $38,950.

There is, however, a salvage value involved and this occurs at the end of the anticipated life span of the project. The calculation called for (*A*/*F*, 9%, 10) indicates that

A = $50,000 * 0.0658

= $3290

Initial outlay minus salvage value is the difference between the two items and is thus

Initial outlay *A* = $35,660

Labour = $14,500

Power, etc. = $41,750

Total EUACF = $91,910

The calculations show that there is little or no difference between Schemes A and C. A choice to proceed with Scheme C would thus depend on irreducibles such as the prospect of labour unrest. This, of course, presupposes that Scheme B is unavailable because it is plainly the most attractive of the three alternatives.

16.6.4 Benefit/cost ratio

The application of the benefit/cost ratio in the analysis of projects stems from the United States Flood Control Act of 1936 in which it was stipulated that 'benefits to whomsoever they may accrue' should exceed 'estimated costs'. Ultimately, this approach was applied to many public works projects other than flood control.

To illustrate the application of benefit/cost analyses: A stretch of rural road is in such poor condition that something drastic in the line of rehabilitation has to be done. Three possible alternatives are considered:

- Alternative H. Retain the present alignment, but rip up, recompact (adding new material as necessary) and resurface. Construction cost, \$1,100,000; annual maintenance cost, \$350,000; residual value after 20 years, \$0.0.
- Alternative J. Relocate. Construction cost, \$7,000,000; annual maintenance cost, \$210,000; residual value, \$3,000,000.
- Alternative K. Relocate. This route is shorter than that proposed for alternative J, but involves heavier earthworks. Construction cost, \$13,000,000; annual maintenance cost, \$170,000; residual value, \$5,500,000.

The value of i^* is assumed to be 7 per cent.

If we had been using NPW as our form of calculation, Table 16.1 would have immediately indicated that location J was the favoured economic solution to the problem.

The fundamental difference between benefit/cost and other forms of comparison is that it is marginal, whereas the others are absolute. What it is saying, in effect, is 'Is the increase in benefit greater than the increase in cost? That is, is the benefit/cost ratio greater than 1?'. This form of calculation comes into its own when it is necessary to calculate marginal improvements. As stated earlier, 'Only the differences between alternatives are relevant in their comparison' and it is sometimes easier to determine differences in cost than absolute costs.

The process followed is to take the first competing alternative and compare it to the null alternative. If it shows a B/C ratio less than 1, it is rejected and the next alternative is compared to the null alternative. If, however, its B/C ratio is greater than 1, it becomes the favoured alternative or 'champion' and the null alternative is rejected. The next competitor is then compared to the current champion until an overall winner finally emerges.

The benefit in benefit/cost is that accruing to the road user and comprises reductions in time costs, vehicle operating costs and crash costs. These are related to a single user and then

Table 16.1 Comparison of competing alternatives

Item	*Location H*	*Location J*	*Location K*
Cost of construction	\$1,100,000	\$7,000,000	\$13,000,000
NPW of residual cost	\$0	\$1,525,000	\$2,796,000
NPW of maintenance costs	\$2,458,000	\$1,475,000	\$1,194,000
NPW of road user costs	\$28,230,000	\$21,170,000	\$18,350,000
Total	\$31,788,000	\$28,120,000	\$29,480,000

applied to the number of road users expected to be using the facility over the next x years. Because of the large multiplier, accuracy in the determination of road user costs is difficult to achieve. In the interest of brevity some road user costs were assumed in Table 16.1.

The first step in the process is to convert all cash flows to their present day equivalent, that is, their NPW. For the purpose of the illustration, all the sums quoted are given in the currency of year zero. Otherwise it would have been necessary to first apply a conversion by considerations of the inflation rate and THEN calculate their NPW using the stipulated value of i^*, which, for the occasion, has been taken as 7 per cent.

One of the weaknesses of benefit/cost analysis is that the definition of benefits or costs can be somewhat arbitrary. As stated previously, the benefit is the reduction in road user cost. The residual cost could be seen as a saving and hence a benefit. However, the residual cost is, in effect, a reduction in overall cost and applies to the cost centre, the Road Authority. Similarly, the reduction in maintenance cost also applies to the Road Authority, hence its inclusion below rather than above the line. The benefit/cost ratio for location J versus location H thus becomes

$$B/C_{JH} = (\$28{,}230{,}000 - \$21{,}170{,}000)/[(\$7{,}000{,}000 - \$1{,}100{,}000) - \$1{,}525{,}000 + (\$1{,}475{,}000 - \$2{,}458{,}000)] = 1.32$$

Location J is thus favoured over location H and becomes the new champion. Location K is then compared to the new champion. The benefit/cost ratio is 0.81 so that the additional benefit is not sufficient to offset the additional cost. Location J remains the champion.

16.6.5 Rate of return

The previous two forms of comparison between competing projects employ a preset rate of interest, i^*, (verbally referred to as i-star) and it can be demonstrated that the selection of value of i^* can influence the outcome of the comparison. It is sometimes preferable to calculate what the actual rate of interest realised by each of the competing proposals really is.

Part of this preference stems from the fact that totally different investment opportunities can be compared whereas the previous methods are employed to compare different methods of providing the same service. The rate of return methods could be used to determine whether an allocation from a limited state budget should be made into a road, a hospital or a school. Where an analysis is directed towards a funding source that has various options of investment available to it, analysis should thus be couched in terms of rate of return as opposed to NPW or benefit/cost analyses.

The rate of return is defined as the interest rate at which the NPW is zero. Alternatively stated, the benefits are equal to the costs or disbursements so that the benefit/cost ratio is 1. This can be done only by resorting to a trial-and-error process. Using a NPW calculation, as described in Section 16.6.2, two values of i are assumed and the NPW calculated for each of these values of i. If the values have been well selected, one will show a NPW that is negative and the other a positive NPW. A series of iterations should then be carried out in which the selected values of i result in the gap between the two values of NPW shrinking towards zero. The value of i at which this is achieved is the desired rate of return.

Such calculations are variously referred to as the discounted cash flow method or the investor's method. Rate of return calculated in this way has been called the profitability index (sometimes abbreviated to PI, under which circumstances it should not be confused with the point of intersection, PI, or the plasticity index, PI), the interest rate of return, the solving rate of return or the internal rate of return. More recently, two other terms, measure of worth and figure of merit, have come into popular usage.

If internal rates of return had been calculated rather than NPW or benefit/cost ratio, the outcome would have been the same EXCEPT that the road authority concerned may decide that some totally different project should be the recipient of funding on the grounds of a still better internal rate of return, leaving the local road users without the benefit even of the relatively modest expenditure required by location H.

The consequence of abandoning the scheme in total and going to some other should include the fact that the effectively zero investment would no doubt be accompanied by very high road user costs. As such, the 'Do nothing' approach should rather be replaced by the 'Do minimum' approach, which is sufficient investment to maintain the benefits at their current level.

In theory, the Road Authority should seek to allocate the expenditure of its budget in such a way that the benefit/cost ratio accruing from its total expenditure is optimised. The reality mitigating against this ideal state of affairs is that road authorities have great difficulty in estimating what their allocation from the state budget is going to be. Predicting with some expectation of reasonable accuracy some 20 years hence is beyond the bounds of feasibility. Furthermore, even if this were known, irreducibles such as concern for the environment and political pressures have a way of overturning sound economic analyses. Finally, the complexity and magnitude of the calculation involved are substantial.

16.7 FACTORS INCLUDED IN ECONOMIC STUDIES

16.7.1 Introduction

The phrase 'to whomsoever they may accrue' with reference to the benefit/cost ratio highlights a problem that has not been addressed so far. In the earlier example, there was no doubt as to who would benefit from the investment. It was the investor himself. Where the general public is the beneficiary, the definition of benefit is less clear cut. Relocating a road to eliminate a dangerous curve will reduce accident costs, which, of course, means that vehicle repair businesses, members of the legal profession and medical practitioners will suffer a loss of income. It follows that cash flow, which can be used in analyses for the private investor, is not appropriate because, as just illustrated, every positive cash flow (receipt) by someone is a negative cash flow (disbursement or loss) by someone else. The algebraic sum of cash flows is then zero and there is no basis for economic evaluation. It is therefore necessary to have a clear idea of what is a benefit and what is not.

In the case of road projects, benefits are taken as those accruing to direct users of the facility in terms of

- Time savings
- Reduction of running costs
- Enhanced safety, that is, reduction in number and severity of crashes
- Delay

Of these, time savings are a major portion of the benefits and their quantification is thus important.

There are also benefits accruing to the community at large other than the road users (Snell, 2011):

- Stimulation of economic activity (which may, in the case, e.g. of a new bypass, be a negative benefit for High Street shopping areas)
- Reduced crashes in terms of the community cost perspectives of health and ambulances services, police, hospitals and aftercare

- Environmental changes which, like economic activity, may be either positive or negative and include loss or enhancement of environmental and recreational facilities or improved access to them
- Wider environmental effects such as increases or reductions in noise levels, atmospheric pollution
- Neighbourhood safety and security

The valuing of costs is as complex as that of benefits. They are thus grouped together as 'impacts' on the basis of 'with project/without project' differences in each category. The distinction between benefits and costs is often blurred so that benefit/cost ratio is seldom a suitable form of analysis and NPW is preferred. Furthermore, placing a monetary value on the benefits or otherwise accruing to the community at large can be a heavily biased process.

As discussed further later with reference to the value of time savings, the customary approach to the determination of benefits is to determine the benefits experienced by a single person and then multiplying this value by the number of persons to whom the benefits accrue. Benefits to road users are thus dependent on the number of vehicles on the road, the duration of their trips and the number of occupants in each vehicle.

These can be determined by traffic surveys but, as benefits accrue over the design life of the road, one of the problems facing the analyst is the matter of traffic growth. In rural areas where changes in traffic patterns are slow, a simple growth rate is assumed and applied to the entire life of the road. A fraction of a percentage difference between what is assumed and what is ultimately experienced could result in a substantial difference in the benefit experienced over the design life of the road. In urban areas, changes in traffic growth vary as the nature and density of adjacent land uses change. These changes are usually stepwise and rapid. Highly sophisticated and complex urban planning and transportation modelling has to be brought to bear on determination of the number of person and goods trips to be accommodated and on which routes between a multitude of origins and destinations. History as shown that, regardless of the methodology brought to bear, traffic growth is usually grossly underestimated.

16.7.2 The worth of time saved

The cost of business trips and freight transport can generally be evaluated from the costing by the employers or hauliers concerned. Time saved by infrastructural improvements to these trips and the corresponding implications in terms of cost could thus be determined with a fair amount of confidence.

Personal trips, that is, commuting, shopping, recreation, and so forth, are not paid for by a third party and the worth of time saved by an improvement in the road network is thus a matter of the perception of the individual. Two approaches to the economic quantification of time savings appear in the literature. These are the 'constant value of time' approach and the discounted approach (Frith, 2012).

Loosely speaking, the constant value approach states that individual time savings have the same unit cost regardless of duration and can be added together. Taken to its logical conclusion, this approach implies that a saving of one second to each of 3600 people is the same as a saving of 1 hour. This is patently unrealistic and, given that time savings are typically of the order of 70 to 90 per cent of the total benefit, could result in a gross skewing of the analysis.

However, it is also argued that small time savings have zero or low value because they are not large enough to be either perceived and/or put to some alternative higher valued activity. This implies a threshold, the lower end of which could vary from seconds to a number of minutes depending on the view of what constitutes a useful packet of time.

A calculation by the Transport Agency of Canada (TAC; Markow, 2012) showed that removing small savings, that is, less than 5 minutes, from the calculation of savings reduced the benefit of time savings from 80 per cent to a third of the total. The issue arose because it was not clear whether these small increments of time should be valued proportionally to larger travel-time savings or whether small time intervals that fall below some threshold might be valued less than proportionally because they are too small to be used productively elsewhere. The approach recommended by TAC is to compute the cumulative small time savings (e.g. the sum of all savings of 5 minutes or less) and value them proportionally to larger savings, but to isolate them from the NPW analysis. In this way they are reserved for separate consideration by management.

The counterargument is that most savings are small and are generally found on trips of short duration. The value of the threshold has yet to be finally determined but it has been suggested that it should be some or other percentage of the total trip time.

The discounted approach produces more realistic values of time. It is quantified by stated preference or revealed preference surveys. These are usually couched in willingness-to-pay or willingness-to-accept concepts. The general consensus seems to be that the value that people attach to their own time is about 50 to 60 per cent of their hourly income level (Fosgerau et al., 2007).

As suggested previously, the designer must have some idea of the average occupancy of vehicles using the road before being able to use these figures. Five minutes saved may mean more in the case of a bus than in the case of the luxury sedan occupied only by the driver. Furthermore, the designer will be confronted by the problem of estimation of traffic growth over the design life of the road.

16.7.3 Operating costs

These are concerned with the direct cost of keeping a vehicle on the road and are thus concerned with consumables such as fuel, oil and tyres. Maintenance costs are a function of distance travelled and are thus also often, but not invariably, included. Annual charges, which would be incurred regardless whether the vehicle moved or not are, most often, excluded. These charges relate to depreciation, insurance and garaging.

A feature of operating costs that cannot be ignored is that the cost of fuel, oil and tyres include a plethora of taxes and levies. These are not true economic costs but are rather devices employed by the government as its funding sources and, as such, cannot be brought into the calculation of the benefits accruing from use of some or other improved facility. Tax revenue ‘lost’ by the government never becomes a ‘benefit’ to the road user.

16.7.4 Crash costs

Crash costs have also been the subject of much study. Various methods of calculation have been developed and those most frequently used are the human capital (or lost output) method and the willingness to pay method. Both of these are income dependent and it has been established that up to 40 per cent of the variation between the values attached to fatalities can be accounted for by variation in gross domestic product per capita. It would thus appear that saving a life in a low-income country is worth less than in a high-income country. This is borne out by calculations of the value of a statistical life (VSL) derived from application of one or other of the methodologies referred to previously (McMahon and Dahdah, 2008). Regarding fatal crashes, Table 16.2 illustrates the variation of VSL and gross domestic product (GDP) per capita between various countries. The contrast is glaring.

Table 16.2 Comparison of VSL and GDP per capita between various countries

Country	*VSL (US $000)*	*GDP/capita (US $)*
Australia	1304	28,935
Austria	3094	35,871
Bangladesh	71	1710
Canada	1427	29,851
France	1252	29,472
Germany	1257	28,953
Iceland	3304	44,679
India	147	2651
Indonesia	92	3125
Latvia	1042	18,140
Lithuania	746	12,027
Malaysia	722	9513
Myanmar	51	1545
Netherlands	1944	31,009
New Zealand	2033	25,024
Poland	574	14,984
Singapore	924	25,034
Sweden	2015	32,394
Thailand	222	6958
United Kingdom	2292	32,555
United States	3000	36,311
Vietnam	53	2475

Crash costs are typically quoted in terms of fatal crashes, crashes causing injuries with these sometimes being split between serious and minor injuries, and property damage only crashes. Values are given annually by the statistical services of various countries and vary according to a number of factors such as the level of development of the country and the standard of living of its inhabitants of the country. For example, there is a significant difference between crash costs in developing countries and developed countries across all levels of severity of crashes. Fatal crashes consider matters such as loss of potential earnings, which are based in turn on the age of the deceased and his or her educational level. Injuries are also concerned with potential loss of income, in addition to the social cost of medical treatment, pain and suffering. Property damage only is self-explanatory including the cost of repairing damage to vehicles (and their contents – other than the passengers) and the environment.

As a form of utility analysis, crashes are rated in terms of the Abbreviated Injury Scale (AIS) illustrated in Table 16.3. The most widely used derivative is the Maximum AIS (MAIS).

Table 16.3 Description of the AIS code

AIS code	*Description*
1	Minor
2	Moderate
3	Serious
4	Severe
5	Critical
6	Maximal

16.7.5 Direct costs

Direct costs are the costs actually incurred by the Road Authority in providing the facility. These costs include construction and maintenance. They do not include the cost of borrowing the money (interest) because, as indicated previously, a present worth is a present worth, regardless of how it is repaid.

The fact that a facility may have a certain salvage worth at the end of its useful life is included in the calculation as a reduction in cost rather than as an increase in benefit, because it accrues to the road authority rather than to the road user.

16.8 UTILITY ANALYSIS

The benefit provided by a road may not include any economic aspects at all and the road could have a purely social benefit such as providing access to a school or hospital or a recreational area or centre. In this case, reference is made to utility analysis. When the benefits provided by the road are social rather than economic, the decision to provide it is essentially political in nature. Without a cash value being attached to a project it is necessary to ask how the community perceives the derived benefit. Would the road be essential or important or merely nice to have? Qualitative judgments such as these can be quantified by the application of utility analysis.

Utility is defined as the power of a commodity or service to satisfy a human want or need. It is thus the satisfaction that is derived by the consumer by consuming the goods. For example, cloth has a utility because it can be worn. A pen has a utility because it can be written with.

As previously stated, every project includes irreducible factors, that is, elements to which it is not possible to attach a cash value. Examples are

- Schools
- Churches
- Hospitals and clinics
- Recreational facilities
- Public open space and parks
- Areas of historical interest including monuments or memorials

These elements all have some greater or lesser value or utility to the community served by the road. Table 16.4 illustrates the calculation of the utility of a road. The values of utility

Table 16.4 Example of the calculation of the utility of a road

Item	*Quantity*	*Utility*	*Total*
Clinic	2	30	60
Community hall	1	10	10
Agricultural holding >100 ha	15	40	600
Primary school	2	20	10
High school	1	50	50
Church	1	10	10
Tourism facilities[a]	0	20	0
Recreational area	1	30	30
Population 1000–2000	3	80	240
Population >2000	1	100	100
Total			1150

[a] Although shown as a zero quantity for this road, it is included in the table purely as a guide to the analyst of the items that can be expected to be included in a utility analysis.

attributed to the listed amenities are purely arbitrary. Normally, they would be established by deliberation and debate by the experts briefed to carry out the analysis.

Utility could be established by the extent of usage of each amenity. A church that is used only for an hour or two by a congregation of 200 parishioners on Sundays would in all probability be assigned a low value of utility. On the other hand, if the church is a hub of the community with ongoing activities it would attract a higher value of utility. A high school used by a staff of 20 teachers and 400 students for 30 hours a week should have a high utility.

Utility analysis could be applied to two different routes between a common origin and destination, the point being that each route may serve different noneconomic activities. It could equally be applied to two totally different routes.

Chapter 17

Statistics

CONTENTS

17.1 INTRODUCTION

Statistics is a branch of applied mathematics and differs from pure mathematics in that the values in statistics are approximations or estimates, but they certainly are not guesses. An exact value may be thought of as being one of the possible values that a variable may assume. In pure mathematics, there are only two possibilities: that the value has a certain value or it does not. In the first case the probability is 1, in that it is certain that the variable has that value and in the second it is zero, meaning that it is certain that the variable does not have that value.

Although it was known in Grecian times that certain mathematical patterns were found to coincide with chance happenings such as occur in card games, it was later found that they also occurred in actual happenings. In the 17th century, one of the first practical uses of statistics was the development of life expectancy tables, for use in life insurance (Gerlough and Huber, 1978). Statistics has been applied for a long time in other sciences. The theory of evolution is heavily based on statistics. Theoretical physics is also based on statistics as employed by Bohr, Einstein and Fermi. The social sciences use statistics in the study of demographics, including population distributions by age, employment, income, morbidity rates and physical location. It was only in relatively recent times that traffic engineers realised that statistics and probability could be powerful tools in the design and operation of transportation networks.

There are three elements that have to be considered in the design of a road: the vehicle, the driver and the road itself. The first two are the purview of the geometric designer and the third that of the materials engineer. The vehicle is, by itself, an inert mechanism and is functional only when the driver is behind the wheel. It is the driver's control that dictates the path that is followed and the speed driven. This control is a function of the driver's characteristics, including age, skills level and physical and emotional characteristics not to mention mindset while driving. These characteristics can really be analysed in a meaningful manner only by the application of statistics and probability. This realisation led to the birth of traffic flow theory, which is discussed in Chapter 18.

In this chapter, the characteristics and the mathematics of probability are discussed. Statisticians notoriously accept that their readers are familiar with the various symbols and notations they employ. The intention in writing this chapter, while focussing on the principles of probability and the various statistical distributions that are used to model data sets, is thus to include some discussion of the mathematical shorthand and symbols typically employed in considerations of the notions of probability.

17.2 DATA ACQUISITION

17.2.1 Introduction

In most areas of research, data acquisition is arguably the most expensive of all the phases of a project. The Large Hadron Collider, for example, is, after all, nothing more or less than a very large and expensive device geared to the acquisition of data. At a more prosaic level, a government intending to conduct a national census needs to print millions of questionnaires and then appoint thousands of temporary agents to visit all the households in the country to help them in the completion of the questionnaire. With all the questionnaires received the workload of analysis would probably be beyond the capacity of the number of statisticians employed by the government and consultants would have to be retained to assist in the analysis process. It is known that, after the considerable financial outlay involved, the desired 100 per cent complete data set is never achieved.

In various fields it is not always desirable to have data on a total population. In manufacturing, testing 100 per cent of all products to destruction is obviously not desirable. The question must be asked as to what is the minimum size of the data set that will provide an informative answer on the quality and performance of any variable that has to be evaluated, and furthermore what the level of confidence is that can be attached to the answer provided by analysis of the data set.

17.2.2 Data sets

As suggested previously, many phenomena of concern to traffic engineers are random so that the outcomes of experiments are invariably different from one experiment to another. Experimental data can be expressed graphically in the form of histograms or probability density functions. Rainfall data for a catchment area in New York State for the period 1918 to 1946 (Ang and Tang, 1975) are summarised in Table 17.1 and illustrated by a histogram in Figure 17.1 and by a probability density function in Figure 17.2.

A data set is simply a group of values of a variable and is usually simply referred to as a set. Without wishing to delve into the details of the analysis, it is widely accepted that a set should comprise at least 30 elements to be meaningful.

17.2.3 Design and decision making under uncertainty

Predictions are based on models of greater or lesser complexity. In a situation in which no single observation is representative of the whole data set, prediction must be based on imperfect models to account for the variability between observations. A policy of decision making should thus be adopted. For example, it may be desirable to assume the consistently worst condition, for example, the highest possible flood of all time, and develop conservative designs based on this basis. Although in terms of safety this may be a suitable approach, it would inevitably be costly, possibly to the point of being unaffordable.

To design for the worst storm in 50 years would obviously result in a more economical design than designing for the worst flood of all time. On the other hand, an optimistic

Table 17.1 Data set of annual rainfall

Interval (mm)		*Number of observations*	*Percentage of total*	*Cumulative percentage*
900	1035	3	10.3	10.3
1035	1170	7	24.1	34.4
1170	1270	5	17.2	51.7
1270	1370	5	17.2	68.9
1370	1470	3	10.3	79.3
1470	1570	3	10.3	89.6
1570	1670	1	3.4	93.1
1670	1780	2	6.9	100.0

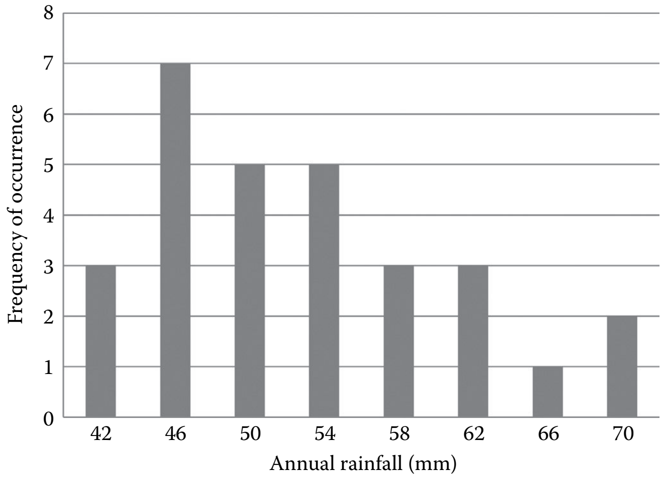


Figure 17.1 Histogram of annual rainfall data set.

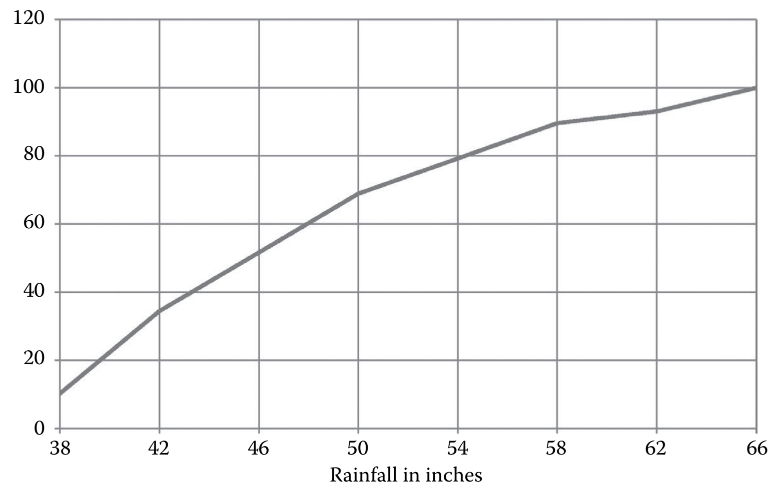


Figure 17.2 Cumulative distribution function (CDF) of annual rainfall data set.

design, for example, on the basis of the assumption of a low rate of traffic growth, may result in the premature failure of a pavement and the consequent cost of rehabilitation. Decisions must therefore be based on a trade-off between cost and benefit as was discussed under the heading of the Design Domain in Chapter 4.

With uncertainties in the data and the analytical models employed, it is necessary to allow in the trade-off for the consequences of decisions made. The unaffordability of designing for the worst storm of all time has been mentioned earlier. An optimistic design may, however, result in the bridge not being able to accommodate the flow in an abnormally wet season with the consequences of damage to the bridge itself or (more usually) its approach fills. The back-up could also result in flooding upstream and higher flow velocities downstream with consequent damage to properties and risk to life and limb on both sides of the structure. These consequences could have cost implications far exceeding any savings made in the design.

Variability of data occurs in all forms of engineering design. Mention has already been made of hydrology and pavement design. Structural design has to accommodate variables such as wind loading and snow on roofs, not to mention damage by hail. Geotechnical design, which is the prime input into structural engineering with respect to foundation design, has the problems of soils being inherently heterogeneous with irregular layers of various materials with widely varying ranges of load-bearing capacity arising from differences in soil density, moisture content and plasticity among other properties. Construction planning and management is concerned with the prediction mainly of time durations and the cost on site of materials. It has often been said that the first time that the cost of a contract can accurately be predicted is when it has been handed over to the client. Cost estimates that are too conservative may result in a tender bid being too high to be acceptable to the client. On the other hand, an optimistic bid may cause the contractor to end up in a seriously loss-making situation. The contractor needs to assess his or her bid with a degree of conservatism to realise an acceptably profitable contract. This also is modified by the personal characteristics of the contractor, who could be a gambler or be as conservative as a banker.

In traffic engineering, problems of variability abound. These include

- Reaction times, which vary widely between drivers and tend to increase with increasing age as drivers become more cautious
- Gap acceptance, where
  - At a priority-controlled intersection, drivers on the minor road have to make a judgment call on when it is safe to enter or to cross the through road
  - On the major road, where drivers on the through road have to await a gap in the opposing traffic to complete a left turn
  - Pedestrians at an unsignalised crossing having to await a gap in the traffic that will be acceptable to them
- Considerations of capacity and level of service, which are riddled with uncertainties. It is impossible to measure delay, either macro- or microscopically, because there are too many vehicles between origin and destination, all of whom are suffering varying degrees of delay

The aforementioned variations are all related to human factors.

Variations in vehicle properties also impact on design. Values of vehicle dimensions and performance of the design vehicles are offered in Chapter 5. The design vehicle does not exist in the marketplace. It is a composite of the values of the various features of the vehicles allocated to a specific class of vehicle such as their length, breadth, width and height, and radius of turning circle or the power output of their engines. These are all determined by statistical analyses and the assumption of a representative percentile value. The percentile adopted is often the 85th. Thus the ostentatiously long stretch limo would never serve as the design vehicle for the demarcation of parking areas.

17.2.4 Sample size

An unnecessarily large sample set will cost more than is necessary to acquire and a sample set that is too small will provide misleading results in the ensuing analysis. It is necessary to provide a data set that will provide acceptable accuracy of modelling without incurring excessive expenditure. Acceptability is expressed as a confidence level and this is described further in Section 17.6.6.

The size of the sample required to provide a meaningful basis for evaluation can be derived from Nezamuddin et al. (2010).

$$n \geq \frac{t_{\frac{\alpha}{2} \cdot n-1} \cdot s}{\varepsilon}$$

where

n = minimum sample size

$t_{\frac{\alpha}{2} \cdot n-1}$ = Student's t-statistic for $(1 - \alpha)$ *100% confidence level and $(n - 1)$ degrees of freedom

s = sample standard deviation

ε = permitted error

Seeing that the required sample size is a function of its standard deviation, which cannot be known until the data set has been created, that is, a circular process is required, it follows that several iterations may be required to determine an adequate sample size. A useful starting point is, however, offered by the relationship as

$$n = \frac{p(1-p)K^2}{E^2}$$

where

n = minimum sample size

p = proportion of the sample representing a specific event as a decimal value

K = constant corresponding to the desired confidence level

E = permitted error in the proportion, p

The value of K for various confidence levels is given in Table 17.2. For example, with a value of $p = 0.5$, $E = 5\%$ and hence, from Table 17.2, $K = 2.0$ the number of data points required in the sample set with a confidence level of 95% is

$$N = \frac{0.5(1-0.5)(2.0)^2}{(0.05)^2} = \frac{0.025(4)}{0.0025} = 400$$

Table 17.2 Constant *K* for various levels of confidence

Constant (K)	*Confidence level* (E)
1.00	68.3
1.50	86.6
1.64	90.0
1.96	95.0
2.00	95.5
2.50	98.85
2.58	99.0
3.00	99.7

17.3 PROBABILITY

17.3.1 Introduction

Probability is defined as the occurrence of an event relative to all other events. There has to be more than one possibility because otherwise the problem is deterministic. Probability is thus a numerical expression of the likelihood of an event relative to a set of alternative events. The formulation of a probabilistic problem commences with the definition of the full set of all possibilities – the sample or possibility space. Probabilities are then assigned to each of the specific events within that space (Ang and Tang, 1975).

Sample spaces may be discrete or continuous. Discrete sample spaces have a finite number of sample points, which are countable and non-infinite, within them. Continuous sample spaces have an infinite number of sample points regardless of the size of the sample space. Numbers of vehicles waiting to turn at an intersection are an example of a discrete sample space whereas the distribution of speeds from the slowest to the fastest may contain any value between the two limits and is thus continuous.

17.3.2 Characteristics of events

Events are studied in depth in set theory and presented in the form of Venn diagrams. This theory may be considered in other texts but is not dealt with in detail here. However, if probability is to be properly understood, there has to be at least some understanding of the basic elements of set theory.

Events of interest may often be combinations of other events. For example, materials for concrete bridge construction would require the delivery to site of cement, sand, coarse aggregate and reinforcing steel, all of which should be on site as foundation excavation comes to completion. The combination of these events is another event and described as the intersection of events and is presented by the notation $E_1 \cap E_2$ meaning that both events, E_1 and E_2, have to be present for the combination to occur. Another form of combination is the union of two events, meaning either of the original events separately or both simultaneously are present. This is denoted by the notational $E_1 \cup E_2$ in the context of Boolean 'or' which is 'and/or'.

Special events include the

- Impossible event, often denoted by the symbol φ, which is the event with no sample point, an empty set within the sample space
- Certain event denoted by S, which is the event containing all the sample points, in other words the sample space itself
- Complementary event denoted by $\bar{E}$, which for an event E contains all the sample points in the sample space other than those relating to the event
- Mutually exclusive events where the occurrence of one event precludes the occurrence of any other events, which is the situation where alternative routes between a common origin and destination are being evaluated
- Collectively exhaustive events, where the union of these events constitutes the entire sample space so that no other events are possible

The following symbols describe the operational rules governing sets of events:

- $\cup$ Union
- $\cap$ Intersection

- $\subset$ Contained
- $\supset$ Superset or 'contains'
- $\bar{E}$ Complement of E

17.3.3 The operational rules of events

These rules govern the

- Equality of sets
- Complementary sets
- Commutation
- Association
- Distribution

and include de Morgan's rule.

17.3.3.1 Equality of sets

Two sets are equal if and only if both sets contain exactly the same sample points so that

- $A \cup \varphi = A$
- $A \cap \varphi = \varphi$

17.3.3.2 Complementary sets

The relationship between complementary sets is given as

- $E \cup \bar{E} = S$
- $E \cap \bar{E} =$
- $(\bar{E}) = E$ (which means that the complement of the complementary event is the event itself)

17.3.3.3 The commutative rule

The union and intersection of sets is commutative in the sense that they are insensitive to sequence so that

- $A \cup B = B \cup A$
- $A \cap B = B \cap A$

17.3.3.4 Associative rule

The union and intersection of sets are associative, meaning that the sequence in which operations are carried out has no effect on the outcome.

The sets $(A \cup B) \cup C$ and $A \cup (B \cup C)$ are equal.

17.3.3.5 Distribution rule

The union and intersection of sets are distributive in a manner appearing similar to the addition and multiplication of conventional algebra in terms of the hierarchy of algebraic operations although not to be confused with them. Similarly, there are operations in set theory that have no equivalent in algebra. In consequence

- $(A \cup B)\ C = AC \cup BC$
- $(A \cap B)\ C = (A \cup C)(B \cup C)$

17.3.3.6 de Morgan's rule

This rule relates sets and their complements in the form that states

$$\overline{E_1 \cup E_2} = \bar{E}_1 \cap E_2$$

What it amounts to is that the intersection of the complements of the sets is equal to the union of the respective sets.

17.4 THE MATHEMATICS OF PROBABILITY

Building on the exposition of events contained in the previous section, there are three axioms that apply to the probable occurrence of events:

1. For every event E in a sample space S, the probability of the event is greater than zero.

   $$P(E) \geq 0$$

2. The probability of the certain event, S, is equal to 1.

   $$P(S) = 1$$

3. For two mutually exclusive events, the union of their probabilities is equal to the sum of their individual probabilities.

   $$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

From the first two axioms it follows that the probability of an event falls in the range between 0.0 and 1.0 so that $0 \leq E \leq 1$.

Applying the associative rule, the third axiom states that if an event E_1 occurs n_1 times and another mutually exclusive event E_2 occurs n_2 times then E_1 and/or E_2 would have occurred $(n_1 + n_2)$ times.

The multiplication rule defines a conditional probability, where the probability of an event depends on the occurrence or non-occurrence of another event. This probability is denoted $P(E_1|E_2)$, which means the likelihood of an event occurring assuming that it belongs to E_2. In terms of the nomenclature adopted in Section 19.3.2, E_2 is the 'super set' and becomes the reconstituted sample space. The probability that E_1 will occur given that E_2 has occurred is

$$P(E_1 | E_2) = \frac{P(E_1E_2)}{P(E_2)}$$

where E_1E_2 has the same significance as $E_1 \cap E_2$. This equation for conditional probability can be rearranged as

$$P(E_1E_2) = P(E_1)P(E_2)$$

which implies that the two events are also statistically independent.

Finally, the theorem of total probability is required if an event, A, cannot be predicted directly but its occurrence is always accompanied by other events, E_1, E_2, E_3 ...E_n, which are mutually exclusive. In combination, these events become the reconstituted space, S, in which event A occurs so that

$$P(A) = P(AE_1) + P(AE_2) + \cdots + P(AE_n)$$

(where AE has the same significance as $A \cap E$).

From the multiplication rule, this can be rewritten as the theorem of total probability defining the probability of the occurrence of A in the original sample space and given as

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \cdots P(A|E_n)P(E_n)$$

17.5 RANDOM VARIABLES

In all branches of engineering, many phenomena are associated with the numerical outcome of some or other physical quantity proportionally in relation to a sample space. One example is the flood level of a river above the mean flow level, which is a continuous event.

However, there are some outcomes that cannot be expressed numerically, such as the state of completion of a road construction project after a given period such as a month or a year. These can be assigned numerical values artificially. As an example, after the given period, the construction project could be definitely complete, definitely not started or at a defined level of completion between these two extremes by arbitrarily attaching numerical values to each condition, going, say from 1 for definitely complete to 5 for definitely not started, with the intervening values rated as completion being poor, fair, or good with values of 2, 3 or 4. In short, a random variable may be a device to identify events in random terms such that the event X may have the probability of a value a. This would be expressed mathematically as $P(X = a)$. $P(a > X \leq b)$ is the probability of a variable falling anywhere in the range of a to b.

As a random variable represents an event, it follows that it can assume a value only if associated with a probability or probability measure. The rule for describing a probability measure associated with all values of a random variable is described as a probability distribution. The probability distribution is invariably described by its cumulative distribution function (CDF) expressed as $F_X(x) \equiv P(X \leq x)$ for all x. The significance of this notation is that the random variable is indicated by an uppercase letter, in this case X, and the value of the event with the corresponding lowercase letter, x.

X is a discrete random variable if and only if certain values of x have positive probabilities. A discrete random variable can also be expressed by its probability mass Function (PMF), which is a function of expressing $P(X = x)$ for all x. The probability distribution function of $(X = x)$ is then expressed as

$$F_X(x) = P(X \leq x) = \sum_{\text{all } x_i \leq x} P(X = x_i)$$

$$\text{also written as} = \sum_{\text{all } x_i \leq x} p_X(x_i)$$

Alternatively, X could be a continuous random variable if probability measures are defined for any value of x. In this case, the probability law is described in terms of a probability density function (PDF) so that, if $f_X(x)$ is the PDF of X, the probability of X within the range of events from a to b is expressed as

$$P(a < X \leq b) = \int_a^b f_X(x)\,dx$$

To sum up, any function used, artificially or otherwise, to represent the probability distribution of a random variable has to satisfy the axioms of probability, namely that

- $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$
- $F_X(x) \geq 0$ and increases with increasing x
- $F_X(x)$ is continuous with x

and any function possessing these properties is a bona fide cumulative distribution function.

17.6 ANALYTICAL MODELS

17.6.1 Introduction

It has been found that the probability of random variables can take on characteristic forms, broadly being described as linear, exponential, logarithmic, S-curves and so on. Importantly, a random variable will always take on one and only one of these forms and not jump from one form to another. Plotting the cumulative density curve of a particular variable will reveal the basic shape to be employed. A misunderstanding that frequently arises amongst non-statisticians is that 'a data set is linear or logarithmic or whatever'. This is not true. It is the model representing the data set that has the derived form and not the data set itself.

Every distribution has two major defining characteristics. These are its mean and its variance. The variance, expressed as σ^2, provides an indication of the dispersion of the distribution around its mean. Very often, reference is made to the standard deviation, σ.

17.6.2 Regression

It is often useful to know what the relationship between two variables is so that the value of one variable can be predicted by the value of others. Examples include the volume of sales of a product in relation to its price, referred to by economists as price elasticity. In traffic engineering, the speed of vehicles as a function of traffic volume is an example.

The relationship may be between a dependent variable and more than one independent variable. The expected number of trips per day generated by a family, for example, could possibly be predicted by the number of vehicles it owns, its distance from the central business district (CBD), the population density where it is physically located and the family income. In a case such as this, care would have to be taken that the independent variables are truly independent. The number of vehicles owned would almost certainly depend on the extent of the family income. Location relative to the CBD may be more a matter of lifestyle and the ability to exercise personal choice, which, in turn, also derives in large measure from income level.

Although it is not possible to exactly predict the value of one variable in terms of another, it is possible to predict averages of expected values, for example, the average runoff from a catchment area for various rainfall intensities. The basic problem is that of determining the conditional expectation, $E(y|x)$, namely the average value of y given that x has occurred. With a single independent variable, reference is made to bivariate regression. The number of trips generated by a family as a function of more than one independent variable is an example of multivariate regression, $E(y|a, b, c)$.

The term 'regression' was coined by Francis Galton (1822–1911), who employed it in a study of the heights of fathers and sons in which he noted a regression (a turning back) from the heights of the sons to that of their fathers.

17.6.2.1 Method of least squares

The regression equation is linear when it has the form $E(y|x) = \alpha + \beta x$, with α and β being constants, called the regression coefficients. The attraction of the linear regression is that it lends itself to further mathematical treatment and that it often provides a good approximation to otherwise complicated regression equations. It is then necessary to know what the values of the parameters α and β actually are.

This can be addressed by using the method of least squares. This is best illustrated by use of an example. Assume that the weight losses of 10 persons after being on a weight-reducing diet for different periods are as shown in Table 17.3 (Freund, 1972). The number of months on the diet (the independent variable) is ordered from lowest to highest and the linear regression carried out. Plotting the weight losses against the appropriate duration of the diet in each case, it would appear that a straight line could provide a reasonably good fit with the data as shown in Figure 17.3.

The method of least squares is a method of curve fitting developed by the French mathematician Adrien Legendre and is used to determine the value of the parameters α and β of the proposed model. The quantity to be determined and minimised is $\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2$. Minimising of the expression is achieved by setting the differential of the expression to

Table 17.3 Weight loss

n	*Months on diet* X	*Actual weight loss (kg)*	x^2	xy	*Modelled weight loss (kg)*
1	1	0.5	1	0.45	2.35
2	4	7.7	16	30.91	7.08
3	4	5.0	16	20.00	7.08
4	7	10.9	49	76.36	11.82
5	9	17.3	81	155.45	14.97
6	10	20.5	100	204.55	16.55
7	12	18.2	144	218.18	19.71
8	14	24.1	196	337.27	22.86
9	17	29.1	289	494.55	27.60
10	22	32.3	484	710.00	35.49
Sum	100	165.5	1376	2247.73	165.49
			10*x	1654.55	
			Difference	593.18	
			β	1.58	
			α	0.77	
			Equation	$y = 0.77 + 1.58x$	

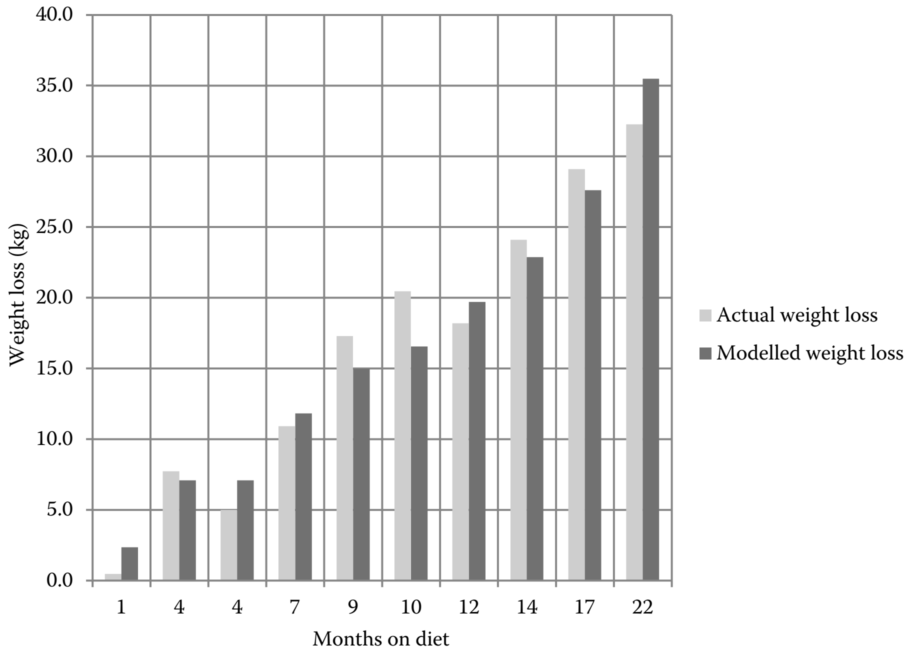


Figure 17.3 Plot of data of weight loss.

zero. Partial differentiation with reference to α and β and equating the derivatives to zero is expressed by the relationships

$$\sum_{i=1}^{n}(-2)[y_i - (\alpha + \beta x_i)] = 0$$

and

$$\sum_{i=1}^{n}(2)x_i[y_i - (\alpha + \beta x_i)] = 0$$

It is, however, more convenient to derive the values of the regression coefficients, α and β, from solving the normal equations

$$\sum_{i=1}^{n} y_i = \alpha n + \beta \cdot \sum_{i=1}^{n} x_i$$

$$\sum_{i}^{n} x_i y_i = \alpha \cdot \sum_{i=1}^{n} x_i + \beta \cdot \sum_{i=1}^{n} x_i^2$$

Reverting to the data set regarding weight loss as set out in Table 17.3,

$$n = 10,\ \sum_{i=1}^{n} x_i = 100,\ \sum_{i=1}^{n} y_i = 364,\ \sum_{i=1}^{n} x_i \cdot y_i = 36{,}400, \text{ and } \sum_{i=1}^{n} x_1^2 = 1376$$

Inserting these values into the above relationships yields

$$364 = 10\alpha + 100\beta$$

and

$$4945 = 100\alpha + 1376\beta$$

Solving these simultaneous equations leads to

$$\alpha = 1.7 \quad \beta = 3.47$$

so that the equation of the least squares regression is

$$y = 1.7 + 3.47x$$

This, it must be noted, does not give us any proof that a straight line provides the best fit of the data. What is actually being said is that, **if** a straight line **does** provide the best fit, the equation of the line that provides this is given by the above equation as deduced from the application of the theory of least squares.

The reason for adopting the square is that the addition of a series of positive and negative numbers would result in a value equal to or close to zero whereas what is required is, in effect, the total length of the differences between the model and the individual elements of the data set – the absolute value – but with the advantage that the square is arithmetically more convenient. This is also shown in Figure 17.3. Although what is illustrated above is the case of a single independent variable, it can be applied to multiple independent variables and applied equally to any probability distribution.

A comparison between the values of the observed data and the least squares regression may show that the scatter of the data may be wide or narrow. The question could then legitimately be asked of the legitimacy of the regression, that is, the extent to which it really models the data. Stated differently, what in reality is the correlation between the dependent and the independent variable? The sample correlation coefficient, normally indicated by r, will indicate what the extent of correlation is. If $r = 0$, there is no correlation between the two variables and they are independent. As the value of r increases towards $r = 1$, the probability of obtaining a value of y outside a strip containing the regression line $y = \alpha + \beta x$ reduces towards zero suggesting that there is a strong relationship between x and y.

The value of r is calculated by use of the formula (Freund, 1972)

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \quad \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i - \sum_{x=1}^{n} x_i{}^{2}} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i{}^{2}}}$$

The concept of the maximum likelihood estimate, $\bar{\sigma}$, will not be discussed here. It addresses the issue of correlation as opposed to regression and its value can be determined from the equation

$$\bar{\sigma}^2 = (1 - r^2)\hat{\sigma}_2^2$$

where $\hat{\sigma}_2$ is the maximum likelihood estimate and r is the sample correlation coefficient.

This equation simplifies the calculation of the maximum likelihood estimate and also links the concepts of regression and correlation. The value $\hat{\sigma}_2^2$ measures the total variation in the values of y and $\hat{\sigma}_2$ measures the conditional variation of the y's for fixed values of x. It follows that $\hat{\sigma}_2^2 - \hat{\sigma}^2$ measures that part of the total variation of y that is explained by its relationship with x. Thus if $r = 0.6$, then 36 per cent of the variation of y is accounted for by its correlation with x. A value of $r = 0.9$ suggests that 81 per cent of the variation of y is related to x. It can thus be stated that a 'value of r of 0.9 is more than twice as strong as a value of r of 0.6'. In the case of traffic engineering, the range of variation of the characteristics of drivers is so wide that values of r (or r^2, which is the value normally quoted) are normally low.

Freund illustrates the application of the sample correlation coefficient with an example whereby the time taken by a secretary to type a certain form in the afternoon is compared with that taken in the morning. Ten samples are taken in each case, as shown in Table 17.4.

Inserting the values of n, ● x, ● x^2, ● y, ● y^2 and ● xy into the above equation provides the value of r as being equal to 0.94. This implies that there is a strong correlation between the times taken in the morning and the afternoon to type the forms, with the morning time as the independent variable and the time in the afternoon the dependent variable.

This example also demonstrates the spurious correlation, '*Post hoc, ergo propter hoc*', loosely translated as 'After this, therefore because of this'. An example of this is 'the cock crows, the sun rises, the cock crows therefore the sun rises'. There cannot be any rational causative effect directly linking the two data sets. However, they suggest that it takes the secretary slightly longer in the afternoon to type the same number of forms as in the morning. It is possible that the real correlation is between the time taken to type the forms and the time of day. It could be that, as the day wears on, the secretary becomes more tired and slows down.

It sometimes happens that researchers 'improve' their research results by trimming off or 'losing' outlying values of the dependent variable until the point is reached where the claim

Table 17.4 Times taken for typing a specified form

n	*Morning* x	*Afternoon* Y	x^2	y^2	xy
1	8.2	8.7	67.24	75.69	71.34
2	9.6	9.6	92.16	92.16	92.16
3	7.0	6.9	49	47.61	48.3
4	9.4	8.5	88.36	72.25	79.9
5	10.9	11.3	118.81	127.69	123.17
6	7.1	7.6	50.41	57.76	53.96
7	9.0	9.2	81	84.64	82.8
8	6.6	6.3	43.56	39.69	41.58
9	8.4	8.4	70.56	70.56	70.56
10	10.5	12.3	110.25	151.29	129.15
Σ =	86.7	88.8	771.4	819.3	792.9

can be made that the null hypothesis supports whatever pet theory is being propounded. The temptation may be strong but must be resisted.

17.6.3 Selection of a model

As an example of the application of the selection of a model, our secretary also produces typographical errors in typing a document. The numbers of errors made per page typed of a document are listed in Table 17.5. This is presented graphically in Figure 17.4.

It would appear possible that this data set could be modelled by the Poisson distribution. The Poisson distribution, discussed later, has the form

$$P(X_t = x) = \frac{(vt)^x}{x!} e^{-vt} \quad \text{for } x = 0, 1, 2\ldots$$

Table 17.5 Data set of typographical errors

Number of errors	*Total number of errors*	*Observed frequency*	*Poisson probability with v = 3*	*Expected frequency* e_i
0	0	18	0.05	21.912
1	53	53	0.15	65.736
2	206	103	0.22	98.56
3	321	107	0.22	98.56
4	328	82	0.17	73.92
5	230	46	0.10	44.352
6	108	18	0.05	22.176
7	70	10	0.02	9.504
8	16	2	0.01	3.564
9	9	1	0.00	1.188
	1341	440		
	v =1341/440			
	= 3.047			

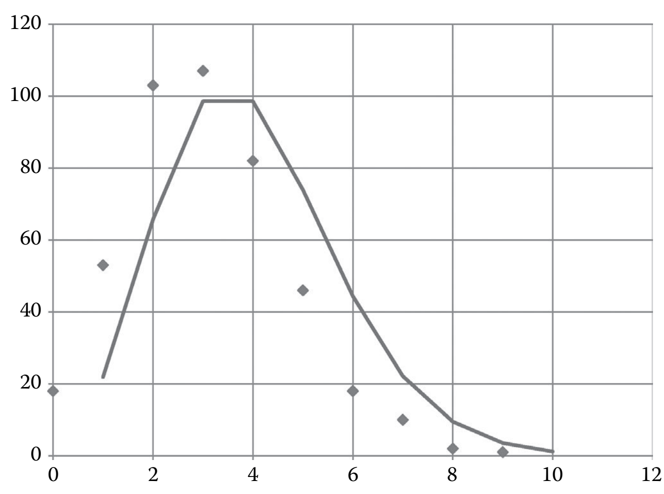


Figure 17.4 Distribution of typographical errors.

where v is the mean occurrence rate or the average number of occurrences of the event per unit time. The mean occurrence rate for this data set is $v = \frac{1341}{440} = 3.05$. Being a discrete data set, values cannot be fractional. On average there are three typographical errors per page. The expected frequency is shown in the right-hand column of Table 17.5 and in Figure 17.4.

17.6.4 Testing of a model for goodness of fit

The extent of the congruence between the data set and its model is typically reflected in measures of 'goodness of fit', which means that what is to be determined is whether a data set may be looked upon as values of a random variable having a given distribution.

Measures of goodness of fit include the χ^2 (chi-squared) criterion and the Kolmogarov-Smirnoff (K-S) test. These are discussed in Section 17.7. Goodness of fit is determined by use of the null-hypothesis test of the findings of the method of least squares, namely that the differences observed between the individual values of the distribution of the data set and those of the observed set are independent. In short, is the difference between the two data sets sufficiently close to zero? If the assumed model fails, the null hypothesis is rejected and another model is to be sought.

17.6.5 Degrees of freedom

Few people seem to have any interest in having an understanding of the statistical construct of degrees of freedom and are satisfied simply to be able to calculate and report it. The degrees of freedom can be explained by taking any set of, say, three numbers. Calculating the mean is simply the sum of the numbers divided by the number of terms in the set. On the other hand, if the mean is given, any two numbers can be freely chosen. The third term is, however, fixed by the mean and the selected term. The set is then said to have two degrees of freedom. The general rule is thus that, for any set, where n equals the number of values in the set, the degrees of freedom equal $n - 1$.

One of the needs for being able to specify the number of degrees of freedom is that it features in the calculation of goodness of fit in terms of the Student t-test discussed later.

17.6.6 Confidence level

Statistical reports often include the statement that 'the chosen distribution with a mean of x and a standard deviation of y reflects the data set with a confidence level of z%'. This can be visualised as a band of a specified width surrounding the distribution that has the calculated mean and standard deviation. The confidence level could be a low value such as 5% or its complement, 95%. In the first case, the value has the significance that the mean of any other sample set could fall *outside* the confidence band 5% of the time and, in the latter case, the mean of any other data set would be *within* the confidence band 95% of the time.

The bandwidth is a matter of choice and the statistician would have to select a value of bandwidth that reflects the importance or otherwise of departing from the chosen value of mean and standard deviation. A patient about to undergo major surgery would not be happy with an airy statement that 'This operation is successful in about 50% of the cases'. A value much closer to certainty would no doubt boost the patient's feeling of confidence.

The width of the confidence band is dependent on the confidence level selected. If no sample set were to have a mean outside the confidence band, the inference would be that the confidence band would have to be very wide. As the acceptable percentage of failing sample sets increases, the band width could decrease.

The greatest level of confidence in the value of a mean is, however, in the case where very few sample sets fail AND the bandwidth is narrow. It must be realised that this level of precision carries a penalty, specifically that of the cost necessarily incurred to achieve it. In setting up a sampling procedure, a manufacturer would have to decide what level of risk is acceptable, for example, what the cost of an out-of-court settlement could be. For light bulbs, a large number of failures would probably be acceptable whereas, in the case of pharmaceutical products, very few failing sample sets would be acceptable and the range of the bandwidth would also have to be narrow.

The width of the confidence band, in other words its boundary values, is calculated by establishing the mean and standard variation as

$$\bar{E} = (E_1 + E_2 + E_3 + \cdots + E_n)/n$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(E_i - \bar{E}\right)^2$$

If the sample set has a normal distribution, then

$$t = \frac{\bar{e} - }{\frac{s}{\sqrt{n}}}$$

where

$\bar{e}$ = the mean of the data set
μ = the mean of the distribution
s = the standard deviation of the distribution
n = the size of the random sample

has the Student t-distribution with $(n - 1)$ degrees of freedom. The name of this distribution derives from its having been obtained by WS Gosset, who published his research under the pen name ‘Student’ (Freund, 1972).

It is to be noted that the confidence band is symmetrical around the statistical mean, so that a confidence level of 5% implies a range of values spanning from μ – 2.5% to μ + 2.5%. If a value, μ, is selected, there is a 2.5% chance that T will be less (or 97.5% more) than –μ and a 2.5% chance that it will be larger than (or 97.5% less) +μ. Thus, the probability that T will be between –μ and +μ is 95%.

From the sample, values of $\bar{e}$ for E and s for S are calculated and the boundaries of the confidence level are thus

$$\bar{e} - \frac{}{\frac{s}{\sqrt{n}}}, \bar{e} + \frac{}{\frac{s}{\sqrt{n}}}$$

17.7 PROBABILITY DISTRIBUTIONS

17.7.1 Introduction

Engineers use probability distributions to analyse data describing observed behaviour or to be able to predict the likelihood of future occurrences. Data may be ordered from lowest to highest, such as, for example, in the case of reaction time, or traffic flow levels from the highest to the lowest across the duration of a year, and a distribution sought thereafter that replicates the resulting curve to an acceptable level of accuracy. Once found, this distribution can then be used to predict the frequency of future events of occurrences.

Various distributions have found useful application in traffic flow theory and geometric design. The Poisson distribution has been demonstrated in the preceding text. Other distributions include the

- Normal distribution, also known as the Gaussian distribution
- Log-normal distribution
- Binomial distribution
- Negative binomial distribution
- Geometric distribution
- Exponential distribution
- Gamma distribution
- Hypergeometric distribution
- Beta distribution

The χ^2 distribution and its application in statistical analysis have already been demonstrated. The *t*-distribution, also known as the Student's *t*-test, is also important in analysis. These distributions are discussed in the sections that follow.

17.7.2 The Poisson distribution

The Poisson distribution is based on the assumptions that

- An event can occur at random and at any time or point in space
- The probability of occurrence of an event in a small interval of time is proportional to the duration of the interval and can be given by $v \cdot \delta t$ where v is the mean rate of occurrence of the event
- The probability of two or more occurrences in δt is negligible
- The occurrence of an event in an interval is independent of that in any other nonoverlapping interval

The flipping of a coin to get a series of heads and tails is an example of this distribution. Another example is the number of vehicles waiting to make a left turn at an intersection.

If X_t is the number of occurrences in a time or space, t, the Poisson distribution has the form

$$P(X_i = x) = \frac{(vt)^x}{x!} e^{-vt} \quad \text{for } x = 0, 1, 2\ldots$$

where v is the mean occurrence rate, that is, the average number of occurrences of the event per unit time.

Table 17.6 Cumulative number of waiting vehicles

Number of waiting vehicles (x)	*1/x!(120/60)^x exp(−120/60)*	*Cumulative probability*
0	0.135335283	0.135
1	0.270670566	0.406
2	0.270670566	0.677
3	0.180447044	0.857
4	0.090223522	0.947
5	0.036089409	0.983
6	0.012029803	0.995
7	0.003437087	0.999

As an example, a road authority may decide that there should be a 95 per cent probability that the left turn lane should be long enough to avoid traffic spilling over into the through lane. The question to be resolved is the number of vehicles waiting to turn left that have to be stored in any one cycle to ensure a sufficient length of auxiliary lane when the average turning flow is, say, 120 vehicles per hour. With a 60-second cycle length, the average number of turning vehicles in any one cycle is 2.

The calculation is shown in Table 17.6. It shows that providing for the storage of four vehicles is barely sufficient and that storage for five vehicles would at least provide a modest margin for error. Assuming that the design vehicle is a passenger car and allowing about 1 metre of clear space in front of each vehicle, the turning lane should have a length of about 40 metres.

The Poisson distribution is useful in the modelling of random arrivals. Zero arrivals in an interval constitute the gap that is available to drivers on a minor road to cross or join the major road at a priority-controlled intersection. A gap may be measured as the distance between the back of a vehicle and the front of the following vehicle, although it is more often measured in units of time. In this case, reference is made to headway and is the interval between the arrivals at a point on the road of any common points of two vehicles, for example, their front or rear overhangs. Headway thus includes the length of a vehicle, in the sense of the travel time elapsed between its front and rear overhangs passing a fixed point on the road. In the case of a vehicle arriving at an intersection between the passage of two vehicles along the major road, reference is made to a lag. A lag is thus the unexpired portion of a gap.

At unsignalised intersections, pedestrians need gaps in the traffic in order to cross the street. As traffic increases, the time in which pedestrians could cross the street will reduce and it will ultimately be necessary to apply traffic control to the crossing. In its simplest form, traffic control could be the alternating black and white stripes on the road surface, universally known as a zebra crossing. In some countries, notably the United Kingdom and the other Commonwealth countries, this grants right of way to pedestrians whereby all vehicles are required to stop in the presence of pedestrians. At higher traffic volumes, this requirement may create a delay unacceptable to the drivers of moving vehicles and they may elect to ignore it. It would then be necessary to force them to stop by providing traffic signals. These may be push-button controlled or controlled by the normal methods such as fixed time or vehicle actuation used at signalised intersections between vehicles.

The application of the Poisson distribution to zero arrivals in this example takes the form

$$\frac{3600}{t}\ e^{-vt} = 60$$

Table 17.7 Traffic flows warranting traffic control

Number of lanes	*Distance to be walked (m)*	*Flow*	
		Per second	*Per hour*
1	3.6	1.111	3999
2	7.2	0.430	1549
3	10.8	0.238	857
4	14.4	0.153	549
5	18	0.106	381
6	21.6	0.077	278

because, in the case of the event of zero arrivals, $x = 0$ so that the term, $\frac{(vt)^x}{x!} = 1$. The probability of one acceptable gap on average in every minute of the hour is expressed as 60 gaps per hour. With a gap duration of t, the hour can be split into $3600/t$ intervals. The required gap is a function of the distance, d, in metres to be walked and the speed of walking, say 1.3 m/s, so that $t = \frac{d}{1.3}$.

The expression above thus becomes $\frac{3600 * 1.3}{t} e^{-v\frac{d}{1.3}} = 60$ from which

$$v = \frac{1.3}{d} \ln \frac{3600 * 1.3}{60d}$$

The vehicular flows above which traffic control is required at a pedestrian crossing are offered for various widths of street, where the width of individual lanes is 3.6 metres, in Table 17.7.

The provision of a median island would reduce the distance to be walked to half that of an undivided street. A single-lane chicane with an island separating the two directions of flow through the chicane would reduce the distance to be walked by half again. In each case the traffic flow warranting the use of traffic control increases.

17.7.3 The normal distribution

The normal distribution is arguably one of the most important of the statistical distributions and was first discovered in the 18th century when scientists observed a degree of regularity in errors of measurement. The distributions they observed were closely modelled by what they referred to as the 'normal curve of errors' and that was attributed to the laws of chance. The French mathematician A de Moivre was the first to investigate the properties of this curve and its theoretical basis. It was subsequently studied further by Carl Friedrich Gauss; in his honour the distribution is often referred to as the Gauss distribution although it is usually known as the normal distribution because so many statistical phenomena appear to follow it (Steyn et al., 1984).

Whereas the Poisson distribution is used to model discrete events, the normal distribution models continuous events. Its probability density function is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \quad -\frac{1}{2} \quad \frac{x-}{\sigma} \quad ^2$$

The graph of this distribution is symmetrical around $x = 0$, otherwise the y-axis, and extends from $-\infty$ to $+\infty$ and is shaped like the cross-section through a bell. The values μ and σ are the mean and standard deviation respectively of the distribution, with the mean being the maximum point of the distribution and appearing on the y-axis. If the mean, μ, equals 0 and the variance, σ^2, equals 1, the normal distribution is known as the standard normal deviation and indicated as the $N(0; 1)$ distribution.

Being a continuous as opposed to a discrete function, the cumulative probability that the variable, X, will fall between two values, a and b, is expressed by the integral

$$P(a < X \le b) = \frac{1}{\sigma\sqrt{2\pi}}\int_a^b \exp\ -\frac{1}{2}\ \frac{x-}{\sigma}^{\ 2}\ dx$$

If people don't remember the calculus that they studied at university the direct calculation of the integral may prove difficult. A change of the variables to $s = \frac{x-}{\sigma}$ and $dx = \sigma ds$ will make it possible to use the table of standard normal probability, which is

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp\ -\frac{1}{2}s^2$$

to solve the probability as

$$P(a < X \le b) = \Phi\ \frac{b-}{\sigma}\ -\Phi\ \frac{a-}{\sigma}$$

where, as stated previously, the symbol Φ has the meaning of 'distribution function' so that $\Phi(s)$ is the distribution function of the standard normal variate, S.

To illustrate: if the annual rainfall in a certain catchment area has a normal distribution N (600 millimetres, 150 millimetres), the probability that the rainfall could fall between 800 and 1000 millimetres in future years could be calculated as

$$\begin{aligned} P(800 < X \le 1000) &= \Phi\ \frac{1000-600}{150}\ -\Phi\ \frac{800-600}{150} \\ &= \Phi(2.6667) - \Phi(1.3333) \\ &= 0.996 - 0.908 \\ &= 0.088 \end{aligned}$$

The probability of above normal rainfall in the range of 800 to 1000 millimetres is thus only 8.8%.

The normal distribution bears a passing resemblance to the Poisson distribution and can even be forced to approximate it by taking shorter and shorter probability intervals until the stage is reached whereby the likelihood of more than one event in the probability interval is very small.

17.7.4 The log-normal distribution

The log-normal distribution arises if ln X (the natural or Naperian logarithm) is normal. In this case, the relationships offered in the preceding section become

$$f_X(x) = \frac{1}{\zeta\sqrt{2\pi}} \exp \; -\frac{1}{2} \; \frac{x-\lambda}{\zeta} \; ^2$$

where λ (lambda) is the mean of ln X and ζ (zeta) is the standard deviation of ln X and

$$P(a < X \leq b) = \Phi \; \frac{b-\lambda}{\zeta} \; - \Phi \; \frac{a-\lambda}{\zeta}$$

The parameters λ and ζ are related to μ and σ as

$$\lambda = \ln \quad -\frac{1}{2}\zeta^2 \quad \text{and} \quad \zeta^2 = \ln \;\; 1 + \frac{\sigma^2}{\;^2}$$

With these substitutions, the calculations proceed, using the table of standard normal probabilities, as discussed in the preceding section. Because of the convenience of calculating the probabilities of log-normal variates and also because the variable is always positive, this distribution may be useful where it is known that the values of the variate will always be positive. Examples include rainfall intensity, as used in the preceding example, and traffic volumes.

17.7.5 The binomial (or Bernoulli) distribution

As its name suggests, this probability distribution applies when an experiment or trial has only two possible outcomes, one successful and the other not. If the probability of success is expressed as θ, the probability of failure is (1 – θ) and the probability function of the distribution can be written as

$$f(x;\theta) = \theta^x(1-\theta)^{1-x} \quad \text{for } x = 0, 1$$

The parameter θ is a constant referring to a specific binomial distribution. In the case of flipping a coin, θ equals 0.5, with 0 and 1 representing heads and tails respectively. If 2% of the apples in a barrel are rotten, θ = 0.02. Zero represents selecting a bad apple and 1 selecting a good one. It is to be noted that selecting a bad apple constitutes a success and a good apple is failure! This terminology derives historically from probability theory applying only to games of chance when one player's failure was the other's success (Freund, 1972). An experiment in which the binomial distribution applies is referred to as a binomial or Bernoulli 'trial' and the repetition of these experiments is 'repeated trials'. Repeated binomial trials play an important role in probability and statistics in the case where successive trials are independent, and the probability of success, θ, is the same for each trial.

With one factor, θ, representing each success and another, (1 – θ), for each failure, the product of all these outcomes $\theta^x(1-\theta)^{1-x}$ and $\genfrac{}{}{0pt}{}{n}{x}$ gives the number of ways in any order in which x successes can occur in n trials. The notation $\binom{n}{x}$ is called the 'binomial coefficient' and has the value $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$

The calculation of factorials can be tedious if there are several values of x. While 5! = 120, 10! = 3.626*10^6 and 15! = 1.307*10^{12}, a useful formula for calculating the factorials of large numbers is

$$n! \sim \sqrt{2\neq n}\left(\frac{n}{e}\right)^{n}$$

derived by James Stirling in the 18th century. This formula is accurate to within 1% even for values of n as low as 10. However, a more efficient way of calculating the probability of an event happening x times in n trials is by the use of recursion. Recursion is the process whereby the value of a probability is calculated from the value of the preceding term in the series. If

$$P(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

then

$$P(x+1) = \frac{n!}{(x+1)![n-(x+1)]!}p^{x+1}(1-p)^{n-(x+1)}$$

from which

$$P(x+1) = \frac{n-x}{x+1}\cdot\frac{p}{(1-p)}\cdot P(x)$$

17.7.6 The geometric distribution

The number of trials in a binomial distribution until a specified event occurs for the first time is modelled by the geometric distribution. If the event occurs during the tth trial, it follows that there are no events in the preceding $(t - 1)$ trials so that

$$P(T = t) = p(1 - p)^{t-1}$$

and this is known as the geometric distribution.

The number of trials until the first occurrence of an event is called the 'first occurrence time'. Because of the statistical independence between successive trials, it follows that the 'first occurrence time' must also be the time between any two consecutive occurrences of the same event so that the recurrence time is equal to the 'first occurrence time'. The recurrence time is generally referred to by engineers as the 'return period'.

This is widely used in hydrological calculations. For example, if a culvert is to be designed to be able to accommodate a flood with a 20-year return period, there may be some interest in knowing what the probability would be of this flood occurring within the maintenance period of 2 years after completion of the structure.

$$P(T \leq 2) = \sum_{t=1}^{2} (0.05)(0.95)^{(2-1)} = 0.0975$$

There is a probability of nearly 10% that the worst storm in 20 years could occur within the maintenance period of the culvert. From bitter experience, engineers know that the worst storm in 100 years could occur the day after tomorrow, with another just like it, 2 weeks hence.

17.7.7 The negative binomial distribution

The geometric distribution is the probability law governing the number of trials until the first occurrence of an event in a binomial sequence. The number of trials until a subsequent occurrence of the same event is then modelled by the negative binomial distribution.

If the kth occurrence of the event happens at the tth trial, there must be $(k - 1)$ occurrences of the event in the preceding $(t - 1)$ trials. From the binomial law it follows that

$$P(T_k = t) = \begin{matrix} t-1 \\ k-1 \end{matrix} \; p^k (1-p)^{t-k}$$

which is the negative binomial distribution.

17.7.8 The exponential distribution

The exponential distribution is a special case of the Poisson distribution. It models the situation of the time, T_1, until the first occurrence of the event so that $(T_1 > t)$ means that no event occurs in time, t. It follows thus that

$$P(T > t) = P(X_t = 0) = e^{-\nu t}$$

In traffic flow theory, this distribution finds application, for example, in the modelling of headways. These describe the successive arrivals of vehicles in a traffic stream for a given average flow in vehicles per time interval in terms of the temporal gap between them.

Figure 17.5 illustrates the distribution of headways from zero seconds to 15 seconds.

The shortest headway is shown as being zero seconds. In practice, this is impossible to achieve, as the headway includes the length of the leading vehicle. To remove this impossibility, one can invoke the shifted exponential distribution. The probability density function of the exponential distribution can start at any value and is expressed as

$$f_X(x) = \nu t \cdot e^{(-\nu t)(x-a)} \quad \text{for } x \geq a$$

$$= 0 \quad \text{for } x < a$$

where

ν = the average occurrence rate in time t
a = the extent of the shift of the exponential distribution

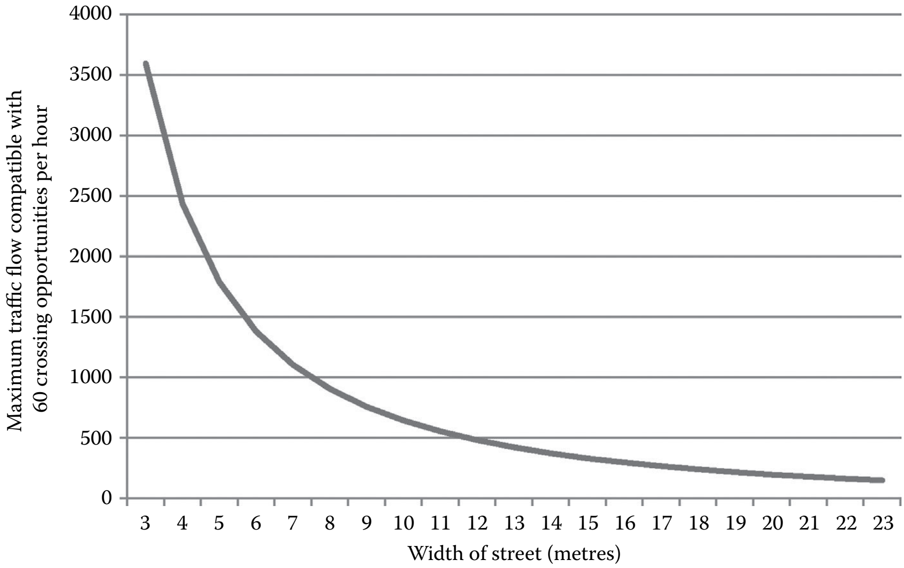


Figure 17.5 The distribution of gaps in a traffic stream.

17.7.9 The gamma distribution

The gamma distribution is similar in principle to the negative binomial distribution The exponential and gamma distributions are the continuous equivalents of the discrete geometric and negative binomial distributions. The exponential and gamma distributions are related to the Poisson process in the same way that the geometric and negative binomial distributions are related to the binomial or Bernoulli distribution.

The gamma distribution models the time until the kth event where the occurrences of events are in terms of a Poisson process. If T_k is the time till the kth event, then $(T_k \leq t)$ is the expression defining that at least k events occur in time t. The distribution function of T_k is

$$F_{T_k}(t) = \sum_{x=k}^{\infty} P(X_t = x)$$

$$= 1 - \sum_{x=0}^{k-1} \frac{(\upsilon t)^x}{x!} e^{-\upsilon t}$$

As an illustration: if a train arrives at a railroad crossing on average once every 6 hours, the time until the first arrival is given by the exponential distribution with $\upsilon = 1/6$ arrivals per hour. This and the times until the subsequent two arrivals are shown in Figure 17.6.

17.7.10 The hypergeometric distribution

The hypergeometric distribution often finds application in connection with quality control in the manufacturing sector, otherwise referred to as acceptance sampling. Samples in a population may test either to be good or to be bad. If the number of failing samples is less

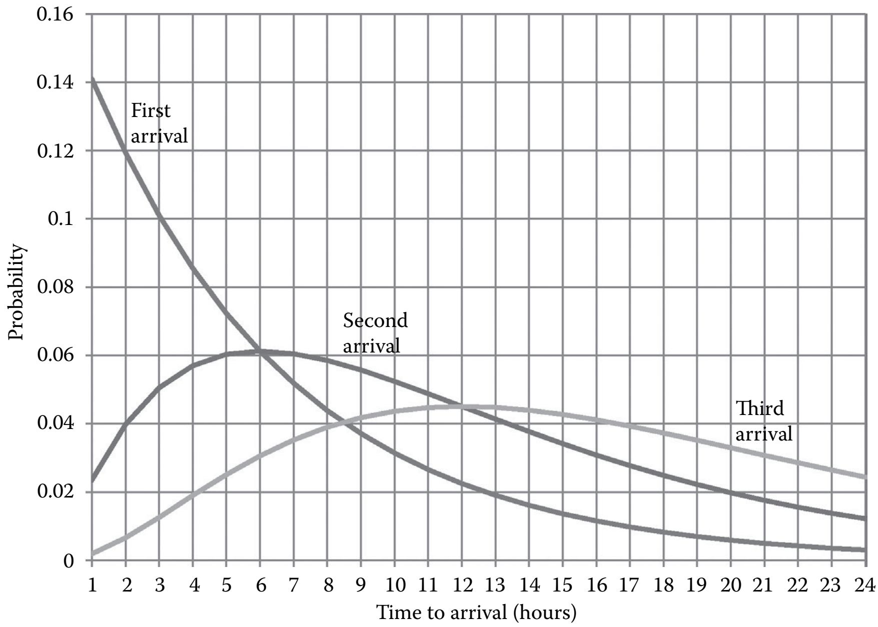


Figure 17.6 Arrivals at a railroad crossing.

than some or other standard value, the population from which the samples were taken can be accepted. A number of samples, say 100, would be taken from a production run of a product and tested. The standard set for the production run to pass scrutiny could be that, say, no more than two of the samples may fail. If the sample set includes three or more failing samples, the production run has to be rejected. Stated mathematically, the hypergeometric distribution addresses the probability of x successes in n trials.

If a population of samples comprises 'a' good samples and 'b' failing samples the total population set, otherwise the sample space, comprises (a + b) samples. The number of ways in which a subset, n, can be taken from the population set is expressed as $\binom{a+b}{n}$ so that the probability of selecting the subset, n, is $\dfrac{1}{\binom{a+b}{n}}$. The number of ways in which a subset of x elements can be selected from the set of 'a' or good samples is $\binom{a}{x}$. The number of ways of selecting '$n-x$' elements from a set of 'b' or failing samples is expressed as $\binom{b}{n-x}$. The probability of 'x' successes in 'n' trials is thus

$$f(x; n, a, b) = \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}} \quad \text{for } x = 0, 1, 2, \ldots, n$$

with the symbol $f(x; n, a, b)$ indicating that the value of x is dependent on the parameters n, a and b.

17.7.11 The chi squared test

A data set has been acquired and a preliminary judgment call has been made on the statistical distribution that best fits it. Using the chi-squared (χ^2) test, the selected distribution is rejected if

$$\chi^2 = \sum_{i=1}^{m} \frac{(f_i - e_i)}{e_i}$$

where
m = the number of terms in the model
f_i = the ith value of the observed data set
e_i = the ith value of the modelled data set

and the value that is obtained exceeds $\chi^2_{\alpha,n-1}$ where α is the selected confidence level of test and $(n - 1)$ is the number of degrees of freedom of the distribution. This value of χ^2 can be read off tables provided in the appendices to most statistics texts.

From the data offered in Table 17.3 the value of χ^2 is 7.8 whereas the value of $\chi^2_{0.05,6}$ from the tables is 12.59. The null hypothesis can thus not be rejected and the Poisson distribution adequately models the data set shown in Table 17.4.

17.7.12 The Kolmogorov-Smirnoff test

Whereas the chi-squared test is applied to the probability distribution, the Kolmogorov-Smirnoff (K-S) test compares the experimental cumulative distribution frequency with that of an assumed theoretical distribution. In traffic engineering, many distributions are presented in the form of the cumulative distribution function, typically where some or other percentile value is of importance. For example, the dimensions of the design vehicle are typically 85th percentile values of the entire data set, as are driver reaction times.

The K-S test requires the set of observed data to be rearranged in increasing order and a stepwise cumulative frequency function developed, usually in a normalised form, that is, with a range of values between 0 and 1. The selected theoretical probability distribution is similarly converted to a normalised cumulative function. The difference between the observed and the modelled probability of each value of the variable is determined. It is the maximum difference, D_n, that is the measure of the discrepancy between the observed data and the theoretical model.

$$D_n = \max_x \left| F(x) - S_n(x) \right|$$

This difference is compared to the critical value, D_n^α, of the K-S test, tabulated in numerous texts on statistics for various values of the confidence level, α. If the observed difference, D_n, is less than the critical value, D_n^α, the null hypothesis cannot be rejected so that the assumed distribution is acceptable at the selected confidence level.

The weakness of the K-S test is that it is based on one data point only. It is theoretically just possible that the vast majority of the differences less than the maximum could actually be either very close to either the maximum or to zero. The goodness of fit in these two extreme cases may thus not compare all that favourably with a test based on the theory of least squares where the entire data set contributes to the comparison. Another possibility is that the maximum difference could be to a data point that is an outlier, which would have the effect of distorting the succeeding portion of the cumulative frequency function. On the other hand, if the differences are normally distributed, these concerns may balance out.

Chapter 18

Traffic flow theory

CONTENTS

18.1 INTRODUCTION

Traffic flow theory is a very broad field and addresses geometric design as well as traffic operations. In this chapter, the focus is limited to those aspects of traffic flow theory that are relevant to the design of roads and the geometric standards that provide the upper and lower limits of the values that designers could adopt. Even with this limitation, it is not possible to provide more than the barest overview of the topic.

Eighty years ago, the discipline of traffic flow theory did not exist. In 1933, Dr. Bruce Greenshield propounded the theory that predicted and explained the trends that are observed in real life traffic flows. He theorised that, under uninterrupted flow conditions, there was a relationship between flow, speed and density and, with any of the qualities known, the third could be calculated. He started his career as a traffic engineering scientist with this publication which led to a PhD thesis at Michigan University in 1934 (Kühn, 2011). This became known to succeeding generations of traffic flow theorists as the Greenshield model and signalled the birth of traffic flow theory.

This model is discussed further in this chapter. It seeks to define the interaction among the driver, the vehicle and the road with the road comprising the links and nodes of the road network with its traffic signs, road markings and other control systems (Gartner et al., 2001). The road is static infrastructure and the vehicle without its driver is inert. The driver is the sole dynamic element of the road transportation system, deciding the direction and speed of movement of the vehicle and its location relative to the road surface and other vehicles. Obviously the driver with all his or her physical characteristics, frailties, desires and emotions is of interest in the pursuit of road safety and operational efficiency of the road network. Traffic flow theory thus seeks first to quantify these characteristics through the acquisition and analysis of data and second to develop statistical models that can be used to predict the behaviour of drivers.

As discussed in Chapter 4, historically geometric design was predicated on vehicular characteristics and vehicle/road interactions, in fact on what a vehicle **can** do. The only driver characteristics considered were driver eye height and reaction time as applied to sight distance in all its myriad forms. Since the late 1990s the emphasis has shifted to what the driver **wishes** to do. The change in approach towards human factors as a design input is a paradigm shift of note. It will almost certainly place added emphasis on the further development of the discipline of traffic flow theory.

In this chapter, the acquisition of data is discussed and the various traffic flow studies, namely those related to speed, flow and density are described. The statistical models that can replicate the observed data are also discussed.

18.2 DATA ACQUISITION

18.2.1 Introduction

The most fundamental measurement of traffic performance is that of traffic volume. It dictates where in the hierarchy of roads a particular road should be located and what its cross-section should be. Furthermore, a high volume of traffic would justify the use of generous standards of horizontal and vertical alignment.

Traffic counts are aimed at seeking a design that can accommodate the operational issues of efficiency of movement. Crash data, on the other hand, are aimed at enhancing the safety of the road. Road safety in developing countries compares poorly with that of First World countries, with crash rates per 100 million kilometres often being five times higher than that of the United States. In spite of this, even in developing countries, crashes are relatively infrequent occurrences so that acquiring a data set of an adequate size is difficult. Recourse is thus often taken in conflict measurement.

Speed measurements are indicative of the quality of the horizontal alignment of a road, specifically the level of consistency of the design as discussed in Chapter 4. Consistency of design has a profound impact on road safety.

Travel time and delay impact on the economic aspects of transportation. The majority of consumer goods move between supplier and consumer by road. Suppliers invariably pass transport costs along to consumers, and an inefficient road system where vehicles are subject to high levels of congestion and the ensuing delays will impact unfavourably on the cost of living of the entire population of a country.

This section of the chapter thus deals with the acquisition of data with respect to

- Traffic volumes in vehicles per unit of time
- Crashes and conflicts as dimensionless numbers
- Speeds as distances per unit of time
- Travel time and delay

18.2.2 Modes of measurement

Five different modes of measurement can be followed (van As and Joubert, 1993). These are measurement

- At a point
- Over a short distance, that is, less than about 10 metres
- Over a length of road, usually more than 0.5 kilometre
- By an observer moving in the traffic stream
- By samples obtained simultaneously from a number of vehicles

These are illustrated in Figure 18.1 using a space–time diagram. The vertical axis of the diagram represents the distance in the direction of travel from an arbitrary starting point. The horizontal axis represents the elapsed time from an arbitrary starting time. Each line within the graph represents the space–time trajectory of a vehicle as it moves down the road. The gradient of the trajectory has the unit of space–time, in other words the speed of the vehicle. The faster a vehicle is moving, the steeper is the gradient of its trajectory. Where lines cross, a faster vehicle has overtaken a slower one. Although it may appear that, contrary to Einstein's theories, two objects are occupying the same space at the same time, they are actually side by side during the passing manoeuvre and, without unnecessary three-dimensional complications, it is not possible for the space–time diagram to show this.

A horizontal line on the space–time diagram represents measurement at a point in space, that is, the distance from the starting point is unaltered but the time varies. Two horizontal lines close to each other thus represent a measurement over a short distance. A vertical line represents a snapshot like an aerial photograph of a length of road at a point in time. Two vertical lines close to each other represent measurement over a short time interval and the further apart they move the longer the time interval becomes. The combination of choices regarding the time and spatial boundaries of the space–time diagram result in a number of possible methods of measurement as shown in Figure 18.1.

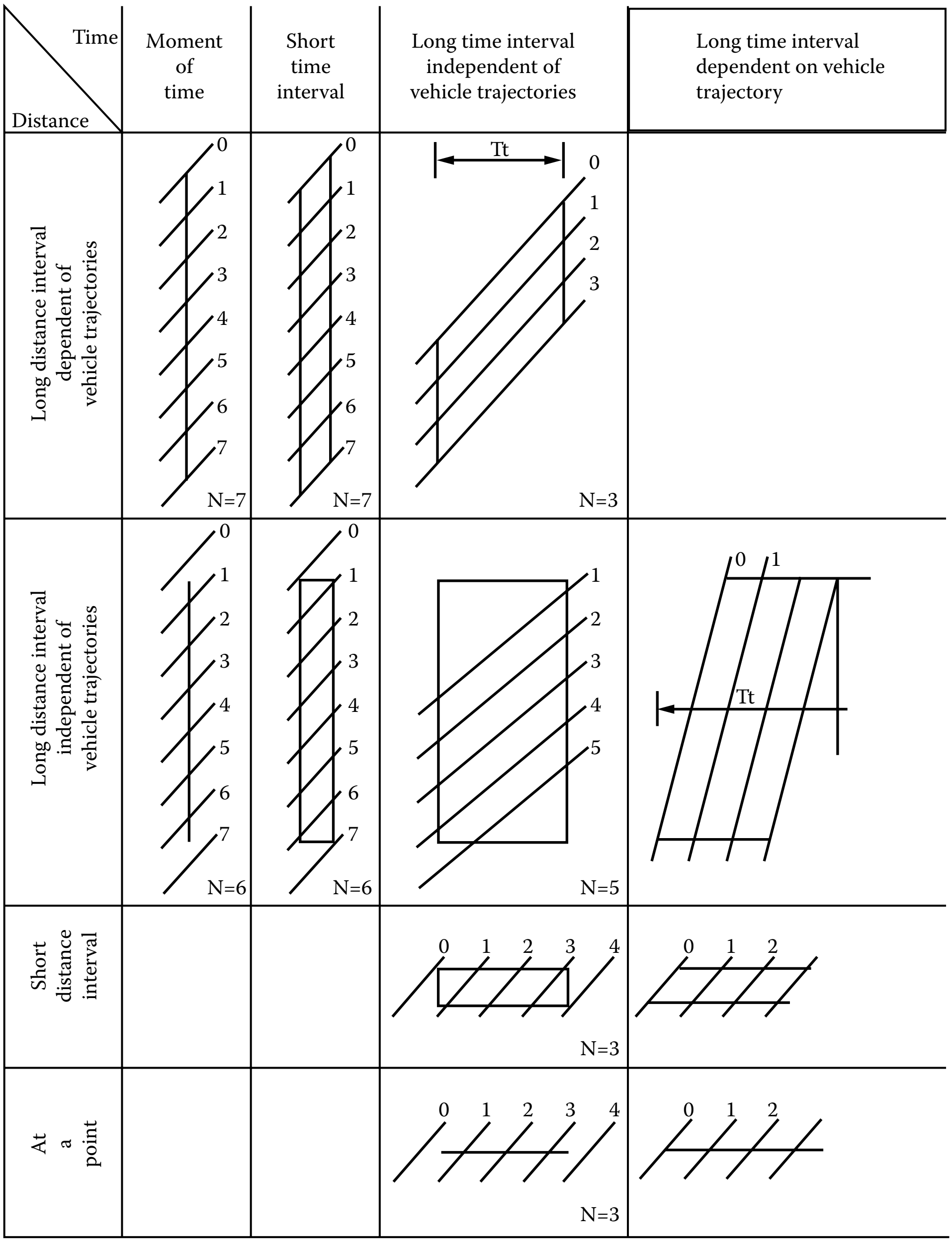


Figure 18.1 Measurement procedures. (From van As and Joubert, *Traffic flow theory*. University of Pretoria, 1993.)

18.2.3 Sample size

As discussed in Chapter 17, the size of the sample required to provide a meaningful basis for evaluation can be derived from

$$n \geq \frac{t_{\frac{\alpha}{2}\cdot n-1} \cdot s}{\in}$$

Table 18.1 Constant *K* for various levels of confidence

Constant (K)	*Confidence level* (E)
1.00	68.3
1.50	86.6
1.64	90.0
1.96	95.0
2.00	95.5
2.50	98.85
2.58	99.0
3.00	99.7

where

n = minimum sample size

$t_{\frac{\alpha}{2},n-1}$ = Student's t-statistic for $(1 - \alpha)$ *100% confidence level and $(n - 1)$ degrees of freedom

s = sample standard deviation

$\in$ = permitted error

Seeing that the sample size is determined by its own standard deviation, a trial and error approach must be followed in the determination of the required sample size (Nezamuddin et al., 2010). A useful starting point is, however, offered by the relationship (Box and Oppenlander, 1976).

$$n = \frac{p(1-p)K^2}{E^2}$$

where

n = minimum sample size

p = proportion of the sample representing a specific event

K = constant corresponding to the desired confidence level

E = permitted error in the proportion, p

The value of K for various confidence levels is given in Table 17.2 and repeated for convenience as Table 18.1. As a very rough rule of thumb, the minimum sample size is 30 elements.

18.2.4 Traffic volumes

Traffic volumes are necessary inputs into

- Considerations of capacity and level of service (LOS)
- The determination of road users costs for application in economic analyses
- Safety analyses where exposure rates have to be established
- Axle load repetitions over the design life of a road as a prime determinant of pavement design
- Determining traffic signal timing
- Executing environmental analyses such as of air quality and noise
- Planning studies

They are also useful for other than the purposes of transportation professionals such as

- Vehicle use as part of revenue forecasts
- Statistics used by the private sector for placement of businesses and services

In the field of geometric design, traffic volumes are required to be known in order to establish, amongst others

- The number of lanes that the road cross-section should comprise
- Indirectly, the desirable width of the individual lanes
- The need for auxiliary through and turning lanes
- The required form of traffic control at intersections and pedestrian crossings

Traffic counts are acquired either from outside the traffic stream or from within it. Measurement from outside the traffic stream can be, and still is in specific cases, acquired by roadside observers. In general, however, use was previously made of mechanical counters, such as pneumatic tubes across the road. These have now largely been replaced by electronic counters.

Electronic counters include piezoelectric sensors and induction loops installed across the road (Klein et al., 2006). In the case of temporary counting stations, loops are simply stuck to the road surface whereas permanent counting stations use induction loops embedded in grooves cut into the road surface. Pneumatic tubes are limited to recording the passage of a vehicle and the time of its arrival at the counter. Electronic counters can provide more detail by providing a stratified traffic count by type of vehicle – passenger cars, buses and various sizes of trucks. As a vehicle goes over an induction loop it distorts the magnetic field created by the loop. This distortion is referred to as the electronic signature of the vehicle and is a function of its size, shape and ground clearance. Some manufacturers claim that their counters can use this electronic signature to identify the make and model of individual passenger cars. Information to this level of detail is seldom necessary for transportation professionals.

In the interest of worker safety and ease of installation, particularly on very busy roads, nonintrusive technologies have been developed. These technologies are generally based on some sort of transmitted energy such as radar waves or infrared beams to detect vehicles.

The ideal situation is to be in possession of each of the flows occurring during the 8760 hours of the year and between all the nodes comprising the road from end to end. The cost of installing permanent stations between all the intersections along a road would be prohibitive and, for the same reason, they definitely would not be installed on every road of the network. The selection of the locations of its traffic counters by a transportation authority requires careful planning on the basis of the variety and nature of traffic flows on its road network. This is for application in the use of mother–daughter relationships as briefly discussed further in Section 18.4.2.

The average of measurements addressing the period 1 January to 31 December is referred to as the annual average daily traffic (AADT) but could actually comprise any period of 12 consecutive months. The averages of measurements covering shorter periods are referred to as average daily traffic (ADT). For these to be meaningful inputs in the design process, it must be noted that traffic volumes fluctuate minute by minute and by the hour of day, day of the week and month of the year. On a weekly basis, the highest flows occur on Fridays and the next highest on Mondays. Weekend traffic can be low except in close proximity to holiday destinations. It follows that, to be meaningful, the minimum duration of a traffic count should be 1 week with a desirable minimum of 2 consecutive weeks. The selection of the period to be counted should exclude public holidays and so-called special periods such as Easter, Christmas and, in the United States, Thanksgiving, when traffic counts would be highly abnormal. Similarly, the start and end of school holidays should also be excluded. Finally, the weather also plays a part. Traffic counting should not be undertaken during periods of inclement weather.

18.2.4.1 Urban network volumes

Although traffic volumes in the design year (typically 20 years hence) on rural roads can usually be estimated with sufficient accuracy by the application of a simple growth factor to present-day counts, the analysis of traffic volumes on urban networks is significantly more complicated. Present-day volumes on the various links of the network are derived from elaborate origin–destination (O–D) surveys as discussed in Chapter 20. These are projected to the future design volumes by regional and town planning exercises that take changes in land use into account.

Traffic counts, as described previously in the case of rural roads, are replaced by cordon counts. Boundaries are drawn around the central business district (CBD) (or any other area of interest such as an outlying industrial area or shopping precinct) on the basis of estimates of travel time from an origin to the CBD. Cordons are typically located at 5- or 10-minute increments of travel time working outwards from the CBD and it is not usual to have more than three cordons simply because the cost of data acquisition would become excessive. Very often, one cordon will suffice.

Cordon counts supply information on trends in movement to and from the CBD and, by totalling the inbound and outbound traffic, it is possible to derive an estimate of the accumulation of vehicles within the demarcated area and hence the extent of the need for parking. These studies also provide information on peak traffic flows during the day and the streets that are most heavily used. Unbalanced directional flows are easily detected and this information could be applied to the generation of contraflow strategies. Counting stations are normally provided at midblock locations to avoid the complications of turning vehicles.

Screen line counts record vehicle movement across major natural or manmade barriers. These have the practical effect of limiting the number of roads and hence counting stations required. Screen line counts are an important part of the accuracy checks on O–D surveys. These surveys are based on data gathered from interviews and provide an indication of the total trips having their origins and destinations on opposite sides of the screen line. These totals are compared with the actual counts as a check of their accuracy and the extent of modification (the fudge factor) required to improve the accuracy of the O–D survey.

18.2.5 Crashes and conflicts

What the driver wishes to do is based largely on what is read from the environment. The width of the road and its horizontal and vertical alignment, the intersections or interchanges along the way and the surrounding landscape and land uses all convey messages to the driver. The interpretation of these messages results in decisions regarding the path to be followed and the speed of travel. It is often found that there are more crashes at certain points along the road than at others, suggesting that a number of drivers are making the same mistakes at the same places. The inference to be drawn from this is that the environment is, in some way or another, being misread. The logical conclusion is that the much maligned 'driver error' could, at least in part, be the consequence of a design fault possibly causing an optical illusion.

The application of traffic flow theory could conceivably pinpoint the cause of the error, suggest a solution to the problem and then, through before-and-after studies, determine whether the implemented remedial action has been effective. It has to be noted that before-and-after studies are very often misinterpreted, sometimes deliberately. A crash rate high enough to be a cause for concern may trigger off the implementation of some or other remedial measure. Traffic counts taken before and after the implementation of the remedial measure may show a dramatic improvement at the site.

Unfortunately, it has also been observed that before-and-after studies *without intervening interventions* can also show improvements in safety! Mathematically, above average values are always accompanied by below average values and the black spot signifies a highly above average number of events. The before-and-after study can thus be a victim of the phenomenon known as 'regression to the mean' and it is necessary to establish whether the change in the number of events is statistically significant or not. In short, the question to answer is whether the after study represents a new situation or is a set of the same population as that of the before study.

Data on crashes are provided by police records. Unfortunately, recording crashes is low on the list of priorities of a normal police force focussed on fighting crime. In the absence of a dedicated traffic police force, these data are thus usually scanty and, where they exist, tend towards the unreliable.

The most reliable and useful source of information derives from the application of the fairly recently developed discipline of vehicular accident reconstruction. This involves the forensic investigation, analysis in depth and the drawing of conclusions about the causes of and the events during a crash. It typically finds application in the analysis of fatal and personal injury crashes for submission to a court of law and is concerned with the role of the driver, the vehicle, the road itself and the environment. Results from accident reconstructions are also useful in developing recommendations for making roads and the vehicles that travel on them safer.

18.2.5.1 Conflict studies

Even in Third World countries, where crashes are at astronomical levels compared to those in First World countries, crashes are relatively infrequent and it is thus difficult to acquire sufficient data to be able to pinpoint the cause of the crashes with any degree of certainty. Recourse can then be had to conflict measurement.

Conflict measurement does not apply to run-off-the-road (ROR) crashes. It addresses the special case of the situation involving two vehicles where one or both drivers have to take (usually fairly violent) measures to avoid a crash. These evasive measures include braking, as evidenced by the brake lights, and/or a change of direction, as indicated by changing of lanes. A conflict is also defined by an infraction of traffic laws even though no other vehicles may be present (Box and Oppenlander, 1976). Examples include the running of a red light, driving in the wrong direction along a one-way road and taking a shortcut to execute a left turn at a roundabout.

The location of sites where conflict studies may be appropriate is often identified by study of crash data, complaints from the public or excessive travel time and delay suggesting congested circumstances. Excessive conflicts are often found at busy intersections, in weaving sections or in merging areas at the on-ramps of interchanges. Studies should be undertaken at times of peak traffic flows and should address both the morning and the evening peak hours on account of the difference in flow patterns. Furthermore, studies are generally conducted when the weather is good and traffic conditions normal.

Conflict measurement is by observers either on-site or through the medium of video recordings and can be grouped under two major headings: those occurring at intersections and rear-end conflicts. There are 10 intersection conflicts and 12 rear-end conflicts. These are listed in Table 18.2. It is to be noted that many of the rear-end conflicts also occur at intersections.

Table 18.2 Summary of possible conflicts

Conflicts at intersections	*Rear-end conflicts*
Wrong lane	Stop on yellow
Left turn from	Slow for left turn
Right turn from	Slow for right turn
Left or right turn into	Previous traffic conflict
Cross-traffic	Driveway beyond intersection
Through left to right	Slow truck
Through right to left	Congestion in intersection
Right turning	Clear intersection
Left turning from left	Stalled vehicle
Left turning from right	Traffic back-up
Other	Merging beyond intersection
Weave	**Pedestrians**
Opposing left turn	Crossing
	Single vehicle pedestrian conflict
	Weave pedestrian conflict

18.2.6 Speeds

While flows are measured at a point, speeds are measured over short distances and hence require measurement of both time and space (Gartner et al., 2001). Prior to the development of electronic devices, measurement was by two observers using stopwatches. The stopwatches would be started by each observer as a vehicle passed them and then stopped simultaneously. The difference between the two times would then be recorded and, with the distance between the two stations being known, the calculation of speed was straightforward. The distance between the observers could be anything up to 100 metres or perhaps even more. In consequence, measurement was prone to be inaccurate because the distance involved made it possible for significant fluctuations in speed between the two observers. In the extreme case, a speeding driver could, on detecting the first observer, come to a dead stop before reaching the second one.

Replacing the stopwatches with two successive induction loops makes it possible to reduce the distance between the two observations to as little as 2.5 metres (Klein et al., 2006). At an initial speed of 100 km/h a driver, realising that he is possibly entering a speed trap, may decelerate under panic stop conditions of 6 m/s^2 and achieve a departure speed of 98 km/h assuming that he had his foot hard on the brake as he crossed over the first of the two loops. In practice, even quick reflexes would still have the driver beyond the second loop before braking commences.

Measurement by radar reduces the distance still further, almost but only almost to the point at which what is measured is the instantaneous speed of the vehicle.

There are two ways of presenting speeds (Gartner et al., 2001). The first is the arithmetic mean of all the observations, referred to as time mean speed because it is an average taken over time. It is presented as

$$\bar{u}_i = \frac{1}{N}\sum_{i=1}^{N} u_i$$

where

$\bar{u}_i$ = average speed
u_i = the ith observed speed
N = the total number of observations

The second way is referred to as space mean speed and seems to have acquired a plethora of definitions. There are two main types of definition, with the more generally accepted one being that it is the speed based on the average time to cross a given distance. The weakness of this definition is that it is getting back to the problem of measuring speed by the use of two stopwatches at a distance from each other. Implicit in this definition is that it is over all the vehicles that cover the full distance, whereas, with a fixed duration of the measurement period, there always will be, except under very light flow conditions, vehicles that are, in effect, trapped within the length of the section being used as the basis for measurement.

The second type of definition involves taking the average of the speeds of all the vehicles on the section of road at one point in time. On the space–time diagram this would be illustrated by a vertical line. Measurement could be accomplished by having two aerial photographs taken in quick succession to measure the speeds of the vehicles that appear in both photographs. A distribution measured in this manner will be identical to the true distribution of speeds.

If the volume of traffic in vehicles per hour and the speed in kilometres per hour is known, the units suggest that the quotient of volume and speed gives the density of the traffic stream from

$$\text{Density in vehicles per kilometre} = \frac{\text{Vehicles}}{\text{Hour}} / \frac{\text{Kilometres}}{\text{Hour}} = \frac{\text{Vehicles}}{\text{Kilometre}}$$

18.2.7 Travel time and delays

Delay is an indication of the LOS provided on two-lane roads and at intersections.

18.2.7.1 Delay on two-lane roads

On two-lane roads, if it is accepted that delay is the difference in travel time between travelling at some or other desired speed and the observed speed, it follows that every driver on the road would be operating at a different preferred speed when volumes are sufficiently low to allow the exercise of this preference. Furthermore, even at these low traffic volumes, the geometry of the road could have an impact on the selection of travel speed. Steep gradients and short radius curves would both have the effect of reducing travel speeds. Delay could thus be determined by acquiring sufficient speed measurements from end to end of the road under conditions of average and low traffic volumes, that is, at the lower levels of flow in LOS A, to be able to construct speed profiles along the length of the road for the two flow conditions and then to compare them.

It would be significantly more convenient, though, to obtain these measurements by the use of the floating vehicle method of measurement in which the vehicle moves with the stream and the driver attempts to pass as many vehicles as pass the test vehicle (van As and Joubert, 1993). Two groups of runs would thus be required: those at very low flows and those at average flow levels. It has been suggested that between 12 and 16 runs in each direction would usually be sufficient to provide reasonable estimates of flow and speed.

As an alternative, the Wardrop method of the moving vehicle could be used. The floating vehicle can only observe speeds but the moving vehicle method can also be used to measure flow and density. It requires the following observations:

- The number of vehicles passed by the moving vehicle in direction i, P_i
- The number of vehicles passing the test vehicle in direction i, S_i
- The number of vehicles, N_i, met in the opposing traffic stream when the test vehicle is returning in direction j, N_i
- The travel times, T_i and T_j, for both direction i and j
- The length of the road, L

From these inputs, if Q_i is the rate at which vehicles enter the road, then

$$N_i = Q_i(T_i + T_j) + P_i - S_i$$

from which

$$Q_i = \frac{N_i + S_i - P_i}{T_i + T_j}$$

The macroscopic speed of the traffic stream is the average speed achieved by the test vehicle because as many vehicles that passed by the test vehicle were passed by it and is expressed as

$$U_i = \frac{L}{T_i}$$

The density of the flow is derived as the quotient of the flow in vehicles per hour and the average speed of the flow in kilometres per hour and is thus expressed as vehicles per kilometre:

$$K = \frac{Q_i}{U_i}$$

Delay is the difference, T_D, between the travel times at the actual macroscopic speed, U_a, and the desired macroscopic speed, U_d. For a single vehicle this is given as (Wolhuter, 1990)

$$T_D = L \quad \frac{1}{U_a} - \frac{1}{U_d}$$

Hence, for an hourly flow in a single direction, delay is given by

$$T_D = L \quad \frac{1}{U_a} - \frac{1}{U_d} \quad QD(1 - P_t - P_s)$$

where

- L = length of road across which delay is being determined
- U_a = actual macroscopic speed
- U_d = desired macroscopic speed
- Q = average two-way flow (vehicles per hour)
- D = directional split as a decimal fraction
- P_t = rigid chassis trucks in the traffic stream as a decimal fraction
- P_s = semitrailers in the traffic stream as a decimal fraction

This relationship assumes that the flow rate stays constant for the entire hour, whereas the variation in flow could be modelled by the Poisson distribution. Delays calculated on the bases of uniform and Poisson flows over a wide spectrum of conditions of gradient, flow, directional split and traffic composition showed (Wolhuter, 1990) that a ratio, R_d, between the two types of flow existed as

$$R_d = e^{0.046+\frac{50.51}{Q}} \quad \text{for} \quad Q > 36 \text{ vehicles per hour}$$

with an R^2 value of 0.99, so that the total hourly delay could be expressed by the relationship shown above with the addition of the ratio, R_d,

$$T_D = \frac{1}{U_a} - \frac{1}{V_d} \; QD(1 - P_t - P_s)e^{0.046+\frac{50.51}{Q}}$$

with the various terms as shown before.

In traffic flow theory, the desired speed is sometimes assumed to be the speed travelled at optimum density, that is, when flow is at a maximum. Maximum flow presumably occurs at a low level of service and sounds remarkably like the capacity of the road. Regarding this as a desired condition deserves to be questioned.

18.2.8 The *Highway Capacity Manual* approach

A wide range of situations can arise in the operation of two-lane roads. The *Highway Capacity Manual* (HCM) 2010 thus proposes three measures of effectiveness:

- Average travel speed (ATS). This is the macroscopic speed shown above and is indicative of mobility.
- Percentage time spent following (PTSF). This represents the freedom to manoeuvre, which, in the case of a two-lane road, is the ability to overtake at will. With increasing opposing flows, this freedom obviously diminishes.
- Percentage of free-flow speed (PFFS). This is essentially an urban measure of quality of flow as it reflects the ability of vehicles to travel at or near the posted speed limit.

PTSF is, in effect, a surrogate for delay because it suggests that a vehicle was travelling at the preferred speed until it caught up with a slower vehicle and is now following it at some lesser speed. PTSF could possibly, like ATS, be determined by use of the floating vehicle methodology. It is difficult, if not actually impossible, to measure from outside the traffic stream. For this reason, a further surrogate to what is already a surrogate is required and this is the percentage of vehicles travelling at headways that are less than 3 seconds. This is only an approximation because it is often observed that drivers are prepared to accept the speed of a platoon that is leading them by several seconds as an alternative to being caught up in the hurly-burly of driving in the congested surroundings of a platoon.

18.3 HUMAN FACTORS

18.3.1 Introduction

The driving task is a hierarchical process (Lunenfeld and Alexander, 1990) comprising three subtasks or levels – control, guidance and navigation – and the main outcome of the execution of these subtasks is the maintenance of a safe speed and proper path relative to the roadway, traffic elements and other vehicles. The first two levels are important in the modelling and design of a road. Navigation previously played, at most, a minor role in the driving patterns witnessed on the road and was normally, where the driver was intending to enter unfamiliar territory, a case of study of road maps prior to the commencement of the journey. This knowledge-based behaviour will become more important in traffic flow theory as intelligent transport systems (ITS) mature. ITS can bring about changes or diversions in proposed routes and either change speed limits or post advisory speeds in addition to providing information on conditions ahead. Currently little is known about the impact of these on traffic flow.

18.3.2 Reaction time

The basic human characteristics of concern in the execution of the driving task are listed and discussed in some detail in Chapter 4. Jointly they involve cognitive and neuromuscular time lags which, between them, add up to driver reaction time. These time lags also impact on control movement time and responses to the movement and location of other vehicles and pedestrians.

Reaction time thus has two components. In the first instance, the driver has to sort out a number of alternatives to decide on the appropriate response and, in the second, the response is implemented. The very long reaction times of the elderly or incapacitated driver would require very high standards of vertical alignment, for example, with a corresponding cost penalty. For this reason, the geometric design of a road is typically based on an 85th percentile reaction time.

The time lag for a simple perception-reaction time where little decision making is involved could be 1.5 seconds with the initiating of the selected response still to follow. On freeways there is more going on that needs to be evaluated and, at busy intersections, there is still more information to be processed. The driver is a notoriously slow single-channel processing device and it is thus prudent to allow the driver additional time in these complex situations to arrive at a decision regarding the action to be undertaken. The value of reaction time that has been internationally adopted for design purposes is 2.5 seconds for simple responses and 'decision sight distance' is based on reaction times listed in Table 18.3.

Statistically, empirically derived reaction times cannot be modelled by a normal distribution because a negative reaction time is patently an impossibility. The log-normal distribution addresses non-negative probabilities as discussed in Chapter 17 and the distribution of reaction time using this model is as illustrated in Figure 18.2.

Table 18.3 Reaction times for calculation of decision sight distance

Manoeuvre	*Action*	*Reaction time (seconds)*
A	Stop on rural road	3.0
B	Stop on urban road	9.1
C	Speed/direction change on rural road	10.2–11.2
D	Speed/direction change on suburban road	12.2–12 .9
E	Speed/direction change on urban road	14.0–14.5

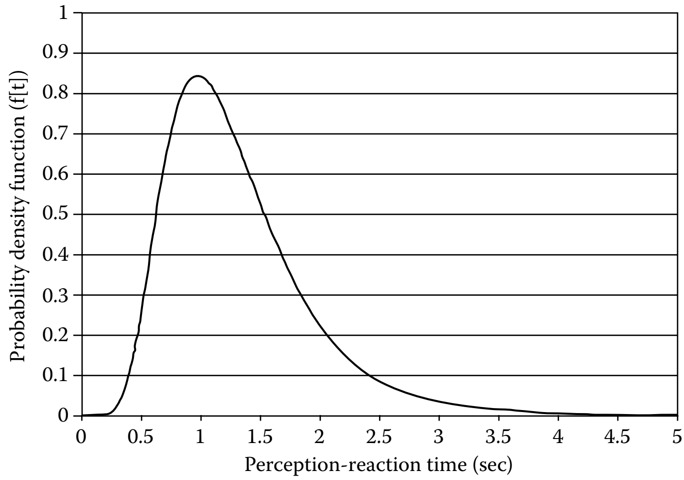


Figure 18.2 Log-normal distribution of reaction time.

18.3.3 Response to traffic control devices

A major input of information that influences the path and speed of the vehicle is that provided by traffic control devices (Gartner et al., 2001). The usefulness of these devices is related to the distances at which they can be

- Detected as objects in the visual field
- Recognized as traffic control devices in need of a response from the driver
- Legible or identifiable so that they can be comprehended and responded to

In the case of traffic signs, many items of information are imparted by the shape and colour of the sign. For example, the octagon of the Stop sign and the inverted triangle of the Yield sign do not require any supporting information. According to the *Manual of Uniform Traffic Control Devices* (2009), the combination of the colour of a legend and the background against which it appears provide guidance to a range of more than 30 types of sign. Finally, every word has a 'gestalt' or shape. This is created by the use of lowercase letters following the leading upper case letter. The series of rectangles that enclose the word illustrate this characteristic as shown in Figure 18.3.

The ability to recognise and understand the meaning of a traffic sign is dependent on the visual angle subtended by the sign at the driver's eye as indicated by the relationship

$$\text{Angle} = 2\arctan\frac{L}{2D}$$

Figure 18.3 The 'gestalt' of a destination name.

Table 18.4 Reaction time to signal change

Perception reaction time percentile	*Time (seconds)*
50	1.3
85	1.5
95	2.5
99	2.8

where

L = diameter of the target (letter or symbol)

D = distance from eye to target

Allowance must be made for reading time and the time required to decide on the appropriate response. Reading time is a function of the amount of information to be read, that is, the number of words, the information order and even the type of text because some fonts are more easily read than others. A sign with a great deal of information may take as much as 10 seconds to comprehend and a vehicle travelling at 120 km/h will cover a distance of 330 metres in the time taken by the driver to read the sign. The size of the sign must be such that it can be read at this range and, furthermore, the sign must be located sufficiently far in advance of where the required action is to be performed to allow for the driver's further reaction time.

An important aspect of response to traffic control devices is the response to changes of information provided by the control device. These include the changes between the various phases of a traffic signal and also drivers' reactions to variable message signs as a feature of ITS.

The reaction triggered by a phase change at a signalised intersection varies among drivers. A phase change from green to yellow may cause one driver to decelerate to a stop whereas another and more impatient driver may accelerate in an attempt to get through the intersection before the onset of the red phase. If the impatient driver is following the more cautious driver, the outcome could be unfortunate.

Drivers' reactions to signal change have been found to be as shown in Table 18.4.

18.3.4 Gap acceptance

Intersection sight distance used to be calculated by the use of a model that drew a distinction between crossing movements from the minor road across the major road and turning movements from the minor road to the left and to the right. Numerous assumptions had to be made including

- The speed of the entering vehicle
- The speed of the through vehicles
- For a crossing vehicle, the length of the vehicle, which affects the distance the vehicle has to travel to clear the intersection
- The speed of the turning movement, which is a function, in part, of the radius of the turn where a right turn will have a smaller radius than the left turn and hence possibly be executed at a lower speed
- The rate of acceleration applied by the driver of the turning vehicle having completed the turn

- The deceleration rate applied by a driver on the through road to reduce speed from the operating speed to whatever speed the left- or right-turning vehicle has managed to achieve starting from rest
- The length of the headway achieved by the following driver by the application of the chosen rate of deceleration

With all of these variables it is a matter of some surprise that researchers were able to come up with values of intersection sight distance (ISD) that were usable in practice. This rather clumsy approach to ISD has, however, now been generally replaced by gap acceptance as the basis of for its calculation. Quite simply, a driver arrives at an intersection, evaluates the gaps available for whatever movement is envisaged in relation to the performance of his or her vehicle and makes a selection of the gap to be accepted.

In practice, values of gap acceptance that can be applied to intersection design are difficult to establish. In general, every driver has some or other minimum gap, below which all gaps are likely to be rejected. Observing the minimum gap (generally referred to as the critical gap) that drivers are prepared to accept basically measures gaps accepted on the basis of gaps that are on offer. Acceptance varies from driver to driver and also depend on the driver's circumstances, for example, a leisurely drive as opposed to a commute to work. It is also observed that a driver will reject several gaps and then, no doubt because of increasing impatience, will accept a gap shorter than those previously rejected.

A distinction must be drawn between the first gap offered when a driver approaches the intersection and subsequent gaps. A gap is measured in time units between the back of the lead vehicle and the front of the following vehicle. If the first vehicle has already passed through the intersection, the driver only has the unexpired portion of the gap between it and the following vehicle available to cross or enter the major flow. This unexpired portion of a gap is referred to as a lag. Where an intersection is subject to Yield control, the vehicle on the minor road could still be moving while considering the gaps being presented and the driver may accept a lag that is smaller than the gap that would otherwise be accepted. This is because the need to start moving from scratch, which does take a little time, is replaced by the vehicle already being in motion even if only by a little over crawl speed. Lag acceptance thus needs to be considered separately from gap acceptance.

In addition to gap acceptance applying to drivers on a minor leg of an intersection who have to decide whether it is safe for them to cross or turn into the major flow, it also applies to drivers who need to judge the gap in opposing traffic on a two-lane two-way road to overtake a slower vehicle. Merging manoeuvres at an interchange on-ramp require the selection of a gap in the through flow. Pedestrians waiting to cross a street will need to assess a gap in the through traffic to establish when it is safe for them to cross. Drivers must, in the absence of a protected left turn offered by signal control, assess the size of gaps in the opposing traffic stream in order to execute a left turn at an intersection. Examples of the two latter cases of gap acceptance are dealt with in Chapter 17.

18.4 TRAFFIC CHARACTERISTICS

18.4.1 Introduction

The three traffic characteristics that define what is happening on the road are the flow, the speed of movement along the road and the density of the traffic stream.

Flow is measured in vehicles per hour or vehicles per second depending on what is being analysed. As one of the triumvirate of important traffic characteristics, it defines the width of the cross-section of the road in terms of the number of lanes that have to be provided. It is also important in the design of intersections as it is the prime input into the selection of type of control brought to bear on traffic movements in the intersection area. Turning flows define the need for auxiliary lanes to ensure that these manoeuvres do not impede the smooth flow of the through traffic, which is normally the major flow at any intersection. Heavy through flows may also dictate the need for auxiliary through lanes which can be dropped downstream of the intersection after completion of the merging movement.

Speed, which is expressed in kilometres per hour or metres per second, plays a major role in the selection of the geometric standards applied to the design of the road. These standards relate to the minimum values of horizontal and vertical curvature which are defined in terms of a selected design speed. Speed also has a bearing on the cross-section of the road insofar higher selected design speeds imply the need for a greater width of lane. Speed is also a measure of the quality of the road as perceived by road users in terms of travel time, economy and safety.

Density is measured in vehicles per kilometre and is a quantitative indicator of the quality of flow being experienced in the traffic stream congestion. It defines the freedom to manoeuvre of individual vehicles and the level of congestion they are experiencing. Density finds application in the identification of sectors of the road experiencing traffic problems.

These characteristics are discussed further below and are interrelated as illustrated by the Greenshields model, which is also discussed.

18.4.2 Flow

Traffic flow is defined as the number of vehicles passing a given point on the road in a given time and is expressed therefore as x vehicles in y days, hours or minutes. The flow **rate** is expressed in vehicles per second or, more usually, in vehicles per hour and may be for any period of time. Traffic flow is one of the more important measures of conditions in a traffic stream and is used in geometric design as a major determinant of the required cross-section of the road and even of the need or otherwise of providing a road with a permanent as opposed to gravel surface. It is also used to rank construction or upgrading of roads in order of priority.

The design of a road is predicated on the annual average daily traffic (AADT). With a sample duration of, say, only a fortnight, it cannot be certain that the average flow during the counting period in fact provides an indication of what the AADT for the road in question should be. This problem is resolved by comparing the data for the project of interest with that of a permanent counting station that has characteristics similar to the road to be designed. This is sometimes referred to as a 'Mother–daughter relationship'. In its simplest form, reference is to the traffic count at the permanent station (mother) during the same calendar period as that of the data acquisition count (daughter). The relationship

$$\mathrm{AADT}_{\mathrm{Daughter}} = \frac{\mathrm{Count}_{\mathrm{Daughter}}}{\mathrm{Count}_{\mathrm{Mother}}} * \mathrm{AADT}_{\mathrm{Mother}}$$

provides an indication of the current AADT of the road to be designed.

For design purposes, it is customary to design for an LOS that is likely to prevail at the end of the design life of the road, typically 20 years hence. This can be estimated in the case of a rural road by assuming a constant growth rate in traffic volume. If there is similarity of characteristics between the mother and daughter roads, it is quite likely that the historical growth rate at the mother would apply also to the daughter. This growth rate could then be applied to derive the traffic flow in the design year, typically 20 years hence.

18.4.3 The relationship between flow and hour of year

Design is predicated on the flow in the design hour, which is an hour occurring in the design year. As stated previously, the design year is typically taken as 20 years in the future. In 1941, Peabody and Norman (Wolhuter, 1990) claimed that the hourly flows ranked from highest to lowest showed a 'knee' in the vicinity of the 30th highest hour of the year. In the 29 preceding hours, flows decreased rapidly with increase in ranking from highest to the 30th highest hour. Thereafter the further decrease occurs at a much lower tempo. They reasoned that adopting the 30th highest hour as the design hour would result in a significant saving, as the additional investment to relieve congestion at the 29 higher hours may not only be prohibitive but would also lead to underutilisation of the facility in the remaining hours of the year. And this has been the basis of selection ever since.

An alternative technique is the flow regime developed by Dawson, who divided the hours of the year into four groups with what could loosely be described as 'constant' flow characteristics. Neither of these can be described as other than being approximate and there is a need for a more reliable determination of hourly volumes, both for the purposes of geometric design and economic evaluation of projects.

In the case of rural roads, if the hourly counts are ranked from highest to lowest and plotted against rank number on a log-log scale, the resulting plot is a virtually straight line between the highest and the 1030th highest hour (Jordaan, 1985). This is represented by the relationship

$$Q_N = 0.072\ \text{ADT} \left(\frac{N}{1030}\right)^{\beta}$$

where

Q_N = two-directional flow in Nth hour of the year (vehicles per hour)
ADT = average daily traffic (vehicles per day)
N = hour of year
β = peaking factor

This relationship was based on analysis of data from 65 permanent counting stations and the the lines emanating from all of the 65 stations pass very closely through the 1030th hour. As Jordaan stated in his conclusion: 'There is, therefore, only one parameter, i.e. β, that determines the peaking characteristics of a given road. In other words, only β is needed additional to the ADT to determine any hourly volume between the highest and the 1030th.'

According to Jordaan, the relationship between K, the ratio between the 30th highest hour and the ADT, and the peaking factor, β, is

$$\beta = -0.283 \ln K - 0.744$$

Values of β for a range of values of K from 0.10 to 0.30 are shown in Table 18.5. A value of β of −0.01 suggests a virtual lack of daily peaking and −0.40 indicates high peaking, with −0.20 being a very typical value.

Table 18.5 Relationship between *K* and β

K	β
0.10	−0.092
0.15	−0.207
0.20	−0.289
0.25	−0.352
0.30	−0.403

In urban areas, peak flows are typically the result of daily commuting, with similar numbers of vehicles on the road on every day of the working week. The 100th highest flow is thus a perfectly adequate measure of the present day flow that has to be accommodated. On planned but not yet built urban roads, assessment is more difficult. The new road will attract traffic from other links in the network serving the same destinations and will also attract new traffic by virtue of people abandoning public transport in favour of the convenience of the new facility. Furthermore, the development of a new destination or enhancement of an existing destination could also serve as an attractor of new traffic. Expansion of the residential area serving as the generator of traffic would similarly impact on the number of vehicles requiring to be accommodated.

Assessment of traffic volumes could require the application of

- Land use planning studies
- Origin–destination surveys
- Cordon counts
- Vehicle occupancy rates
- Trip allocation algorithms

These transportation studies are typically large, expensive and time consuming and require the services of experts in the field of transportation planning.

In rural areas, changes in land use are typically slow. As such, increases in traffic volumes can be assessed to acceptable levels of accuracy by applying a simple annual growth rate to present traffic volumes to determine the design traffic volume. In the absence of any regional planning initiative such as a change in agricultural practices, that is, from pastoral to agrarian, or the opening of a new agricultural industry or a mine, growth rates are typically of the order of 3 to 5 per cent per annum.

In rural areas, the design flow is selected to operate at LOS B and in urban areas at LOS C or even LOS D. The belief is that designing for the highest LOS occurring in a year 20 years in the future would result in the road being permanently underutilised, with the cost implications that this implies. Even using the 30th highest instead of the highest hour, the road is still going to be significantly underutilised in the intervening years.

To summarise the steps required to derive the design flow on a rural road:

1. A traffic count with a minimum duration of 2 weeks is undertaken.
2. This is compared to traffic data at a permanent counting station (the mother) with similar operating conditions for the same period and a correction applied to scale the ADT of the road to be designed (the daughter) up or down as the case may be.
3. The growth factor to be applied to the traffic on the daughter road is assumed to be the same as that shown by the mother station.

As mentioned earlier, the derivation of the design flow on an urban road is a complex exercise in regional planning and transportation engineering and beyond the scope of this book.

18.4.4 Speed

Speed of travel is an indication of the consistency of design of the road as discussed in Chapter 4. In the absence of impeding vehicles, free flow speed is dictated by the horizontal curvature of the road. Apart from this constraint, free flow speeds are modelled by the normal distribution. It has been found that, even though speeds may vary considerably from place to place, the standard deviation is generally of the order of 0.17 of the mean speed.

Speed data are normally presented in an aggregated form ordered from lowest to highest, usually at 5 km/h intervals as shown in Table 18.6 (van As and Joubert, 1993). A cumulative distribution curve is obtained by plotting the cumulative values against the upper bound of each speed interval (As and Joubert, 1993) as shown in Figure 18.4.

Table 18.6 Aggregated speed data

Speed group (km/h)	*Middle value (km/h)*	*Number of observations*	*Percentage of observations*	*Cumulative percentage*
30–35	32.5	2	0.8	0.8
35–40	37.5	12	4.8	5.6
40–45	42.5	40	16	21.6
45–50	47.5	73	29.2	50.8
50–55	52.5	71	28.4	79.2
55–60	57.5	38	15.2	94.4
60–65	62.5	11	4.4	98.8
65–70	67.5	3	1.2	100

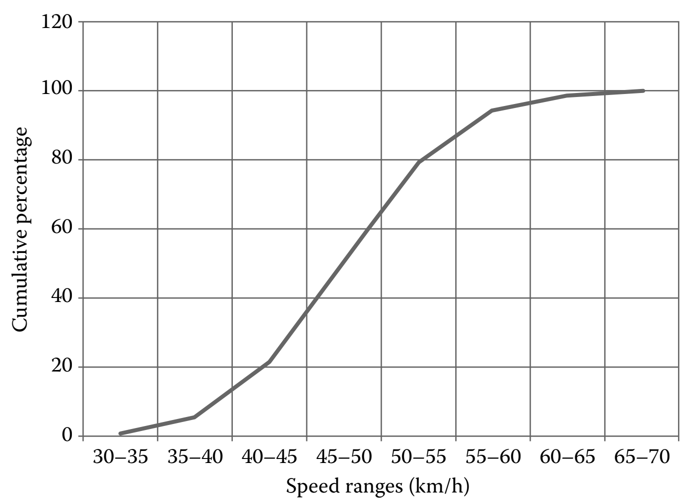


Figure 18.4 Cumulative speed data.

18.4.5 Density

From the fact that density is defined as number of vehicles per kilometre, it follows that density can be measured only along a length. Otherwise it needs to be calculated from the fundamental relationship

$$Q = U \cdot K$$

where

Q = flow
U = macroscopic (or space mean) speed
K = density

Concentration has been used as a synonym for density and density is a spatial concentration. Occupancy is a temporal concentration and is measured over a very short distance, typically shorter than the length of the minimum vehicle length. The argument in favour of using occupancy instead of density is that density refers only to a number of vehicles and ignores the effect of vehicle length and traffic composition. For the same density, the spatial gaps between vehicles would be greater if all the vehicles were passenger cars than if they were semitrailers. Occupancy is affected by both vehicle length and traffic composition and is, therefore, a more useful indicator of the amount of road being used by vehicles than is density.

18.4.6 The Greenshield model

Mention has been made previously of the relationship between the basic characteristics of a traffic flow as modelled by Greenshield. As stated previously, these characteristics are flow, speed and density. With the measurement of flow and speed, the value of density is fixed. The relationship is illustrated in Figure 18.4. Three variables imply a three-dimensional surface. However, efforts made to illustrate the relationship in three dimensions have proved uniformly unsuccessful and it is more convenient to present the model as a two-dimensional orthographic projection (van As and Joubert, 1993). Terms used in the figure are

Q_m = the maximum flow rate in vehicles/hour
U_m = the speed at which flow rate is a maximum
K_m = the density when flow rate is a maximum
U_f = the free flow speed when flow tends towards zero
K_j = the jammed density when vehicles are stopped

The figure illustrates that, as density approaches zero, that is, when flows are very light, the macroscopic speed approaches the free flow speed, U_f. If the density is zero there can be no flow so the flow-density curve must pass through the origin. At the far end of the density axis, the jammed density, K_j, occurs when all the vehicles are stationary so that the flow is again zero. At some value, K_m, of the density between the two limits, the maximum flow, Q_m, occurs and the speed has a value U_m.

It must be noted that Greenshield used very few data points in the construction of the linear relationship between speed and density and it is rather surprising that this relationship was unchallenged for so long (Figure 18.5). According to Drake (Gartner et al., 2001), a better fit with the data is obtained by accepting the existence of two flow regimes, one between zero density and 30 vehicles per kilometre and the other between 30 vehicles per kilometre and jammed density at about 80 vehicles per kilometre. There is an instantaneous

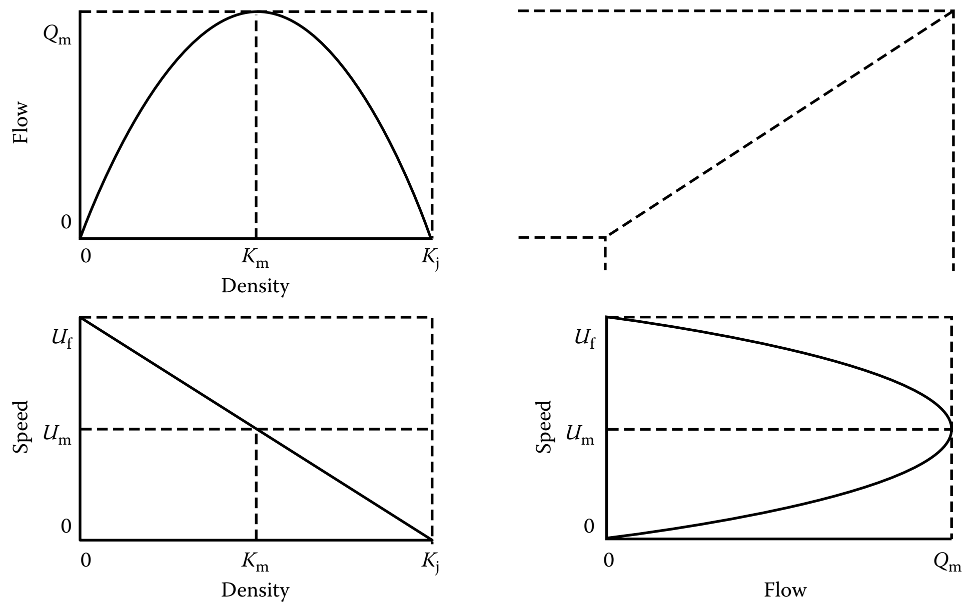


Figure 18.5 The Greenshield model.

drop of about 10 km/h from the low-density speed curve to the higher density curve at the transition point of 30 vehicles per kilometre.

18.5 TRAFFIC ARRIVALS

18.5.1 Random arrivals

Arrivals are defined as events, with random events being defined as being completely independent of any other event. Furthermore, equal intervals of time are equally likely to contain a certain number of events. If traffic is free-flowing, arrivals are indeed random but the hypothesis tends to break down with increasing congestion.

The assumption of traffic being free-flowing suggests that arrivals could be modelled by the Poisson distribution, which is often said to be memoryless insofar future events are completely independent of past events. This is usually demonstrated by flipping coins where each flip of the coin can come down either heads or tails regardless of what the previous flip(s) showed.

With arrivals being modelled by the Poisson process, headways are distributed according to the exponential distribution, expressed as:

$$P(X_i = x) = \frac{(\upsilon t)^x}{x!} e^{-\upsilon t} \quad \text{for } x = 0, 1, 2, \ldots$$

where υ is the mean occurrence rate, that is, the average number of occurrences of the event per unit time.

The special case of no arrivals simplifies the exponential distribution to

$$P(X_t \geq \mathrm{t}) = e^{-\upsilon \cdot t}$$

because the first term on the right-hand side of the above equation reduces to 1.

Non-arrivals are gaps between successive vehicles in the traffic stream. If the gaps are measured between similar points on each vehicle, reference is to the headway between them. Headway is normally measured in units of time. The safety rule of observing a gap of two seconds is between the back of the leading vehicle and the front of the following vehicles and is not to be confused with headway.

As discussed in Chapter 16, Section 16.7.2, the distribution finds application in determination of

- The required length at an intersection of a left-turn lane long enough to avoid traffic backing up into the through lane
- The sufficiency of gaps in a traffic stream long enough for traffic on the minor legs of the intersection either to cross or to turn into the major road without having to resort to signalized control
- The need or otherwise for traffic control measures at a pedestrian crossing
- The need for signalisation of an intersection

As is pointed out earlier, the headway includes the length of a vehicle. If headways are measured from the front of a vehicle to the front of the following vehicle, it is necessary to consider a whole range of combinations of vehicles. Because of the operating characteristics of their vehicles, drivers of heavy vehicles are inclined to leave a sizable gap between their vehicles and the leading vehicle, which may be a passenger vehicle. This is to allow for the longer time they need to decelerate should the leading vehicle brake. In this case, the headway comprises the time needed by the passenger car to travel its own length plus the time needed by the truck to travel the following distance selected by the driver of the truck. With 7 design vehicles, it follows that 49 combinations of 2 vehicles result and each combination could result in a unique headway. The problem is resolved by selecting the rearmost projection on each vehicle as the point to measure. This means that the headway for a passenger car would include the length of the car and the preference of following distance by the passenger car driver, the headway for a truck would include the length of the truck and the preference of following distance by the truck driver and so on. This reduces the number of combinations to 7 from the original 49.

A zero headway, although mathematically possible, cannot exist in practice. There is a minimum physical distance that drivers prefer to maintain between themselves and the leading vehicle and this can be taken into account by using the shifted exponential distribution. The probability function acquires a term, τ, equal to the desired length of the shift and is expressed as

$$f(t) = \frac{1}{T-\tau} \cdot e^{-(t-\tau)/(T-\tau)}$$

18.5.2 Nonrandom arrivals

Random arrivals apply only to low traffic volumes where there is a minimum of interaction between the various vehicles in the traffic stream. When traffic is becoming congested, the lack of passing opportunity causes the hypothesis of randomness to fail and successive headways are no longer independent. Any stream of traffic includes vehicles unconstrained by others and those captured in platoons. Reference is thus made to following or constrained vehicles and non-following or free vehicles. The numbers of each vary across the length of the road and are in a constant state of flux as vehicles catch up with platoons and then finally overtake the vehicle causing the platoon.

The question of when a vehicle is following is in need of resolution. If the driving behaviour of one driver is influenced by that of the driver of the leading vehicle, he or she is clearly a follower. If the leading vehicle is travelling very slowly and the following vehicle is at the speed limit or above, the following driver will commence deceleration at a considerable distance behind the lead vehicle. However, the HCM ignores this transition phase and bases its calculation of percentage time spent following as a measure of LOS on the number of vehicles travelling at a headway of 3 seconds or less. One weakness of this approach to following is the observation that some drivers will avoid the turbulence associated with driving in a platoon by allowing a substantial gap between themselves and the platoon ahead of them and then adopting the speed of that platoon to maintain this gap. They are, thus, in effect following even though falling outside the compass of the definition.

18.5.3 Speeds on curves

The speed at which drivers are prepared to drive at on curves is an important indicator of the quality of design of the road. As discussed in Chapter 4, variations in speeds are divided into three ranges, good, fair and poor, as an indication of whether it is necessary to reconsider the design of the road.

Repeating for convenience what is stated in Chapter 4, the curvature change rate (CCR) (Lamm et al., 2001) is calculated as

$$CCR_s = \frac{\frac{L_{Cl1}}{2R} + \frac{L_{Cr}}{R} + \frac{L_{Cl2}}{2R}}{L} \cdot \frac{200}{\pi} \cdot 10^3$$

where

CCR_s = curvature change rate of the single circular curve with transition curves (gons/km) (Note: The European 400 gons = 360 degrees. The CCR can be expressed in degrees/km by replacing the value 200 in the equation by 180)

L = overall length of unidirectional curve (m)

L_{Cr} = length of circular curve (m)

L_{Cl1}, L_{Cl2} = length of clothoids preceding and succeeding the circular curve (m)

As listed in Chapter 4, regression models were developed for the estimation of operating speeds in relation to radius of curvature in nine countries (Lamm et al., 1999, 2001; Operational Effects of Geometrics Committee, 2011):

- Australia
- Canada
- France
- Germany
- Greece
- Italy
- Lebanon
- United Kingdom
- United States

The highest operating speed is found in Italy and the lowest in Lebanon, with the others running between these two extremes and generally parallel to them as illustrated in Figure 18.6.

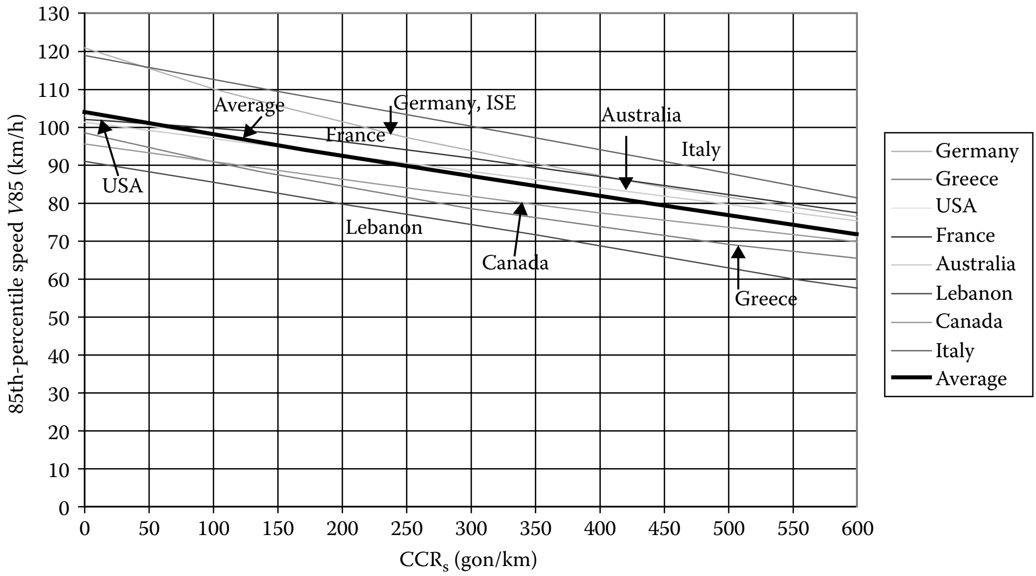


Figure 18.6 Operating speed backgrounds for two-lane rural roads.

As stated in Chapter 4, the average relationship in the case of a gradient equal to or less than 6 per cent between the curvature change rate and the operating speed, *V*85, is given by the regression

$$V85 = 105.31 + 2 * 10^{-5} * \mathrm{CCR}_s^2 - 0.071\mathrm{CCR}_s$$

with an R^2 value = 0.98.

For gradients steeper than 6 per cent, the regression becomes:

$$V85 = 86 - 3.24 * 10^{-9} * \mathrm{CCR}_s^3 + 1.61 * 10^{-5} * \mathrm{CCR}_s^2 - 0.0426 * \mathrm{CCR}_s$$

with R^2 = 0.88.

18.6 MICRO- AND MACROSCOPIC MEASURES

Traffic flow theory comprises two levels of interest. The macroscopic level considers the traffic stream as a whole. Reference is often made to the hydraulic analogy in which the road can be represented by a pipe containing flowing water. It has a certain capacity, that is, when flowing full, but is not always utilised to its maximum, for example, when flowing half or quarter full. This, loosely, is the equivalent of the various LOS scores employed in the *Highway Capacity Manual*. The concern with individual drivers and their vehicles is referred to as the microscopic approach as it addresses the behaviour of individual particles in the traffic stream. Microscopic analysis is aimed at modelling this behaviour in their passage through the individual components of the network. The macroscopic level normally applies to analyses at the network level.

Speed may be measured in either of these two ways depending on the application of the speed study. Microscopic speed refers to the speeds of individual vehicles and microscopic mean speed, also known as time mean speed, is defined as the average of the individual speeds of a sample of vehicles expressed as

$$\bar{v} = \frac{\sum_{i=1}^{n} v_i}{n}$$

where

$\bar{v}$ = microscopic mean speed
v_i = microscopic speed of the ith vehicle
n = the number of observed vehicles

Macroscopic speed, also known as the harmonic mean speed, is expressed as

$$U = \frac{L}{\sum_{i=1}^{n} t_i}$$

where

U = macroscopic mean speed
L = the distance over which the speed is measured
n = number of observed vehicles
t_i = travel time of the ith vehicle

It has been shown (van As and Joubert, 1993) that there is a relationship between microscopic and macroscopic speed given as

$$v = U + \frac{S^2}{U}$$

$$S^2 = \frac{\sum k_i \cdot (U_i - U)^2}{K}$$

where

v = time mean speed
U = space mean speed
U_i = space mean speed of the ith traffic stream
S = variance of the space mean speed
k_i = density of the substream i
K = density of the total stream

Chapter 19

Capacity analysis

Interrupted flow

CONTENTS

19.1 INTRODUCTION

Capacity is defined as the maximum sustainable flow of vehicles or people that can reasonably be expected to pass a point on the road under prevailing roadway, environmental, traffic, and control conditions within a given time period.

The time period involved may be measured in hours, days or weeks depending on the form of analysis that is intended. Although flow may fluctuate quite heavily within the compass of an hour, it can be fairly constant over a period of about 15 minutes. In the case of data acquisition for analysis of an operational issue, this is the normal counting interval for the acquisition of data within the compass of the overall traffic counts. The measurement of delay at a signalised intersection is typically based of counting intervals of 15 seconds. In selection of the counting interval, the analyst should thus have a clear idea of the data needs of the analysis that is intended to be carried out.

Roadway conditions are the three-dimensional combination of the elements of horizontal and vertical alignments with the cross-section. It is to be noted that where *Highway Capacity Manual* (HCM) *2010* (Transportation Research Board, 2010) uses the term 'grade' throughout, in this Handbook a distinction is drawn between the grade, which is the equivalent in the vertical alignment to the tangent of the horizontal alignment, and the gradient, which is the steepness of the grade. It can be expressed either as a number of metres or feet of climb or descent per unit length of the alignment or as a percentage.

The environmental conditions that impact on use of the road are essentially those of weather and light. Adverse weather conditions such as strong wind, rain or snow will all have an effect on reducing the speed of travel with a knock-on effect on flow rate. After dark or under conditions of heavy overcast, fog, or windblown dust, speeds will also be low.

Traffic conditions refer to the composition of the traffic stream in terms of the percentage of heavy vehicles in it and also the directional percentage split between the two directions of flow. Drivers being familiar or not with the road being travelled will also play a role in affecting the ease of operation and hence speed and density of flow. Large numbers of turning vehicles, particularly if turning to the left, can dramatically reduce the capacity of a road if auxiliary turning lanes have not been provided.

Control conditions refer to the presence of intersections, whether signalised, roundabouts or priority controlled.

In this chapter, discussion is focussed on some background to capacity analysis and also interrupted flow. Uninterrupted flow is discussed in Chapter 20.

19.2 THE *HIGHWAY CAPACITY MANUAL 2010*

These chapters serve only as a brief introduction to the 2010 edition of the HCM. Reducing four volumes to two chapters suggests that they are limited in scope to being only the broadest of overviews.

The first edition of the HCM was published in 1950. It was updated in 1965 and, like its predecessor, this was a modestly sized paperback document. The latter document was of historical importance because it introduced the concept of level of service. Since then there have been updates in 1985, 2000 and now 2010. These three editions have incorporated new research as it became available during the intervening years. An important feature of the 2010 edition is that, whereas the previous editions principally focussed on the driver and the vehicle, this edition is concerned with all road users with a greater emphasis being placed on pedestrians, cyclists and public transport, the last-mentioned being referred to in HCM 2010 as 'transit'.

Transit includes

- Buses, either diesel or petrol-driven
- Buses receiving their power from overhead power lines and sometimes referred to as 'trackless trams' or 'trolley buses'
- Light rail, including trams and rail buses

A distinction is drawn between pedestrians and cyclists operating in a shared mode with the other vehicles on the road as introduced in HCM 2010, Chapter 16, and operating in separate facilities, discussed in HCM 2010, Chapter 23.

From it modest beginnings, the HCM has grown to four large volumes:

- Volume 1: Concepts — Chapters 1 to 9
- Volume 2: Uninterrupted Flow — Chapters 10 to 15
- Volume 3: Interrupted Flow — Chapters 16 to 23
- Volume 4: Applications Guide — Chapters 24 to 35

It continues the discussion of the fundamental flow characteristics of flow, speed and density and is concerned with the various classes of road from the freeway to the local access street and their ability to accommodate moving vehicles. In addition, the needs of pedestrians, cyclists and transit users are dealt with in significantly greater detail than was previously the case. Of interest is the measurement and analysis of the observable fundamental characteristics and also the quality of the service provided by the road as perceived by the road user. This quality is quantified in terms of levels of service.

While this Handbook has been written using metric units, the HCM has yet to be metricated. In view of the fact that these chapters serve as a very brief introduction to HCM 2010, it has been decided that, to avoid confusion and where appropriate, imperial units will apply to this and the following chapter.

19.3 THE APPLICATION OF HIGHWAY CAPACITY ANALYSIS TO GEOMETRIC DESIGN

HCM 2010, like its predecessors, addresses many end uses covering the spectrum from planning to operations but the focus in these chapters is limited to its application to geometric design. Typically, this includes determining the width of the cross-section in terms of the number of basic lanes required and the widths of these lanes. Other applications include

- Determination of the impact of steep gradients on roadway performance; followed by
  - The determination of the need for climbing lanes
  - The determination of the need for passing lanes
- The selection of the form of intersection control
- Assessment of the need for turning lanes and auxiliary through lanes at intersections
- Consideration of the need for and nature of provision for pedestrians in terms of
  - The number of passenger queuing areas at pedestrian crossings
  - The form of control needed at pedestrian crossings
  - The required width of sidewalks
- The evaluation of the need for facilities for cyclists such as
  - Widening of the travelled way
  - Separate lanes additional to the normal travelled way
  - Paths removed from the carriageway

It must be noted that design is typically based on the assessment of conditions likely to occur in the design hour. The design hour occurs in the design year, which is usually some 20 years in the future. It follows that calculation to several decimal places implies a level of precision that cannot be matched in practice. The difference between the multipliers applied to the base year to determine flow in the design hour between successive selections of percentage annual growth can be dramatic. This is shown in Figure 19.1. Rural growth rates usually fall in the range of 3 to 5 per cent, suggesting that the multiplier applied to present day traffic counts could differ by a factor of from 1.8 to 2.7. This difference implies a significant difference in the traffic flow being designed for.

Clearly, the level of precision implied by calculation to five decimal places is nonsense, although software achieves this quite easily. Nevertheless, the designer should strive to make the determination of the present-day values of whatever variable is being measured as accurate as possible. It follows that, in the curve illustrated in Figure 19.1, the accuracy of the determination of the current growth rate and traffic count should also be as high as possible

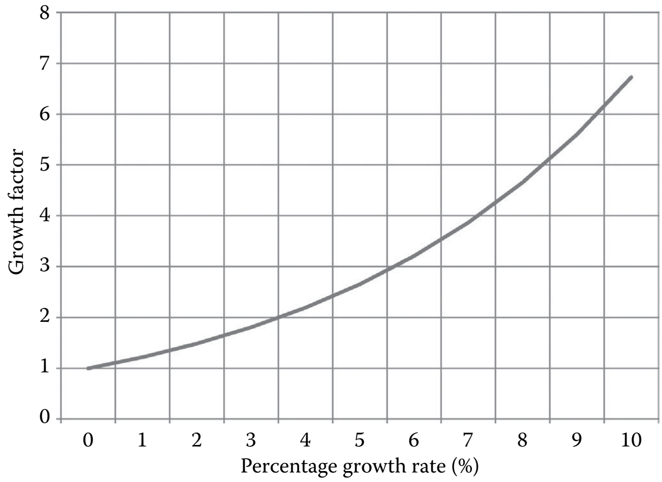


Figure 19.1 Growth factor for various assumption of percentage growth rate.

and an at least reasonable attempt be made to assess the likely growth rate over the design period. The duration of the design period is also in need of careful consideration. If the project is the addition of a climbing lane to a road that is already halfway through its design life, the design life for the climbing lane could perhaps be shorter than the customary 20 years. On the other hand, if the project is the construction of a new road through rugged terrain, that is, where construction and road user costs are likely to be high, the adoption of a longer design life may be desirable.

19.4 THE LEVEL OF SERVICE

Road users include the drivers of motorised vehicles as well as pedestrians and cyclists. There are many different classes of vehicles on the road including

- Passenger cars
- Trucks
- Buses
- Recreational vehicles

The dimensions and operational characteristics of the individual vehicles in each class cover a wide range and the design values of each, as discussed in Chapter 5, are usually their 85th percentile values. Passenger cars predominate by a large margin in the traffic stream and discussions of the perceptions of drivers are tacitly thus those of the drivers of passenger cars.

In the case of the drivers of vehicles, level of service (LOS) is a concept aimed at the perceptions of the driver of the quality of the service provided by the road. A road with numerous curves of short radius would, because of the low speed it forces, be perceived as having a poor quality of service. On the other hand, a road with long tangents and long radii of curvature would support the consistent maintenance of high speeds and thus be perceived as having a high quality of service. LOS is thus the quantification of a quality. It is very useful as it encapsulates a variety of variables into a simple picture describing the overall

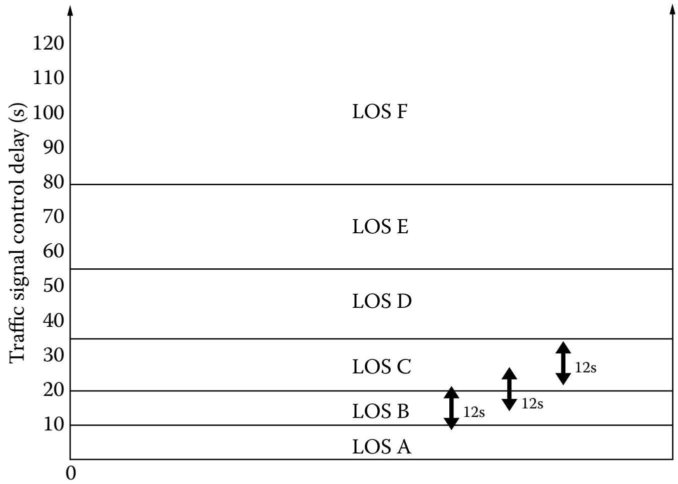


Figure 19.2 The impact of the stepwise progression of levels of service (LOS). (From Transportation Research Board. *Highway Capacity Manual 2010*. Washington, DC, 2010, Figure 5.1.)

impression the road creates in the mind of the driver. Furthermore, it simplifies the explanation of the quality of service to non-technical people who may be road users, the community at large or political decision makers.

Importantly, the definition of capacity stresses that it is the flow that can *reasonably* be expected to pass a point on the road during a period of time. An isolated incident of LOS F flow would thus not generate the implementation of any remedial measures, whereas if it becomes a regular daily occurrence over an extended distance or time, some or other intervention would, in all probability, be required.

LOS is a stepwise rather than a continuous progression, with each LOS addressing a range of values of the criterion of interest. As illustrated in Figure 19.2, a fairly modest range of variation in a value could thus signal either no change in the LOS or a change from one LOS to the next either up or down. In the extreme case, a change through a LOS to the second one up or down could be involved. The last-mentioned would, however, be possible only if the range of the variable being measured was large. It follows that the designer should not blindly accept that a change in LOS will automatically require some or other geometric intervention.

As a general principle, the derivation of the LOS provided by any facility is based on the calculation of the value of the criterion under ideal conditions followed by its modification by factors to accommodate conditions that are less than ideal. The value thus derived is then compared with the defined boundary values of each LOS to establish the LOS range in which the performance of the facility falls.

19.5 INTERRUPTED VERSUS UNINTERRUPTED FLOW

All flows on a road fall into one of two regimes: interrupted or uninterrupted flow. Interrupted flow involves traffic being slowed or stopped by external circumstances such as traffic signals or priority control. Uninterrupted flow arises when disruptions to smooth flow arise inside the flow itself, being caused by slow-moving vehicles and/or congestion. It is important to note that a platoon of traffic moving smoothly without stopping through a series of traffic signals in a 'green wave' operation is said to be operating in an interrupted flow regime whereas a stream of traffic on a freeway at jam density and thus stationery is describing uninterrupted flow. Interrupted-flow facilities have causes of periodic delay or traffic stream interruption such as traffic signals, roundabouts and Stop or Yield signs. The delay caused by traffic signals arises from the fact that flow through them is in platoons, with significant gaps between successive platoons but can be only for part of the hour during the green phase of the cycle. Stop control releases vehicles in small dribs and drabs dependent on gaps in the opposing flow.

The flow regime thus describes the facility and not the quality of flow.

19.6 MEASURES OF PERFORMANCE

19.6.1 Motorised vehicles

There are two basic forms of performance measures. In the first instance there are physical measures that can be determined through observation in the field and, in the second instance, measures aimed at providing an insight into the perceptions of road users of the quality of service being provided by the road, the LOS.

19.6.1.1 Physical measures

Three criteria of effectiveness are

- Average travel speed (ATS). This is the macroscopic speed described in Chapter 20 and is indicative of mobility.
- Percentage time spent following (PTSF). This represents the freedom to manoeuvre, which, in the case of a two-lane road, is the ability to overtake at will. With increasing opposing flows, this freedom obviously diminishes.
- Percentage of free-flow speed (PFFS). This is essentially an urban measure of quality of flow as it reflects the ability of vehicles to travel at or near the posted speed limit.

In the case of interrupted flow along an urban street, ATS is the primary measure of LOS. As such, it represents the level of mobility provided by the road. It reflects the factors that influence running time along each segment of the road and the delays incurred at each intersection with the rest of the network. A segment is defined as that portion of the road between the functional areas of the intersections on either side of it and includes the functional area of the upstream intersection. A signalised intersection always constitutes a segment boundary whereas the uncontrolled legs of a priority controlled intersection can be contained within the segment. Functional areas and their dimensions are discussed in Chapter 10.

PTSF is, in effect, a surrogate for delay because it suggests that a vehicle travelling at the preferred speed has caught up with a slower vehicle and is now following it at some lesser speed. PTSF could, like ATS, be determined by use of the floating vehicle methodology described in Chapter 17. It is inconvenient to measure from outside the traffic stream. For this reason, a further surrogate to what is already a surrogate is required and this is the percentage of vehicles travelling at headways that are less than 3 seconds.

The determination of the percentage of vehicles travelling at headways that are less than 3 seconds long is easily accomplished during the traffic counting process when what is recorded is the time of arrival of successive vehicles and not simply their passage over the counting device.

This acceptance of a 3-second headway as a measure of following is only an approximation because it is often observed that drivers are prepared to accept the speed of a platoon that is leading them by several seconds in preference to being caught up in the stressful situation of driving in an environment of closely spaced vehicles. Furthermore, delay actually commences at the time when a vehicle decelerates from the desired speed to match that of the platoon and terminates at the end of the acceleration back to the desired speed. The expectation that a desired speed will instantaneously drop from free-flow speed to the platoon speed when the headway is only 3 seconds is patently unrealistic. However, at a headway of only 3 seconds, the following vehicles will have little choice but to travel at the same speed as that of the leading vehicle. It may be an imperfect measure but it is better than no measure at all. Engineers are sometimes prepared to forgo the precision of the scientist in favour of the more practical approach of 'what is good enough'. The HCM frequently is a glowing example of the application of this philosophy.

19.6.1.2 The level of service

Each LOS comprises a range of values of the criterion being employed. The significance of the various levels is as follows:

- LOS A provides for unimpeded freedom to manoeuvre and the attainment of speeds higher than 85 per cent of free-flow speed. Delays at intersections are minimal.
- LOS B describes reasonably unrestricted but not totally unimpeded freedom to manoeuvre. Control delay at intersections is still not significant. Speeds are between 67 and 85 per cent of free-flow speed.
- LOS C operation is still stable but the ability to manoeuvre and change lanes is more restricted than at LOS B. Longer queues at intersections cause a reduction in speed to between 50 and 67 per cent of free-flow speed.
- LOS D represents the onset of unstable conditions in which small increases in flow may result in substantial increases in delay and reductions in travel speed to as low as 40 to 50 per cent of free-flow speed.
- LOS E is characterized by unstable operation and significant levels of delay with speeds of 30 and 40 per cent of free-flow speeds.
- LOS F flow is at extremely low speeds, typically anything less than 30 per cent of free-flow in addition to frequently operating in a stop-and-start mode. Congestion at intersections is more or less inevitable and delay thus considerable. Regardless of the travel speed through the road section, if the volume/capacity ratio is greater than 1.0 then LOS F is assigned.

For ease of reference, the ATS associated with the various LOS scores are shown in Table 19.1.

Table 19.1 Travel speeds as a percentage for the various LOS

ATS as a percentage of free-flow speed	*LOS*
>85	A
>67–85	B
>50–67	C
>40–50	D
>30–40	E
≤30	F

It is pointed out that LOS E conditions need not be a result of geometric constraints but could arise from operational conditions. Poor selection of phasing at signalised intersections and lack of signal progression between successive intersections could be at fault.

The capacity of a facility is defined as being at the boundary between LOS E and LOS F. LOS F is defined as being unstable flow and is subject to breakdowns in flow. It is subdivided further but not as LOS F-1, LOS F-2, and so on but rather by an added descriptor such as the distance over which the breakdown applies or the duration of the breakdown. The reason for this is that a breakdown limited to a point may not require much in terms of financial outlay to resolve whereas, if over a considerable length, it may require extended upgrading of the cross-section of the road. Similarly, a breakdown over a short duration may perhaps be simply accepted as a passing nuisance and not warrant any remedial measures.

Importantly, the definition of capacity stresses that it is the flow that can *reasonably* be expected to pass a point on the road during a period of time. An isolated incident of LOS F flow would thus not generate the implementation of any remedial measures, whereas, if it becomes a regular daily occurrence over an extended distance or time, some or other intervention would, in all probability, be required.

19.6.2 Intersection delay

19.6.2.1 Delay at signalised intersections

The geometric design application of the determination of delay at a signalised intersection is the evaluation of the need for auxiliary lanes that may be either for through or for turning traffic. Delay commences at the point at which a vehicle's speed is reduced below the operating speed on the approach to a traffic signal and terminates once through the signalised intersection and back to operating speed. The methodology described in the text that follows determines the delay caused by vehicles being stationary in a queue and replaces measurement of the deceleration and acceleration portions of the delay by the application of a correction factor.

Delay at intersections is usually measured by on-site observers, although area control systems automatically acquire information that can be used to calculate delay.

HCM 2010 defines delay at intersections as the increase in travel time due to traffic signal control and suggests that it provides a surrogate measure of driver discomfort and fuel consumption. The methodology proposed in HCM 2010, Chapter 18, for the evaluation of the LOS provided by a signalised intersection is, in its own words 'computationally intense', and requires the use of software. Analysis is described with respect to three modes of travel, by passenger car, bicycle or walking. In the case of analysis of LOS for passenger cars, the determination of delay is but one step in a series of ten in the derivation of the LOS.

A simpler form of measurement is provided in HCM 2010, Chapter 31. It requires the use of two observers. One observer counts the vehicles actually in queues and the other keeps track of three counts:

- The total number of vehicles arriving during the study period
- The number of these vehicles that stop one or more times
- The number of signal cycles during the study period

The average queued time per vehicle, that is, delay, is given by the relationship

$$d_{vq} = I_s \cdot \frac{\sum V_{iq}}{V_{tot}} * 0.9$$

where

d_{vq} = vehicles stopped in a queue
I_s = time between the start of successive counts, that is, 15 seconds
V_{iq} = the queue length (in vehicles) in time period i
V_{tot} = total number of vehicles arriving during the survey period

It follows that any vehicle that is stationary for longer than the 15-second counting period will be counted more than once. Seeing that total delay is the sum of the number of vehicles stationary during each 15-second period multiplied by 15 seconds this should not be a cause for surprise. The adjustment factor, 0.9, corrects the error made when delay is estimated by the use of queue counts. A further correction factor shown in Table 19.1 is offered to allow for deceleration and acceleration which cannot be measured by this manual application. This factor is multiplied by the fraction of vehicles stopping and the product added to d_{vq} calculated as shown to obtain the estimate of the control delay. The form of the calculation of the total delay caused by the traffic signal is also shown in Figure 19.3.

19.6.2.2 Delay at priority-controlled intersections and pedestrian crossings

The calculation of delay at priority–controlled intersections and pedestrian crossings is to determine whether signalisation is required and, furthermore, to establish the required length of the storage areas at turning roadways and lanes. To this end, the methodologies described in Chapter 17 are adequate. The HCM 2010 methodologies are thus, in the interests of brevity, not described.

19.6.3 The LOS for a facility

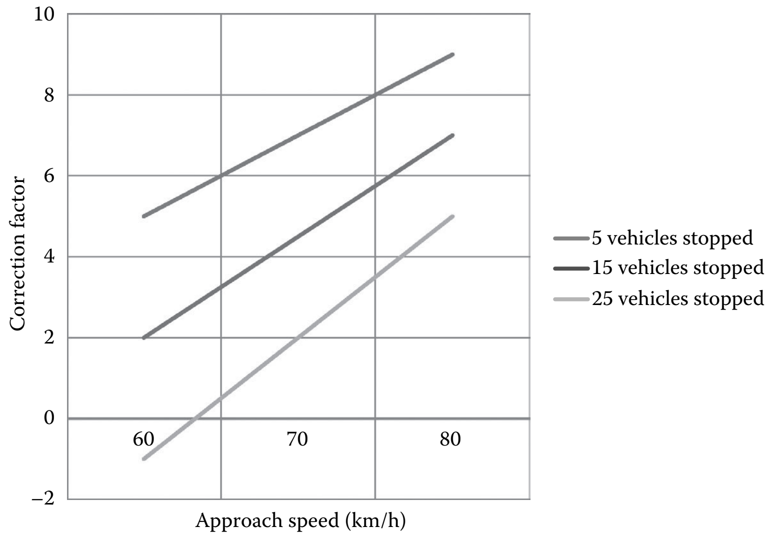
The LOS score provided by a facility is the weighted average of the scores for the individual segments. As previously stated, a segment comprises a link and its upstream intersection. It is thus necessary to calculate the score for each component of the segment. The LOS score for the segment is calculated by application of the methodology explained in HCM 2010, Chapter 30. This methodology comprises five steps:

- Running time
- Proportion of vehicles arriving during the green phase of the cycle
- Through control delay
- Through stop rate
- Travel speed and spatial stop rate

Worksheets are provided for each of the calculation steps and an example of the calculation is shown in HCM 2010, Chapter 30, Exhibits 30-8 to 30-12.

The LOS of the facility comprising a series of segments is the weighted average of the LOSs of the individual segments. It is based on a combination of travel speed and delay calculated by following a sequence comprising the steps of determination of

- Base free-flow speed
- Travel speed
- Spatial stop rate
- LOS



Data acquired

Total vehicles arriving	=	V_{tot} (vehicles)
Total vehicles stopped	=	V_{stop} (vehicles)
Counting period	=	I_q (seconds) (usually 15 second intervals)
Number of lanes	=	N
Number of cycles surveyed	=	N_c

Calculation

Total vehicle seconds spent in queue	=	ΣV_{iq} (seconds)
Time in queue per vehicle	=	$d_{vq} = (I_q \frac{\Sigma V_{iq}}{V_{tot}}) * 0.9$ (secs/vehicle)
Fraction of vehicles stopping	=	$\text{FVS} = \frac{V_{st}}{V_{tot}}$
Correction factor from Table 20.3	=	CF
No. of vehicles stopping/lane/cycle	=	$\frac{V_{stop}}{N_c * N}$
Acceleration-deceleration correction delay	=	$d_{ad} = \text{FVS} * \text{CF}$
Control delay	=	$d = d_{vq} + d_{ad}$

Figure 19.3 Correction for deceleration and acceleration.

Base free-flow speed is calculated as the quotient of the overall length of the facility and the sum of the free-flow travel times across all the segments and shown in the relationship:

$$S_{fo,F} = \frac{\sum_{i=1}^{m} L_i}{\sum_{i=1}^{m} \frac{L_i}{S_{fo,i}}}$$

where

$S_{fo,F}$ = base free flow speed for the facility (miles/hour)
L_i = length of segment i (ft.)
m = number of segments comprising the facility
$S_{fo,i}$ = base free-flow speed for segment i

The base free flow speed is calculated using the procedure described in HCM 2010, Chapter 17. The travel speed is calculated by the use of a similar formula except that the base free-flow speed in each segment is replaced by the actual speed achieved.

Delay is reflected by the number of full stops determined for each vehicle in relation to the distance travelled as shown in the relationship:

$$H_F = \frac{\sum_{i=1}^{m} H_{seg,i}}{\sum_{i=1}^{m} L_i}$$

where

H_F = spatial stop rate for the facility (stops/mile)
$H_{seg,i}$ = spatial stop rate for segment i

with the other variables as previously defined.

19.6.4 Pedestrians

In both cases, pedestrians and cyclists, the derivation of the LOS is based on scores reflecting the perceptions of the respective road users. The process followed in determination of the LOS has three steps in common, which are the determination of

- Travel speed
- User score, followed by
- LOS

In the case of pedestrians, the level of personal space required is also an issue. No one likes being crowded but circumstances can and do cause a change in outlook. For example, in a crowded elevator, people are prepared to accept others standing so close to them that they are almost touching, whereas, if walking along a sidewalk, a clear metre between people would be considered necessary. In the calculation of LOS, a fourth step is thus required, specifically preceding the three in the preceding list to address this issue.

The full analysis of LOS for a pedestrian facility as described in HCM 2010 requires a process involving 10 steps as outlined in HCM 2010, Chapter 18.

Steps 1 and 4 determine free-flow pedestrian speed and pedestrian travel speed respectively. Step 2 determines the average pedestrian space. This step includes four sub-steps:

1. The determination of the effective sidewalk width, which is basically the total width with reductions applied for obstacles such as benches and trees within the sidewalk. No one is prepared to walk so close to a wall to be continuously brushing against it so that one of the allowances is for shy distance.
2. Determinations of the pedestrian flow rate per unit width of the sidewalk.
3. Calculation of the average walking speed.
4. Calculation of the average pedestrian space.

The effective sidewalk width is illustrated in Figure 19.4.

Step 3 relates to pedestrian delay, for which there are three possible sources:

- Crossing an intersection parallel to the centreline of the segment
- Crossing the segment itself at right angles to the general direction of flow
- Waiting for an opportunity to cross over the segment at an uncontrolled intersection

With steps 1 to 4 completed, steps 5 and 6 calculate the pedestrian LOS scores for the intersection and the link respectively. Step 7 determines the LOS for the link. Step 8 follows on after step 5 by determining the roadway crossing difficulty factor and hence, in step 9, the overall pedestrian LOS score for the segment. Finally, in step 10, the total LOS of the segment is determined.

The pedestrian facility LOS score is computed from those of the individual segments of the facility using the relationship

$$I_{p,F} = \frac{\sum_{i=1}^{m} I_{p,seg,i}}{\sum_{i=1}^{m} L_i}$$

where

$I_{p,F}$ = pedestrian LOS score for the facility
$I_{p,seg,i}$ = pedestrian LOS score for segment i
m = number of segments in the facility
L_i = length of segment i

Finally the pedestrian LOS is determined by application of the LOS score calculated as shown and the average pedestrian space followed by comparison with the values associated with each as shown in Table 19.2.

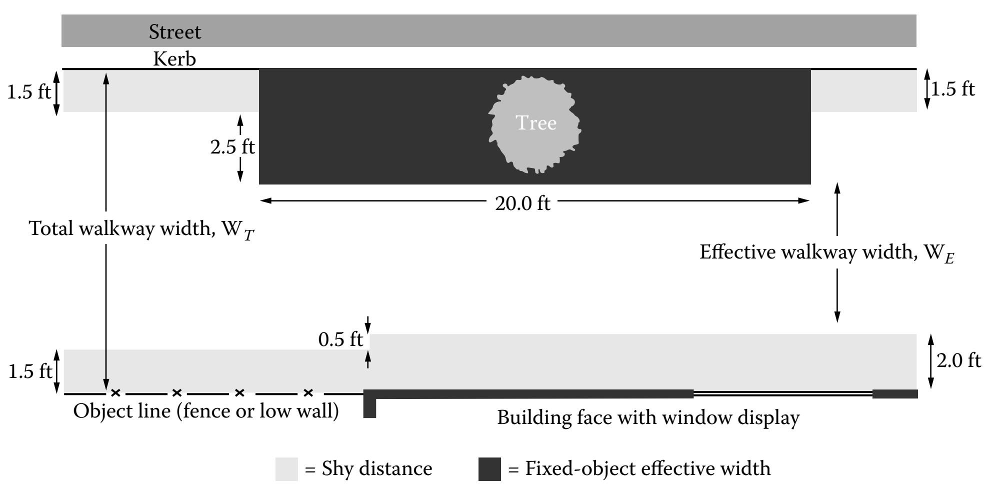


Figure 19.4 Effective pavement width.

Table 19.2 LOS criteria for pedestrian mode

Pedestrian LOS score	*LOS by average space (ft²/p)*					
	<60	*>40–60*	*>24–40*	*>15–24*	*>8–15*	*8*
≤2.00	A	B	C	D	E	F
>2.00–2.75	B	B	C	D	E	F
>2.75–3.50	C	C	C	D	E	F
>3.50–4.25	D	D	D	D	E	F
>4.25–5.00	E	E	E	E	E	F
>5.00	F	F	F	F	F	F

Note: p, pedestrian.

19.6.5 Cyclists

As is the case with determination of the LOS for pedestrians, the LOS for cyclists is based on the determination of a LOS score and then comparison with the thresholds provided in Table 19.3.

The calculation comprises eight steps. Steps 1 and 3 relate to the determination of running speed and travel speed respectively and step 2 to intersection delay. Step 4 involves the determination of the LOS score for the boundary intersection of the segment and step 5 the LOS score for the link, leading to determination of the link LOS in step 6. The combination of the LOS scores for the intersection and the link provides the LOS score for the segment and hence, in step 8, determination of the segment LOS. The calculations involved in each of these steps are shown in HCM 2010, Chapter 17. The LOS score is compared to the boundaries of the various LOS scores as shown in Table 19.3 to determine what the LOS of the segment is.

The LOS of the facility is the weighted average of the LOSs of the individual segments and calculated as

$$I_{b,F} = \frac{\sum_{i=1}^{m} I_{b,seg,i} L_i}{\sum_{i=1}^{m} L_i}$$

where

$I_{b,F}$ = cyclist LOS for the facility
$I_{b,seg,i}$ = cyclist score for segment i
m = the number of segments comprising the facility
L_i = the length of segment i

Table 19.3 LOS criteria for cyclists and transit users

LOS score	*LOS*
≤2.00	A
>2.00–2.75	B
>2.75–3.50	C
>3.50–4.25	D
>4.25–5.00	E
>5.00	F

Source: Transportation Research Board. *Highway Capacity Manual 2010*. Washington, DC, 2010, Exhibit 17–4.

19.6.6 Transit

The performance of transit operating in mixed traffic or exclusive bus lanes and stopping along the street is described in HCM 2010, Chapter 16. It is based on transit travel speed, calculated for each segment of the facility as described in HCM 2010, Chapter 17. This is used to determine a LOS score, which is compared with the boundary conditions for the various LOS scores as shown in Table 19.2 to determine the LOS provided by the segment and subsequently the LOS of the facility. The process is very similar to that applied to determination of the LOS for pedestrians and cyclists and comprises a sequence of seven steps, which are the determination of:

- Transit vehicle running time
- Delay at intersections
- Travel speed
- Transit wait-ride score
- Pedestrian LOS score for link as described previously
- Transit LOS for segments
- LOS for the facility

The procedures for calculating the LOS scores for the facility segments are described in HCM 2010, Chapter 17. The LOS score for the facility as a whole is the weighted average:

$$I_{t,F} = \frac{\sum_{i=1}^{m} I_{t,seg,i} L_i}{\sum_{i=1}^{m} L_i}$$

where

$I_{t,F}$ = transit LOS score for the facility
$I_{t,seg,i}$ = transit LOS score for segment I
m = number of segments comprising the facility
L_i = length of segment i

Chapter 20

Capacity analysis

Uninterrupted flow

CONTENTS

20.1 INTRODUCTION

As described in Chapter 19, flow may be either interrupted or uninterrupted, where the adjective describes the facility on which it occurs and not the quality of flow experienced. Uninterrupted flow occurs on freeways and on multilane and two-lane roads where the intersections are all priority controlled and the roads of interest are the major roads at each intersection.

The geometric design application of analysis of freeways is to determine the number of lanes required to achieve a desired level of service (LOS) either on an existing or proposed freeway facility on the basis of a forecast design value and basic design parameters such as lane width and lateral clearance. Freeways are free in the sense of freedom from interruption of mobility and not freedom from payment as on a toll road. A toll road can therefore also, but not necessarily, be a freeway.

Similarly in the case of multi-lane roads operating under uninterrupted flow conditions, the geometric design application of analysis is determination of the number of lanes required to achieve a target LOS.

With two-lane roads, the number of lanes is a given except that, at certain points along the road, gradients may cause speeds and hence the LOS to drop below the target value, requiring the adoption of climbing lanes at the trouble spots. Similarly, topography resulting in a lack of passing sight distance or short radius curvature at frequent intervals may also cause a drop in speed and hence LOS. Lack of passing opportunity may cause a drop in LOS because the volume of opposing traffic may require passing lanes to support the dispersal of platoons behind slow-moving vehicles.

The difference between climbing lanes (also referred to as truck lanes, crawler lanes and, totally incorrectly, as passing lanes) and passing lanes is that the former is used to match the LOS on the grade to that achieved on the level sections elsewhere along the road whereas the latter is intended to increase the LOS of the road above that of the conventional two-lane road. The passing lane thus serves as a transitional stage towards a multi-lane road. The 2 + 1 cross-section is nothing more or less than a two-lane road with successive passing lanes alternating between the two directions of travel. Not entirely facetiously, a four-lane road could be described as being a two-lane road with continuous passing lanes in each direction, hence illustrating the increase in capacity resulting from the provision of passing lanes.

20.2 FREEWAY SEGMENTS

As in the case of interrupted flow, a freeway facility comprises segments, and the LOS of the freeway is thus the weighted average of the LOSs of the individual segments. A facility is not intended to be the entire freeway from end to end. It is simply that portion of the freeway comprising the various segments that have been identified as in need of analysis. Three types of segments are identified:

- Basic freeway segments, discussed in *Highway Capacity Manual* (HCM) *2010* (Transportation Research Board, 2010), Chapter 11
- Freeway weaving segments, discussed in HCM 2010, Chapter 12
- Merge and diverge segments, discussed in HCM 2010, Chapter 13

They are illustrated in Figure 20.1.

HCM 2010 draws a distinction between sections and segments of a freeway. The freeway section is defined by road markings as being the points where the inner edges of ramp tapers meet the outside edges of the through lanes of the freeway. These points are sometimes referred to as 'yellow line break points' (Wolhuter et al., 2005) because, in South Africa, where their research was done, edge markings are yellow as opposed to the solid white lane specified by the *Manual of Uniform Traffic Control Devices* (Federal Highway Administration, 2009). They are also sometimes referred to as 'edges meet'.

Basic freeway segments are based on the influence areas of the merge and diverge tapers and the weaving areas as illustrated in Figure 20.1. The section of freeway between an on-ramp and the following off-ramp could thus include a merging segment, a basic segment and a diverging section. The influence areas of merges and diverges are, respectively 1500 feet downstream of a merge and 1500 feet upstream of a diverge measured from where edges meet. As the distance between the ramp tapers decreases, the situation could be achieved whereby the basic segment falls away and the merging and diverging segments in fact over-

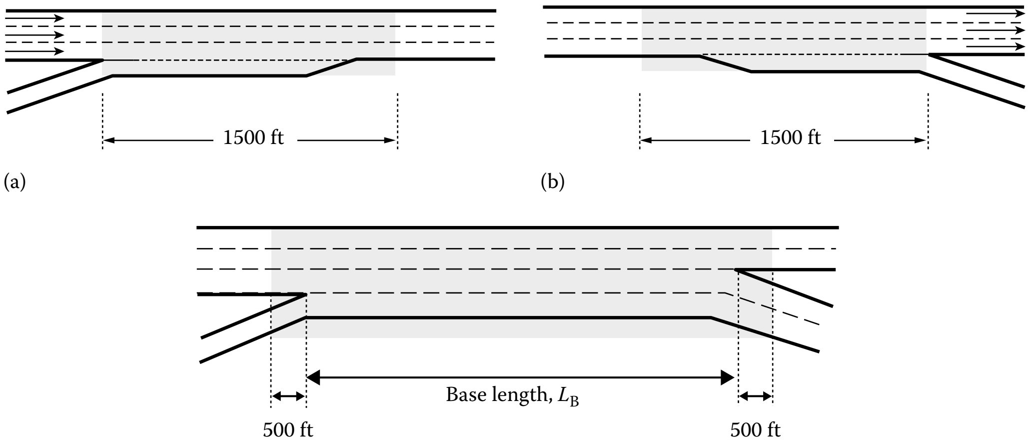


Figure 20.1 Freeway segments. (a) Merge influence area and (b) diverge influence area.

lap. The influence area of a weaving segment extends from 500 feet upstream of the merge to 500 feet downstream of the diverge.

The various algorithms used in the methodologies discussed in these chapters are described in detail in HCM 2010, Chapter 25. A section of HCM 2010, Volume 4: Applications Guide entitled Methodological Details includes an executable spreadsheet, FREEVAL 2010, which implements the complex calculations involved. It and the user's guide to it are to be found in Chapter 10 of Methodological Details.

20.2.1 Basic segments

Base conditions under which capacity of a basic segment can be achieved include

- 0 per cent trucks in the traffic stream
- A driver population that is familiar with the facility
- Lane widths of at least 12 feet
- Right side clearance widths of at least 6 feet

In addition, ideal conditions have to apply and these include

- Good weather
- Good visibility
- No crashes or incidents
- No maintenance activity
- A pavement quality good enough not to have an impact on preferred speed

Flow within a basic segment can be categorised as being

- Undersaturated, insofar as there are no impediment's either up- or down-stream of the segment
- Queue discharge, in which vehicles that have just cleared a bottleneck are accelerating to speeds appropriate to general current conditions
- Oversaturated, where vehicles are trapped in a queue generated by a downstream bottleneck

LOS scores are appropriate to undersaturated flow because oversaturated flow is a LOS F phenomenon. Queue discharge starts at LOS A immediately downstream of the bottleneck and continues to change down through successive LOSs until an equilibrium condition similar to the unsaturated condition upstream of the bottleneck is reached. The various levels are described in the following list and differ only slightly from those described in the case of interrupted flow.

- LOS A describes the condition whereby free-flow speeds are attained and the ability to change lanes is unimpeded.
- LOS B also describes the condition whereby free-flow speeds are maintained and the ability to change lanes is only slightly restricted. The general level of comfort is still high.
- LOS C provides for speeds very nearly equal to free-flow speed but freedom to change lanes is becoming restricted and requires care.
- LOS D shows speeds declining rapidly with increasing flows and density increasing sharply. Freedom to change lanes is severely restricted. The general level of comfort is low and drivers are starting to feel stressed.
- LOS E describes operations at capacity. There are virtually no gaps in the traffic stream. Changing lanes can be achieved only if a driver in the adjacent lane is prepared to drop back to make a space into which the lane change can be made.
- LOS F describes unstable flow with speed being very low and with stationary queues forming in the traffic stream even in the absence of specific bottlenecks.

Application of the density criterion employs the relationship:

$$D = \frac{v_p}{S}$$

where

D = density (passenger cars per mile per lane)
v_p = demand flow rate (passenger cars per hour per lane)
S = mean speed of traffic stream under base conditions (miles per hour)

Estimating or measuring S is difficult, and there is considerable variation in observed and predicted values. In general, speeds are normally estimated in steps of 5 mi/h from 55 mi/h to 75 mi/h.

As pointed out in Chapter 19, HCM 2010 has yet to be metricated. Accordingly, it was believed that for convenience Chapter 20 should also employ Imperial units, also referred to as US Customary units, although there are differences in the magnitude of some of the units of measurement.

The demand flow rate, v_p, is calculated from

$$v_p = \frac{V}{\text{PHF} \times N \times f_v \times f_p}$$

where

v_p = demand flow rate under equivalent base conditions
V = demand volume under prevailing conditions (vehicles per hour)
PHF = peak-hour factor
N = number of lanes in direction of interest
f_v = adjustment factor for heavy vehicles in the traffic stream
f_p = adjustment factor for unfamiliar drivers

Table 20.1 LOS criteria for basic freeway segments

LOS	*Density (passenger cars per mile per lane)*
A	≤11
B	>11–18
C	>18–26
D	>26–35
E	>35–45
F	>45 Demand exceeds capacity

With all these variables known and v_p calculated, the density is calculated using the formula given above and compared to criteria offered in Table 20.1 to establish the LOS of the basic segment.

20.2.2 Merge and diverge segments

In the analysis of merge and diverge sections for the purposes of geometric design, what is sought are the geometric characteristics of the merge or diverge that are needed to deliver the target LOS. These characteristics are

- The free-flow speed of the ramp
- The length of the acceleration and deceleration lanes
- The required number of lanes on the ramp

Because of the number of variables involved, there is no convenient way to directly calculate the value of any one of them without being able to attach values to all the others. Design analysis is thus a trial-and-error application of the operational analysis procedure shown in HCM 2010, Chapter 13, which suggests that a spreadsheet be set up to carry out the calculations.

Although much research has already been done on the analysis of the operation of on- and off-ramps, it would appear that more has still to be done to be able to determine the LOS applying to them with confidence. It would appear that a major factor in the determination of this LOS is the tendency of vehicles on the freeway upstream of the ramp to move to the left to avoid the turbulence associated with ramp operation. An estimate of the number of vehicles remaining in the outside lane(s) is required. This, in combination with the number of vehicles on the ramps themselves, will provide an indication of the density and hence the LOS of the merge and diverge sections of the freeway.

The density within the influence area of an on-ramp is given by

$$D_R = 5.475 + 0.00734v_R + 0.0078v_{12} - 0.00627L_A$$

where

D_R = density in the ramp influence area (passenger cars per mile per lane)
v_R = flow rate on ramp (passenger cars per hour)
v_{12} = flow rate in freeway lanes 1 and 2 (passenger cars per hour)
L_A = length of acceleration lane and taper (feet)

In the case of an off-ramp the density within the influence area is given by

$$D_R = 4.252 + 0.0086v_{12} - 0.009L_D$$

where
L_D = length of deceleration lane and taper (feet)

and the other variables are as previously defined. In the case of a four-lane freeway, the flow rate in lanes 1 and 2 is simply the total approach flow rate so that $P_{FM} = 1$. In the cases of six-lane and eight-lane freeways the flow rate in lanes 1 and 2 is

$$v_{12} = v_f \times P_{FM}$$

where
v_f = total flow rate on freeway upstream of the on-ramp
P_{FM} = proportion of freeway traffic remaining in lanes 1 and 2

In the case of a six-lane freeway, the proportion is the greatest of

$$P_{FM} = 0.5775 + 0.000028L_A$$

or

$$P_{FM} = 0.7289 - 0.0000135(v_f + v_r) - 0.003296S_{FR} + 0.000063L_{UP}$$

or

$$P_{FM} = 0.5487 + 0.2628 \frac{v_D}{L_{DOWN}}$$

where
L_A = length of acceleration lane plus taper
v_f = total flow rate on freeway upstream of the on-ramp
v_r = total flow rate on ramp
S_{FR} = free flow speed on ramp
L_{UP} = length of freeway to preceding ramp
L_{DOWN} = length of freeway to following ramp

Similar relationships are offered in HCM 2010, Chapter 13, for off-ramps.
LOS criteria for merge and diverge segments are shown in Table 20.2.

Table 20.2 LOS criteria for merge and diverge segments

LOS	*Density (passenger cars per mile per lane)*	*Comments*
A	>10	Unrestricted operation
B	>10–20	Merging and diverging manoeuvres noticeable to drivers
C	>20–28	Influence area speeds begin to decline
D	>28–35	Influence area turbulence becomes intrusive
E	>35	Turbulence felt by nearly all drivers
F	Demand exceeds capacity	Ramp and freeway queues form

20.2.3 Weaving segment

Weaving is defined as the operation in which two streams of traffic travelling in the same direction cross each other over an extended length of the freeway without the benefit of traffic control devices. If the on-ramp and the off-ramp are sufficiently removed from each other, they can function independently. However, as this distance shortens, the operation of each ramp begins to impact on the operation at the other and a weaving section is formed. The weaving operation can take on two different forms:

- One-sided ramp weave (previously type A), where both streams of traffic have to undertake a lane change to complete their weave
- One-sided major weave (previously type B), where one stream of traffic has to undertake a lane change to complete its weave

These weaving operations are illustrated in Figure 20.2.

In the previous editions of HCM, reference was also made to a type C weave where one stream had to undertake two lane changes to complete its weave. This has now been dropped, presumably because of its rarity but also possibly because it does not lend itself to analysis.

Two-sided weaves could also exist but are not recommended practice. Left-side merges and diverges do not match driver expectancy. In the absence of lane balance (as illustrated in HCM 2012, Exhibit 12-4) the diverge could have the problem of a vehicle in the fast (or left) lane being deflected off the freeway into a local area. Drivers in the leftmost lane of the freeway would be paying more attention to vehicles to their right and certainly would not expect a vehicle to suddenly appear in front of them from the left. Furthermore, a driver entering from the right and intending to exit to the left would, if on a three-lane carriageway, have to execute a merge followed by two lane changes and a diverge in quick succession to achieve his or her intention (Figure 20.3).

As in the case of isolated merges and diverges, the criterion of LOS for weaving is density. The procedure followed is to calculate the speeds of the weaving and nonweaving flows and apply these to get a weighted average representing the speed of all the vehicles traversing the weaving section. With this speed and the number of vehicles in the weaving section, the density can be calculated and compared to the criteria for weaving sections shown in Table 20.3.

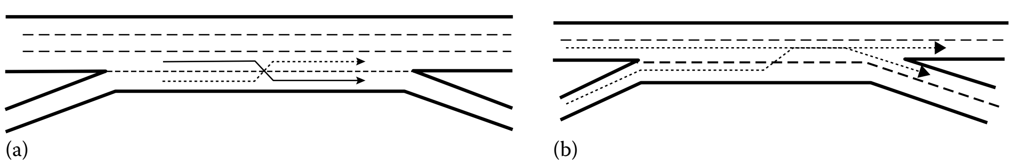


Figure 20.2 One-sided weaves. (a) One-sided ramp weave and (b) one-sided major weave.

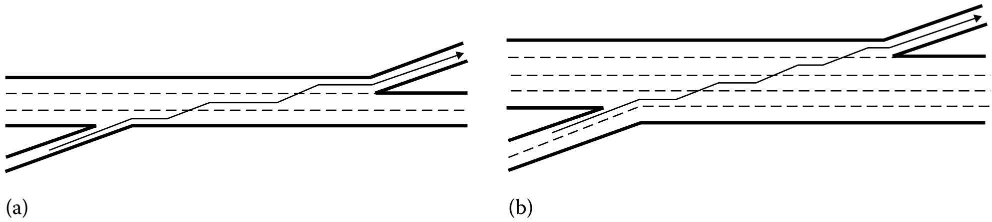


Figure 20.3 Two-sided weaves. (a) One merge - two lane changes - one diverge and (b) one merge - three lane changes - one diverge.

Table 20.3 LOS criteria for weaving segments

LOS	*Density (passenger cars per mile per lane)*
A	0–10
B	>10–20
C	>20–28
D	>28–35
E	>35

20.3 FREEWAY CAPACITY

Freeway capacity is defined as the capacity of the critical segment of those comprising the defined facility. The critical segment is the segment most likely to break down first as long as all traffic, roadway and control conditions do not change. The methodology to determine capacity relies on the adoption of a basic capacity that is shown in Table 20.4.

The free-flow speed is predicated on the basis of there being no trucks in the traffic stream and that all drivers are familiar with the freeway section being studied. The most important aspect of analysis of freeways is the free-flow speed, as it is the basis for analysis of basic segments, merge and diverge segments and weaving segments.

The free-slow speed is estimated by the relationship

$$\text{FFS} = 75 - f_{\text{LW}} - f_{\text{LL}} - 3.22\ \text{TRD}^{084}$$

where

FFS = free-flow speed (miles/hour)
f_{LW} = adjustment for lane width (miles/hour)
f_{LL} = adjustment for lateral clearance (miles/hour)
TRD = total ramp density (ramps/mile)

Delay is the criterion for determination of LOS of basic segments, diverge and merge segments and hence for the freeway facility in total. The LOSs for various values of density are shown in Table 20.5.

20.4 ADJUSTMENT FACTOR FOR TRUCKS ON FREEWAYS

The adjustment for trucks addresses the mathematical problem of analysis of a heterogeneous traffic stream. It is more convenient to have a traffic stream comprising a single class of vehicles, a homogeneous traffic stream. Each truck in the traffic stream is thus converted to an equivalent number of passenger cars (PCEs). The equivalence is dependent on the gradient and length

Table 20.4 Basic capacity for various free-flow speeds

Free-flow speed (miles/hour)	*Basic capacity*
75	2400
70	2400
65	2350
60	2300
55	2250

Table 20.5 Density on freeway segments for various LOS

LOS	*Density (passenger cars per mile per lane)*
A	≤11
B	>11–18
C	>18–26
D	>26–35
E	>35–45
F	>45

of grade, whether the gradient is positive (uphill in the direction of travel) or negative (downhill in the direction of travel) and the percentage of trucks in the traffic stream.

The factor is calculated from the percentage of trucks and recreational vehicles in the traffic stream and their PCEs based on the gradient and length of the grade:

$$f_{\mathrm{HV}} = \frac{1}{1 + P_{\mathrm{T}}(E_{\mathrm{T}} - 1) + P_{\mathrm{R}}(E_{\mathrm{R}} - 1)}$$

where

f_{HV} = heavy-vehicle adjustment factor
P_{T} = proportion of trucks and buses in traffic stream
P_{R} = proportion of RVs in traffic stream
E_{T} = passenger-car equivalent (PCE) of one truck or bus in traffic stream
E_{R} = PCE of one RV in traffic stream

The PCEs for a truck are offered in HCM 2010, Exhibit 11-10, for general terrain segments and in HCM 2010, Exhibit 11-11, for upgrades. PCEs for recreational vehicles on upgrades are offered in HCM 2010, Exhibit 11-12. PCEs for trucks and buses on downgrades are shown in HCM 2010, Exhibit 11-13.

Analysis as described in the preceding text is for specific grades that may have positive or negative gradients. Alternatively, analysis could be for general terrain that addresses an extended length of freeway containing a number of up- or downgrades with no single grade being long enough to have a significant impact on the operation of the freeway. As an indication of the limitations of this form of analysis, it should not be applied over a distance that is longer than 0.5 mile or steeper than a gradient of 3 per cent. The topography traversed by the freeway may be

- Level or flat terrain which comprises gradients of a magnitude that does not prevent trucks from maintaining speeds similar to those of passenger cars. In many countries, differential speed limits are placed on passenger cars and trucks, for example, 75 miles/hour for cars and 50 miles/hour for trucks. This is purely an administrative decision and the definition is based on what heavy trucks can do and not on what they are allowed to do.
- Rolling terrain, which forces trucks to operate at speeds achievable in the lower ranges of their gears but does not force crawl speeds for any significant length of road.
- Mountainous terrain, which forces trucks to operate at the lowest ranges of their gears over significant lengths of road.

Steep grades would typically be longer and/or steeper than would normally be accepted for a general terrain segment.

20.5 MULTI-LANE ROADS

20.5.1 Passenger car mode

Multi-lane divided roads have capacities similar to freeways when in rural areas where intersections are widely spaced. In urban and suburban areas, the intrusion of closely spaced intersections and driveways justifies analysis that includes the various correction factors required to adjust the free-flow speed downwards. The geometric design application of these analyses is to determine the number of lanes required to deliver a target LOS given a forecast demand volume and definition of the various parameters including

- Lane width
- Lateral clearance
- Number of intersections and driveways per mile

The methodology does not, as yet, include correction factors for the presence of on-street parking or pedestrian activity as found in urban and suburban areas.

The calculation of the number of lanes required commences with the assumption of the free-flow speed being equal to design speed. Normally there would not be a need to consider adjustment factors because a new design should be in terms of ideal conditions of lane and lateral clearance widths. Only in the most stringent of topographic conditions or limited right of way would it be necessary to adopt lesser widths followed by the application the appropriate correction factors. Furthermore, there would seldom be any need for climbing lanes because the slow-moving vehicles would tend to gravitate towards the right-hand lane, with the left lane providing adequate opportunity for passing these vehicles. Under flow conditions approaching capacity it may be necessary to provide passing lanes but, at this stage, consideration should maybe be given to increasing the number of basic lanes.

The 2 + 1 cross-section could perhaps be considered to be a multi-lane road but HCM 2010 does not explicitly address it as such. It is dealt with in the Appendix to HCM 2010, Chapter 15, as analysis of a two-lane road with passing lanes.

The number of lanes required to provide the target LOS is calculated from the relationship

$$N = \frac{Q}{D \times S}$$

where

N = required number of lanes
Q = traffic flow assumed for design (vehicles per hour)
D = lower boundary density for target LOS
S = desired operating speed (miles/hour)

No analysis methodology is provided in HCM 2010 for the consideration of the possible need for climbing lanes on multi-lane roads. It is, however, suggested that a possible way of evaluating this need could be to assume that all trucks would be concentrated on the right-hand lane and all passenger cars in the left-hand lane. This new value of Q for the left-hand lane applied to the above relationship would provide an indication of the need for a passing lane by showing N as being greater than 1. Obviously, the device of splitting the traffic stream into two separate streams eliminates the need for calculating the PCEs of the

Table 20.6 Lower boundary densities for multi-lane road LOSs

LOS	*Lower boundary density (vehicles per mile)*
A	11
B	18
C	26
DS	35
E	40

Note: In the case of LOS E, lower boundary densities are offered for slower FFS as shown below:

Free-flow speed (miles per hour)	*Lower boundary density*
60	40
55	41
50	43
45	45

truck traffic for all the gradients that may be encountered during the design of the vertical alignment of the road.

If the value of N is only slightly higher than 1, the decision should probably be to forgo the need of a climbing lane simply because of the inaccuracy inherent in the prediction of traffic flows 20 years hence. Any value closer to 2 may merit serious consideration of the provision of a climbing lane and a value greater than 2 should be considered to be a warrant for a climbing lane.

The densities serving as the lower boundaries of the various LOSs as input to the above relationship are shown in Table 20.6. The lower boundary of LOS E represents capacity so that LOS F doesn't warrant inclusion in the table.

20.5.2 Bicycle mode

Cyclists are concerned only with the right-hand or outer lane of a multi-lane facility and the number of lanes to the left of it is a matter of indifference. It is unlikely that a multi-lane road would be built with other than surfaced shoulders. If the shoulders are, in fact, surfaced, cyclists would use them in preference to sharing the right lane with fast-moving motorised vehicles.

As in other cases, the LOS is established by the calculation of an LOS score. This score is based on a traveller perception index calibrated by using a linear regression model as described in Chapter 16. The score is compared to the boundary values of the various LOSs shown in Table 20.7 to determine the LOS for the bicycle mode.

Table 20.7 Boundary values of LOS scores for bicycles

LOS	*Boundary values of LOS scores*
A	1.5
B	2.5
C	3.5
D	4.5
E	5.5

20.6 TWO-LANE ROADS

20.6.1 Introduction

Two-lane roads are the Cinderellas of the road network. Seventy to eighty per cent of the roads in developing countries are either earth roads or have gravelled surfaces. In neither case are there distinguishable shoulders, and the width of these roads is typically 20 feet or even less. Two-lane surfaced roads often have gravel shoulders and have a similar width, although the more important routes can have lanes up to 12 feet in width.

The unique characteristic of two-lane roads is that the passing manoeuvre is undertaken in the opposing lane. In consequence, the opportunity for passing manoeuvres is limited by the combination of availability of sight distance and gaps in the opposing traffic stream. As traffic flows increase, passing opportunities diminish while, simultaneously, demand for passing increases. In consequence, operating conditions decline rapidly with increasing flows and reach levels unacceptable to road users before capacity levels are reached.

20.6.2 The services provided by two-lane roads

Where freeways and multilane roads are typically geared towards the function of mobility, two-lane roads address the entire spectrum from rural mobility to urban accessibility including access to the individual homes on a residential street.

As a link between towns, two-lane roads serve the needs of long-distance commercial and recreational travel. Substantial lengths of road may be free of intersections and sustained high travel speeds are possible. Infrequent passing delays are desirable.

Two-lane roads also serve as links between villages and other sparsely populated areas and nearby larger towns. Travel demands are usually low and cost effective all-weather access is the prime function of these roads. Rural two-lane roads also serve a recreational function for motorists, cyclists and walkers. Enjoyment of the scenery suggests that high speeds should be actively discouraged although difficulties in passing may be an unwelcome distraction.

These roads may pass from time to time through small villages such as in the Lake Districts of the United Kingdom, where villages are located at distances of 3 miles or less apart. Speed limits through these areas are often very low, typically of the order of 25 to 30 miles per hour. These areas are usually short in length and are characterised by relatively high volumes of pedestrian and cyclist activities. Other road users expect not to have much in the way of passing opportunities.

20.6.3 The classification of two-lane roads

As illustrated previously, two-lane roads serve the entire functional spectrum from mobility to accessibility and the analysis methodology and criteria are thus geared to addressing a wide range of road user expectations and not only those of drivers. This spread of functionality is addressed in HCM 2010 by considering two-lane roads in three classes:

- Class 1 roads, which are primarily geared to the mobility function. These address the expectation of relatively high speeds on long-distance trips in rural areas and serve as
  - Major links between towns
  - Primary links between major traffic generators
  - Commuter routes between nearby towns and cities

- Class 2 roads, where their functionality is a mix of the mobility and accessibility functions and there isn't the expectation of high-speed travel. These roads often serve relatively short trips or are the beginnings and endings of trips on higher order facilities. They serve as
  - Links between local areas and the roads comprising the Class 1 roads and also the roads that are higher in the functional hierarchy
  - Scenic or recreational routes where attention must also be given to road users other than drivers
  - Pass-through adverse topography where the cost of construction to achieve high-speed operation would be unaffordable or uneconomic
- Class 3 roads, which are essentially urban in nature. They may be the portions of Level 1 or Level 2 roads passing through a small town and usually have a mix of local and through traffic. The frequency of local intersections and other access points is much higher than in a purely rural area. Class 3 roads often have reduced speed limits matching their urban, low-speed function.

20.6.4 Analysis of the performance of two-lane roads

There are three criteria of the LOS of two-lane roads:

- Average travel speed (ATS). This is the macroscopic speed described in Chapter 20 and is indicative of mobility.
- Percentage of time spent following (PTSF). This represents the freedom to manoeuvre, which, in the case of a two-lane road, is the ability to overtake at will. With increasing opposing flows, this freedom obviously diminishes.
- Percentage of free-flow speed (PFFS). This is essentially an urban measure of quality of flow as it reflects the ability of vehicles to travel at or near the posted speed limit.

The road has to be considered as a series of uniform segments and each direction of travel has to be analysed separately. Uniform segments have similar traffic and roadway conditions, with their boundaries being at points where a change occurs in

- The terrain
- Width of lanes
- Widths of shoulders
- Intersections where there is a significant change in demand flow rate

Analysis for the purposes of design can, at best, be considered only rather rough and ready as most of the traffic demand data input are default values and each default applied makes the final outcome of the analysis that much more approximate. Clearly, the concept of 'what is good enough' is stretched to the limit. The design application of the analysis is to determine the LOS achieved at various flow levels, hence establishing the need for climbing and/or passing lanes.

As is the case with the other analyses, the analysis of two-lane roads is based on ideal conditions in combination with correction factors addressing situations that are less than ideal. The base conditions for ideal conditions include

- Lanes that are 12 feet or wider
- Shoulders that are at least 6 feet wide
- Level terrain
- No impedance by turning or parking vehicle

Table 20.8 Boundary values of LOS for two-lane roads

LOS	*Class 1 highways*		*Class 2 highways*	*Class 3 highways*
	Average travel speed	*Percentage of time spent following*	*Percentage of time spent following*	*Percentage of free-flow speed*
A	>55	≤35	≤40	>91.7
B	55	50	55	91.7
C	50	65	70	83.3
D	45	80	85	75.0
E	≤40	>80	>85	≤66.7

The criteria of effectiveness of the performance of two-lane roads broadly are in line with the classes into which two-lane roads are divided. For Class 1 roads, the criteria are average travel speed (ATS) and percentage time spent following (PTSF). This is because both travel speed and delay due to lack of passing opportunities are important to drivers. On Class 2 roads, speeds are less important and the criterion of effectiveness is thus only PTSF. On Class 3 roads, travel distances are usually fairly short so that passing opportunities are less important. However, there is a preference by drivers to be able to maintain speeds that are at or close to the posted speed limit so that the criterion of LOS is the percentage of free-flow speed (PFFS) achieved.

The lower values of the various criteria for the three classes of two-lane roads are shown in Table 20.8.

Drivers' perceptions of the quality of service by two characteristics: passing capacity and passing demand. Passing capacity reflects the ability to pass in the face of opposing traffic, which declines as opposing traffic volumes increase. Passing demand increases with increases in traffic volume in the direction of interest. Both capacity and demand are related to flow rates. The various LOSs are defined as

- LOS A. High operating speeds can be maintained and there is little difficulty in passing on Class 1 highways. Platoons of three or more vehicles are rare. On Class 2 highways, a small amount of platooning can be expected. On Class 3 highways vehicles should be able to maintain speeds close to the signposted speed.
- LOS B. Passing demand and passing opportunity match each other and, on Classes 1 and 2, platooning becomes noticeable. On Class 3, speeds close to the signposted speed can be maintained.
- LOS C. A large percentage of vehicles are traveling in platoons on all three classes of highway and speed is noticeably reduced.
- LOS D. Almost all vehicles are travelling in platoons and PTSF is noticeable. On Class 3 highways the reduction in FFS is marked.
- LOS E. Passing on Classes 1 and 2 highways is virtually impossible and PTSF is more than 80 per cent. On Class 3 highways, speed is less than 65 per cent of FFS.
- LOS F. The demand flow exceeds capacity in both directions and heavy congestion exists. On all three classes, movement is at very low speeds and stop–start conditions prevail.

20.6.4.1 LOS in terms of ATS

The general procedure is to establish the class of the two-lane road and then to assess the ATS if the road falls into either a Class 1 or Class 3. Class 1 roads address high-speed flow for extended distances in rural areas and design is aimed at achieving a target LOS, usually LOS B in the design hour. Class 3 is essentially an urban application where geometric design is typically aimed at the achieving of LOS C or, in heavily developed areas, LOS D.

The procedure followed is to estimate the free-flow speed from the base free-flow speed (BFFE) and apply corrections to it on the basis of the peak hour factor (PHF) with adjustments for lane and shoulder width and access point density and using the relationship

$$\text{FFS} = \text{BFFS} - f_{\text{LS}} - f_{\text{A}}$$

where

FFS = free-flow speed (miles/hour)

BFFS = base free-flow speed (miles/hour)

f_{LS} = adjustment for lane and shoulder width (miles/hour)

f_{A} = adjustment for access point density (miles/hour)

For design applications, the BFFS can be assumed to be the design speed plus 10 per cent. The adjustments for lane and shoulder widths are read off HCM 2010, Chapter 15, Exhibit 15-7. The access point density is the quotient of the number of unsignalised intersections and driveways on both sides of the road (as both can impact on the speed in the direction of interest). The adjustment is read off HCM 2010, Chapter 15, Exhibit 15-8.

With the free-flow speed calculated, it is necessary to also determine the demand flow rate. This is derived from the relationship

$$v_{\text{i,ATS}} = \frac{V_{\text{i}}}{\text{PHF} \times f_{\text{g,ATS}} \times f_{\text{HV,ATS}}}$$

where

$v_{\text{i,ATS}}$ = demand flow rate, with i being either d for analysis direction or o for opposing direction

V_{i} = demand volume for direction i

PHF = peak hour factor

$f_{\text{g,ATS}}$ = gradient adjustment factor

$f_{\text{HV,ATS}}$ = heavy vehicle adjustment factor

Exhibit 15-10 in HCM 2010, Chapter 15, lists adjustment factors, $f_{\text{g,ATS}}$, for specific gradients and Exhibit 15-9 provides correction factors for general gradients. The heavy vehicle adjustment factor, $f_{\text{HV,ATS}}$, is defined by the relationship

$$f_{\text{HV,ATS}} = \frac{1}{1 + P_{\text{T}}(E_{\text{T}} - 1) + P_{\text{R}}(E_{\text{R}} - 1)}$$

where

$f_{\text{HV,ATS}}$ = heavy vehicle adjustment factor for ATS estimation

P_{T} = proportion of trucks in the traffic stream (decimal)

E_{T} = PCEs for trucks from Exhibits 15-11 or 15-12

P_{R} = proportion of recreational vehicles in the traffic stream (decimal)

E_{R} = PCEs for recreational vehicles from Exhibits 15-11 or 15-13

On extended downgrades, drivers of heavy vehicles often engage their lowest gear to use the braking power of the engine because continued use of the brakes would tend to lead to overheating, brake failure and the consequent runaway. A different calculation of $f_{HV,ATS}$ from that given above must therefore be applied and this is given by the relationship

$$f_{HV,ATS} = \frac{1}{1 + P_{TC} \times P_T(E_{TC} - 1) + (1 - P_{TC}) \times P_T \times (E_T - 1) + P_R(E_R - 1)}$$

where

$f_{HV,ATS}$ = correction factor for heavy vehicles for ATS estimation
P_{TC} = proportion of trucks operating at crawl speed (decimal)
P_T = proportion of trucks in the traffic stream (decimal)
E_{TC} = PCEs for trucks operating at crawl speed from Exhibit 15-14
E_T = PCEs for trucks in the traffic stream from Exhibits 15-11 or 15-12
P_R = proportion of recreational vehicles in the traffic stream
E_R = PCEs for recreational vehicles in traffic stream from Exhibits 15-11 or 15-13

With the FFS and the demand flow rate calculated it is possible to estimate the ATS from the relationship

$$\mathrm{ATS_d} = \mathrm{FFS} - 0.007766(v_{d,ATS} + v_{o,ATS}) - f_{np,ATS}$$

where

$\mathrm{ATS_d}$ = average travel speed in the analysis direction
FFS = free-flow speed
$v_{d,ATS}$ = demand flow rate in the analysis direction
$v_{o,ATS}$ = demand flow rate in the opposing direction
$f_{np,ATS}$ = adjustment factor for the percentage of no-passing zones in the analysis direction from Exhibit 15-15

20.6.4.2 LOS in terms of PTSF

The general approach to the calculation of LOS with respect to PTSF is similar to that for calculation with respect to ATS. The base PTSF is calculated and this is used in the calculation of the PTSF, with the latter relationship requiring calculation of the demand flow rates in the direction of interest, $v_{d,PTSF}$, and the opposing direction, $v_{o,PTSF}$. These in turn require the input of corrections for gradient and heavy vehicles as well as adjustment of the percentage no-passing zones. The PTSF is then compared to the boundary values of PTSF for the various LOSs to establish the LOSs of the segment being evaluated.

The relationships used are thus

$$\mathrm{BPTSF_d} = 100 \quad 1 - \exp\left(a v_d^b\right)$$

where the factors *a* and *b* are drawn from Exhibit 15-20 of HCM 2010, Chapter 15.

$$\mathrm{PTSF_d} = \mathrm{BPTSF_d} + f_{np,PTSF} \quad \frac{v_{d,PTSF}}{v_{d,PTSF} + v_{o,PTSF}}$$

where
- $PTSF_d$ = percentage of time spent following in the direction of interest
- $BPTSF_d$ = base percentage of time spent following in the direction of interest
- $f_{np,PTSF}$ = adjustment for the percentage of no-passing zones from Exhibit 15-21
- $v_{d,PTSF}$ = demand flow rate in the direction of interest
- $v_{o,PTSF}$ = demand flow in the opposing direction

The demand flow rate is calculated from

$$v_{i,PTSF} = \frac{V_i}{PHF \times f_{g,PTSF} \times f_{HV,PTSF}}$$

where
- $v_{i,PTSF}$ = demand flow rate for direction i with i = d or i = o for the direction of interest or for the opposing direction
- V_i = demand volume for direction i
- $f_{g,PTSF}$ = grade adjustment factor for PTSF determination from Exhibit 15-16 or Exhibit 15-17
- $f_{HV,PTSF}$ = heavy vehicle adjustment factor for determination of PTSF

The heavy vehicle adjustment factor, $f_{HV,PTSF}$, is calculated from the relationship:

$$f_{HV,PTSF} = \frac{1}{1 + P_T(E_T - 1) + P_R(E_R - 1)}$$

where
- $f_{HV,PTSF}$ = heavy vehicle adjustment factor for PTSF determination
- P_T = proportion of tracks in the traffic stream (decimal)
- E_T = PCEs of trucks on specific grades
- P_R = proportion of recreation vehicles in the traffic stream (decimal)
- E_R = PCEs of recreational vehicles in the traffic stream

with E_T and E_R read off Exhibit 15-19. For general road segments they are read off Exhibit 15-18.

The boundary values of the various LOSs are as shown in Table 20.7.

20.6.5 Passing lanes on two-lane roads

The operation of two-lane roads can be improved by the addition of auxiliary lanes, specifically climbing lanes (also referred to as crawler lanes, truck lanes and, totally incorrectly, as passing lanes) and passing lanes. The two forms of auxiliary lanes have similar appearances comprising entry and exit tapers and a length of parallel lane. The parallel lane should have a width equal to that of the main line lane, particularly in the case of the passing lane. Because of the low speeds anticipated on them, the width of climbing lanes can be reduced if forced by adverse topography. The width of the adjacent shoulder can also be reduced if necessary. These reductions must, obviously, be borne in mind in the analysis. The functions of the two auxiliary lanes are totally different insofar as the climbing lane seeks to match the LOS of the grade with a gradient usually steeper than 3 per cent to that of the level sections of the road whereas a passing lane can improve the capacity of the two-lane road.

Passing lanes are normally constructed on level sections of the road because construction costs are likely to be lower on level as opposed to hilly sections of the road. Furthermore, a passing lane on a level grade allows for higher acceleration rates as a precursor to passing a slower moving vehicle than does a climbing lane. The 2 + 1 cross-section has a capacity higher than that of the conventional two-lane road and, in fact, a four-lane road could be considered to be a two-lane road with continuous passing lanes in both directions.

Passing lanes serve to disperse the platoons built up behind slow-moving vehicles. The effect of a passing lane thus extends for a significant distance downstream of its end, specifically to the point where the length of the platoon is once again similar to its length upstream of the passing lane.

To evaluate the effect of the passing lane, it is necessary to analyse the segment without a passing lane. The segment must then be divided into four subsegments:

- The length upstream of the passing lane, L_u
- The length of the passing lane, L_{pl}
- The length downstream of the passing lane but within its effective length, L_e
- The length downstream of the passing lane beyond its effective length, L_d

The length of the passing lane must be sufficient to disperse the entire platoon that has built up behind the slow-moving vehicle and it follows therefore that the optimum length of the passing lane is a function of the directional demand flow rate as shown in Figure 20.4.

As a general rule, the PTSF across the length of the passing lane is about 60 per cent of that upstream of the passing lane. Thereafter it increases again until, at the end of the effective length, L_e, it has reverted to the original value. The effective length can be read off Figure 20.5.

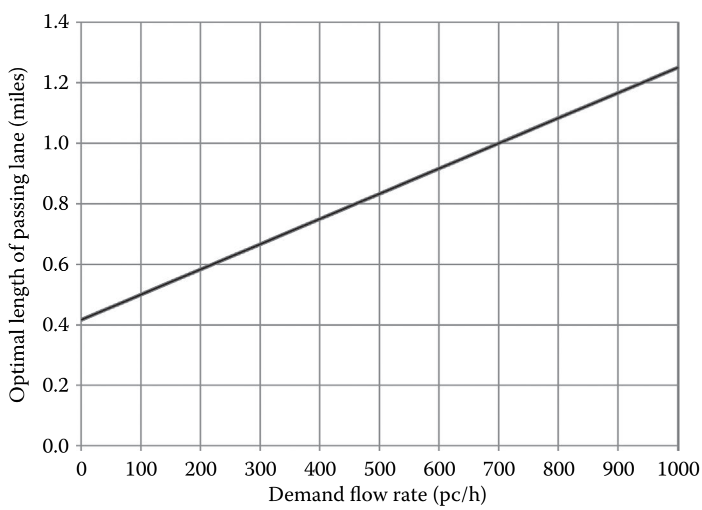


Figure 20.4 Optimum length of passing lane.

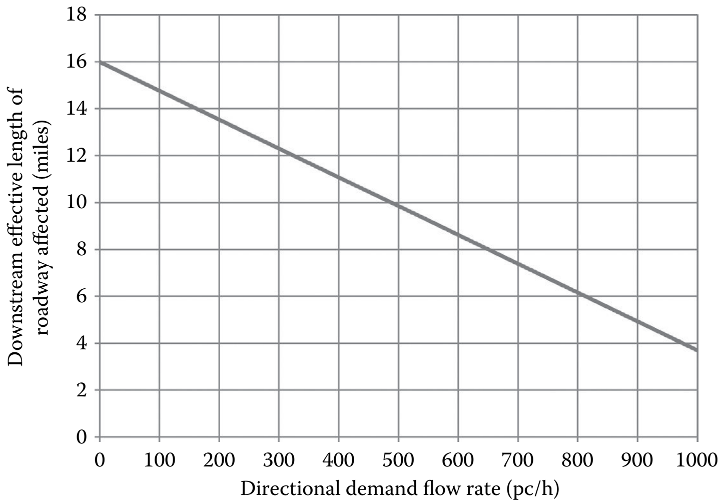


Figure 20.5 Downstream effective length, L_e, of road segment.

20.6.6 LOS for cyclists

The quality of service to cyclists is always poor on gravel or earth roads. Because of the width of these roads, cyclists are continuously at risk of being hit by a passing vehicle in addition to being covered in dust or pelted by flying stones flicked up by the tyres of the vehicle. On surfaced roads the quality of service is better but, if the shoulders are gravelled, their riding quality is usually such that cyclists are still in the position of having to share the travelled way with passing vehicles. The quality of service to cyclists on two-lane roadways with surfaced shoulder is on a par with that on multi-lane roads. In short, the extent of separation between the cyclists and passing motor vehicles is a major determinant of the LOS experienced by them. Furthermore, the blast of wind generated by a vehicle passing cyclists at high speed is considerable and can cause them to feel very insecure.

20.6.6.1 Analysis of LOS for cyclists

The analysis of LOS for cyclists is identical to that for multi-lane highways because cyclists stay as far as possible to the right of the travelled way and lanes beyond that closest to them are a matter of indifference.

The LOS for cyclists is a function of the characteristics of the traffic in the outside travelled lane, the effective width of this lane and the condition of its surface. The traffic characteristics of interest are

- The directional demand flow
- The speed of the traffic in the outside lane
- The percentage of trucks in the traffic stream

The directional demand derives from the design speed and the assumed directional split between the two flows on the road. The percentage of trucks in the traffic stream is the percentage assumed by the designer for the traffic composition in the design hour. The effective width of the lane is a combination of the actual widths of the lane and the shoulder because, if the shoulder is rideable by cyclists, they would use it in preference to sharing the outside lane with vehicles possibly travelling at high speeds.

The effective width is calculated from

$$W_v = W_{OL} + W_s$$

or

$$W_v = (W_{OL} + W_s) \times (2 - 0.005V_{OL})$$

where
W_v = effective width as a function of traffic volume (feet)
W_{OL} = width of outside lane (feet)
W_s = paved width of shoulder (feet)
V_{OL} = hourly directional volume (vehicles per hour)

In HCM 2010, Chapter 15, a correction is also offered with respect to the percentage of the road having on-street parking that is occupied. However, with respect to on-street parking where the shoulder width is less than 4 feet, it would appear, if the formula offered is to be believed, that the presence of on-street parking actually increases the effective width of the road! The effects of on-street parking on the effective width of the roadway in the cases of a shoulder width of greater than or equal to 8 feet and a shoulder width of between 4 and 8 feet seem, based on engineering judgment, to be too generous. These corrections have thus not been shown in the equations for effective width.

The impact of the speed of the traffic in the outside lane is reflected by the relationship

$$S_t = 1.1199\ \ln(S_p - 20) + 0.8103$$

where
S_t = effective speed factor
S_p = posted speed limit

In the absence of a posted speed limit, the design speed should be used as the input into the above equation.

With the calculation of the effective width of the outside lane and the effective speed factor in combination with the assumptions of the directional demand flow rate, the percentage of heavy vehicles (expressed as a decimal fraction) and the assessment of the pavement condition in terms of the FHWA 5-point surface condition rating, the LOS for cyclists is calculated from the relationship

$$\text{LOS} = 0.507\ln(v_{OL}) + 0.1999S_t(1 + 10.38\text{HV})^2 + 7.0661\,\frac{1}{P}^{\,2} - 0.005W_e^2 + 0.057$$

where
v_{OL} = directional demand flow in the outside lane (vehicle per hour)
HV = percentage of heavy vehicles (decimal)
P = FHWA pavement condition rating
W_e = effective width of the outside lane

The FHWA pavement rating is a 5-point scale:

1. Very poor
2. Poor
3. Fair
4. Good
5. Very good

The calculated LOS score is compared to the LOS lower boundary values shown in Table 20.7 to derive the LOS achieved.

Chapter 21

Transportation planning

CONTENTS

21.1 INTRODUCTION

Planning is an extremely complex discipline, arguably the most difficult of all those relating to transportation. It addresses sustainability of the transport plan through integrated policies. Like geometric design, it has shifted from a focus on moving traffic to being a multi-disciplinary activity. In this chapter, the focus is, however, on the mechanisms of technical analysis of transportation planning rather than on its broader philosophies. This is to ensure that the geometric designer will, at least, have some understanding of the outputs of the transportation planning process and how they impact on project design.

In the introduction to Chapter 4, reference was made to the fact that geometric design historically was focussed on the twin functions of mobility and accessibility, which were perceived as a continuum ranging from pure mobility and a total control of access as typified by the freeway to the pure accessibility of the local cul de sac. Little or no attention was paid to the social aspects of the road network in the sense of the impact that it had on the lives of the people served by it, and the physical environment was also not treated with much sympathy. However, the 'Thinking Beyond the Pavement' Workshop convened in 1998 by the Maryland State Highway Administration and supported by the Federal Highways Administration and the American Association of State Highway and Transportation Officials heralded a new approach to all aspects of road planning and design.

21.2 THE CONTEXT-SENSITIVE SOLUTION APPROACH TO PLANNING AND DESIGN

Context-sensitive solutions (CSSs) are seen as part of all phases of program delivery comprising long-range transportation planning, programming, environmental studies, design, construction, operations and maintenance. It considers the total context within which a transportation improvement project will exist. It thus requires that a broad view be adopted that relates the transportation system to local cultural and historical values. The intention with CSS is to plan, design and develop a transportation facility that addresses the needs of the community it seeks to serve. Furthermore, the planning and design of the facility should be such that the landscape would appear to be the poorer without it. It must therefore fit its physical setting, simultaneously preserving scenic, aesthetic and environmental resources without sacrificing the safety of the road users and local inhabitants and maintaining the mobility of traffic using the facility.

21.3 MODELLING

21.3.1 Introduction

Transportation planning is basically a modelling process in the realm of town and regional planning. It seeks to define and link the transportation needs that are generated by the residential structure, employment and services of an urban area and to develop integrated transport policies. The transport plan thus seeks integration with land use planning, the environment and between the various modes of transport. The need for transport planning was first realised when the population of passenger cars started to grow exponentially. In its earliest manifestation, planning simply comprised building more roads in an effort to match or even eliminate congestion. It was ultimately realised that this would never be effective because urban areas are not just places where people live and work but are part of a complex and usually expanding regional system. Proper long-term planning is essential if extreme congestion is not to bring traffic in urban areas to a complete standstill.

Most transportation planners used the rational model of problem solving. This model comprises eight distinct stages:

- Definition of the goals and objectives
- Problem identification
- Generation of alternative solutions to the problem
- Evaluation of the alternatives
- Selection of the preferred alternative

- Development of the selected alternative into a project plan
- Implementation of the project
- Evaluation of the outcome of its implementation

Various models used in transportation planning are described in this section.

21.3.2 The Lowry model

One of the first land use/transportation models was the Lowry model, which saw the light of day in 1964 (Rodrigues, 2013). It was based on the assumption that urban growth is a function of goods and services intended partly for consumption internally in the urban area but principally for export beyond the boundaries of the planned area. It acknowledges that transportation is a derived demand based on the fulfilment of other demands. It thus seeks to quantify these other demands, the most fundamental of which is employment, and then model the transportation needs that this develops.

The Lowry model postulated three sectors:

- The basic sector, which was employment aimed at export demand. This results in a contraflow of wealth into the urban area, thus generating growth and surpluses available for improvement in the quality of life in the broadest sense of the concept. Export principally involves industrial effort but also includes the provision of services to the external area. In general, this sector is not sensitive to location as its market is outside the area of interest.
- The retail sector, which is employment aimed at providing services to match the local demand. As such, location is important. The retail sector accounts for retailing, food and professional and civic services.
- The residential sector, which is the local residents and this is related to the job market in the basic and retail sectors. Generally, people choose to live in areas that are convenient to their chosen place of work.

The Lowry model assumes that total employment is a function of employment in the basic sector. Employment in the retail sector is thus seen as the outcome of a multiplier effect on the basic sector. Like the basic sector, the residential sector has an impact on land use and this is determined on the basis of the multiplier effect of basic and retail employment. Each job is linked to a number of people. Where people elect to live in relation to their choice of employment is a function of the cost of getting to their place of work. This cost is partly financial but also includes values attached to time and convenience and the preparedness and ability to pay for living in an aesthetic and pleasant environment.

The model seeks to represent the spatial distribution of basic and retail employment and residential location by zones. It divides the urban area into zones and derives a set of transport costs between them, thereafter calculating the distribution of the population and employment by zone.

21.3.3 The urban transport modelling system

In terms of the rational modelling described previously, modelling follows various phases. In all of these, the planner is required to

- Consider environmental issues such as air quality and noise
- Consider planning within the limitation of fiscal constraints

- Seek to involve the authorities concerned with various aspects of the urban environment
- Actively involve the general public in the execution of the planning process

The first phase seeks to define the issues that the urban area faces and the goals and objectives that can be selected to address these issues. Data acquisition is a major and very costly component of this phase and includes addressing a large variety of regional characteristics such as demographics, employment opportunities and levels of unemployment, levels of income, and so forth. Broadly speaking, the data collected refer to the transportation system itself and to the land use that triggers the utilisation of the transport system. Data acquisition is an ongoing exercise because it is necessary not only to formulate the problem and prepare alternative solutions but also to evaluate the alternative solutions and finally to monitor the success or otherwise of the outcome of implementation of the selected alternative. Adequate data acquisition is essential to ensure that the model used in the second phase is robust, accurate and complete.

The second phase addresses technical analysis. The model used in the United States is the urban transportation modelling system, also known as the four-step process. These four steps are

- Trip generation
- Trip distribution
- Modal choice
- Route assignment

These steps are discussed in more detail in the sections that follow.

While the complexity of the modelling system makes it difficult to understand without considerable education and training in the field of mathematics, specifically matrix manipulation, two of the outcomes of the analysis are traffic flows and speeds on each link in the road network. The road network supports transport by various modes including private passenger cars, public transport by bus, minibus taxis or light rail, cycling and walking. The geometric designer must be able to produce a road network that supports the flows and speeds of all these modes of transport in a safe, convenient and economical way with a minimum of side effects.

There are obviously other outcomes as well, including the volume of traffic on modes of transport other than the road network. One of these outcomes relates to the environmental impact of the transportation system, which is generally negative in the sense of exhaust emissions, noise and visual intrusion. It may also be negative in the severing of a once cohesive community into alienated portions by the inappropriate location of a major arterial, or through traffic invading a quiet residential area to escape the congestion on the logical commuter route.

21.4 TRIP GENERATION

People engage in many activities, and several of these result in trips. Trips are understood to be person trips rather than vehicle trips and are separated into categories in terms of their origin and purpose. All trips have two ends, an origin (the generator) and a destination (the attractor), and one of these is usually the home. The most pervasive trip purpose is the daily commute. This is because these trips occur typically in peak hours and have a longer distance or travel time than the others, although they do not necessarily constitute the majority of trips made from any home. They typically define the ‘worst case’ scenario in terms of traffic volumes to be accommodated by the road network and hence its various features such as the cross-section and forms of traffic control. Other trips include transporting children to school or after-school extracurricular activities, shopping, recreation, church attendance or visiting friends. The classifications generally employed are thus the home-based work trip, the home-based–other and non–home-based.

To acquire data on trips, the urban area is divided into small homogeneous units sometimes referred to as traffic analysis zones (TAZs). Homogeneity is based on family size, type of dwelling (i.e. high-rise apartment or single dwelling, drawing a distinction between large and expensive dwellings and more modest levels of accommodation) and income level. Questionnaires are usually sent to all homes in the TAZ and a sample of homes is visited for calibration of the dispatched questionnaires. Much of the information required can be extracted from census returns but the full extent of required information includes:

- The number of employed people at each home
- Nature of employment of each
- Address of workplace
- Individual income levels
- The number of people living in the home split between adults and school goers
- The number of private vehicles available for use by these people
- Average number of home-based work, home-based–other and non–home-based trips per day for all residents
- Preferred choice of mode of transport

This information is entered into a database and leads to the derivation, typically by statistical regression, on a number of variables, of a 'typical' family located at the centroid of the TAZ and the number of families resident in the TAZ.

21.5 TRAFFIC ATTRACTION

Many ways of determining the number of trips attracted to a zone are possible. At one stage, the attractiveness of a shopping centre could be determined with fair accuracy by a knowledge of the number of people it employed. With time, shopping patterns changed and the supermarket with its check-out tills replaced the corner shop where the goods were displayed on shelves behind the counter. The number of people employed dropped substantially and newer criteria have subsequently been adopted.

Trips to a shopping centre can now be determined on a macroscopic level by considering the centre as a whole and account taken of features such as the total floor area, the number of stores in the centre and the number of parking spaces provided. At a microscopic level, the trip attraction rates of individual stores could be aggregated to provide the number of trips the centre would attract (Kikuchi et al., 2004).

Shopping centres are also traffic generators insofar as the shopping sooner or later has to end and the customers return home. Furthermore, shopping centres attract traffic through the delivery of goods for sale to customers. Trucks arrive at the shopping centres, discharge their loads and then return to their points or origin.

The traffic attraction of zones not normally accessed by the general public such as factories could be assessed by knowledge of the number of employees on the payroll. Manufacturing concerns also attract traffic through the delivery of raw materials to them and generate traffic through the shipping out of manufactured articles. Raw materials would typically have their origins outside the urban area and include mines, farms and other manufacturers such as producers of components.

21.6 MODAL SPLIT

The modes on offer generally are

- Private passenger cars
- Buses
- Light rail
- Animal-drawn vehicles
- Walking
- Cycling

Of these, in developed countries, walking and cycling are over relatively short distances and are often limited to remaining within the bounds of a single TAZ.

In developing countries, the possession of cars is a luxury available only to relatively wealthy people who form a small minority of the population. Public transport is often by poorly maintained and frequently unroadworthy buses. The road networks are inadequate and, where roads do exist, they are unsurfaced and in a poor condition. Bus operators are reluctant to entrust their vehicles to the remoter links in the network. In consequence, cycling and walking are the only modes available even for trips over extended distances.

The majority of trips in urban areas are by still by passenger cars. It is, however, neither possible nor even desirable to continuously upgrade the road network to resolve the problem of congestion created by increasing levels of car ownership. There is thus an increasing focus on retaining the accessibility of destinations by promoting the use of other modes. Trip assignment is on the basis of the modes available and it is reasonable to assume that, with time, more modes will become available to the travelling public.

Trips may involve more than one mode and it can be anticipated that multimodal trips will become more popular in future. Park-and-ride facilities are being provided by transportation authorities in increasing numbers and these make provision for commuters to drive their cars to a point on a bus route, park there and catch a bus to their destinations. The alternative is the drop-off facility, also known as kiss-and-ride as a form of differentiation from the park-and-ride facility. The issue of public transport is discussed in Chapter 15.

Modal split is dependent on personal choices as modified by the modes available at the origin of any trip. Choices may be based on

- Convenience
- Economics
- Perceptions regarding the value of time

The passenger car is extremely convenient because it is available right at the point at which a trip is initiated; it offers a wide variety of route choices; and it is not restricted to a timetable of travel times. Its convenience is often reduced by an inability or difficulty in finding parking in the vicinity of the destination.

Economics may be related to the actual cost of the trip but, equally, could be simply a perception of cost, with the perception applying particularly to the cost of transport by passenger car. Car owners or users typically have little appreciation of the actual cost of transport by passenger car. The cost of filling the petrol tank may cause a transient pain but is soon forgotten. In any event, it is usually no more than about 25 per cent of the total cost of car ownership. The total cost of car ownership includes

- The original cost of purchase or, more accurately, the annual depreciation of the vehicle
- Routine servicing and repairs
- Garaging
- Insurance
- Running costs, of which fuel costs are a part

The value of time is discussed in some depth in Chapter 21. In short, if a trip takes a significantly longer period of time by bus than by passenger car, it is possible that the traveller may decide that his or her time could be better spent than by sitting on a bus. It is believed that income level is a major driver of choice. At low income levels, people may not have the means to purchase even the most modest of passenger cars and they have no choice but to walk for short trips or to use public transport for longer trips. They are typically described in the literature as being captive to public transport.

Statistical models have been developed to determine the choices travellers make regarding their modal preferences where choice is, in fact, possible. Disaggregate models based on individual preferences, which stem from economics and from psychology, have seen the light of day. The most commonly used of these models is the multinomial logit model which basically seeks a regression between a single dependent variable and multiple independent variables.

The prime function of the analysis of modal split is to convert person trips to vehicle trips. A passenger car may result in one passenger car trip being the equivalent of 1.5 person trips whereas a bus trip may be the equivalent of anything from 10 to 80 person trips. The application of modal analysis will thus provide the geometric designer with information on the number of vehicles that the road network must be able to accommodate. It does not, however, provide any indication of which element of the road network would be occupied by these vehicles. This function is addressed by the process of route assignment.

21.7 ROUTE ASSIGNMENT

Route assignment is the process whereby vehicle trips are assigned to various links in the network. It is not likely that there would be only one possible route between two TAZs, and assignment of the traffic using the various routes is based on the achievement of an equilibrium state. Regarding equilibrium, Wardrop (1952) postulated two principles:

- The journey times on all routes actually used between a common origin and destination are equal and less than those that would be experienced by a single vehicle on any unused route. This principle states that each road user would seek to minimise his or her cost of transportation.
- At equilibrium, the average journey time is minimum, implying that each road user cooperates with all others in choosing his or her route to ensure the most efficient use of the system as a whole.

The equilibrium assignment can be determined by application of the Frank–Wolfe algorithm, which has been built into the widely used Emme transportation planning software (at time of this writing Emme 4). For each link in a highway network, there is a function describing the relationship between resistance and volume of traffic. This could take the form determined by the Bureau of Public Roads.

$$S_a(v_a) = t_a \left(1 + 0.15 \left(\frac{v_a}{c_a}\right)^4\right)$$

where

$S_a(v_a)$ = average travel time for a vehicle on link a
v_a = volume of traffic on link a
t_a = free-flow travel time on link a per unit of time
c = capacity of link a

The volume of traffic on each link of the road network can be determined by means of a series of iterations commencing with all-or-nothing assignments of traffic. These suggest that all the traffic on a desired line between an origin, TAZ_i, and a destination, TAZ_j, is allocated to a single link. It may then be found that the capacity of this link has been exceeded. The difference between capacity and the traffic hypothesised for the link is moved to a different link between the two terminals. The travel time on these two links is assessed and adjusted until parity of travel time between them is reached. A complication arises insofar as a particular link may appear on routes between other origins and destinations. Some of the traffic originally located and subsequently modified on the link will have to be relocated to a different link to accommodate this new source of traffic.

21.8 THE MATCHING OF GENERATION AND ATTRACTION

Having stated that every trip has an origin and a destination, it follows that the sum of traffic emanating from all generators should equal that arriving at all the attractors. Unfortunately, this will almost certainly prove not to be the case and there will be a mismatch between the two numbers.

At a closer level of detail, all the trips originating from one TAZ and terminating at another should equal those attracted to the other TAZ from the first. In short,

$$TG_{ij} - TA_{ji} = 0$$

where

TG_{ij} = trips generated at TAZ i and going to TAZ j
TA_{ji} = trips attracted to TAZ j from TAZ i

It must be realised that in a large metropolitan area, which would be the region most likely to be in need of a transportation study, the number of TAZs defining the study could run into the high hundreds or even thousands. Matching the trips' ends is thus a laborious task but simply splitting the difference between the numbers of generated and attracted trips between two TAZs would probably be good enough. Alternatively, the mismatch can be eliminated by the application of highly sophisticated hill-climbing algorithms although this is not really a necessity. Given the number of estimates that have to be made in reconciling the trip ends, excessive computation does not necessarily result in an increase in accuracy.

From the point of view of geometric design, it is not necessary for the transportation study to be all that accurate anyway. Bearing in mind that transportation studies are more concerned with volumes than with speeds it is pointed out that the addition of a lane implies a jump of about 2000 vehicles per hour in capacity.

21.9 TRAFFIC IMPACT ASSESSMENTS

21.9.1 Introduction

The geometric designer will have little involvement in the transportation analysis described previously but is a recipient of the information it provides. However, it may be necessary for the designer to extend the information available to answer the question 'What if …?' This may be to determine the effect on the traffic conditions on the surrounding road network of a new development such as a new township, or a shopping centre, school or hospital, all of which are noted attracters and generators of traffic. A traffic impact assessment (TIA) must be carried out.

21.9.2 Overview of a TIA

The functions of a TIA (Stantec Consulting Limited, 2005) are to

- Forecast the traffic impacts created by a proposed development
- Establish which improvements to the local road network are required to accommodate a proposed development
- Relate decisions regarding changes in land use to traffic conditions both at present and in the future
- Evaluate the number and location of access points to a development and make recommendations regarding their design
- Provide traffic projections to the authorities charged with responsibility for ensuring that the regional transport plan remains current
- Identify network improvements required to accommodate the proposed development
- Provide a basis for determining the developer's financial responsibility for specific off-site improvements

In essence, a TIA is a mini-transportation planning exercise. It does not normally use the highly sophisticated software required for region-wide planning and is usually reliant on manual calculation.

A distinction is drawn between a traffic impact statement (TIS) and a traffic impact assessment (TIA). The only difference between them is the level of detail required. Information that should be provided in both (Wepener et al., 1995) includes

- The name and brief description of the development in terms of the proposed land use
- Current operational conditions, such as traffic volumes and speeds, on the surrounding road network
- Analysis of operation of the accesses to the development
- Analysis of the operation of the first intersections upstream and downstream of the proposed accesses to the development
- Professional input on the expected traffic impact of the proposed development

The required analysis should include analysis of

- Capacity
- Need for signalisation
- Need for street lighting in the vicinity of the accesses to the development
- Pedestrian movements

21.9.3 Extent of a TIA

A TIA would normally be required where the development could generate more than 150 peak hour trips. If between 50 and 150 peak hour trips are likely to be generated, a TIS would suffice. If fewer than 50 trips are generated, the need for a study falls away except if the surrounding network is already running at or beyond capacity. Even if the above conditions are not met, the responsible transport authority may elect to have a TIS for other reasons such as response to political pressure or long-term economic growth plans.

The areal extent of a TIA may be fairly restricted as traffic disperses from the development into the surrounding network and eventually reaches a level where no further action is required to maintain the integrity of the network. The accuracy of the trip assignment process also decreases as the distance from the development increases. International tendencies are to terminate study areas at a specified distance from the development and, in South Africa, for example, it has been found that study areas seldom extend more than 1.5 km away from the development.

21.9.4 The estimation of trip generation

Trip generation can be assessed by one of four methods:

- Analytical – involving estimation of
  - The number of people likely to use the facility
  - Their arrival times to and departure times from the development
  - Selection of mode of travel, that is, modal split
  - The occupancy of the vehicles involved
- Trip generation tables, such as those published in the ITE trip generation document (Institute of Transportation Engineers, 2012) which provides trip generation information on the basis of 170 different land uses
- Trip rate formulae, where the trip generation rate is a function of the size of the development
- Trip rate algorithms using complex land use/transportation models as discussed in the previous sections of this chapter

It was previously assumed that all trips to a development were new. It has, however, now been shown that many of these trips are already on the network so that they don't change the extent of traffic generated by the development. What could change is the number of vehicles on the links closest to the development. There are thus

- Primary trips – defined as single-purpose trips, that is, home – development – home
- Nonprimary trips – multipurpose trips that call in at the development on the way to another destination

Nonprimary trips are frequently work – development – home trips or in the opposite direction and may be divided into diverted and pass-by trips. Pass-by trips are those that do not divert significantly from the route they would follow anyway between work and home to access the development. Pass-by trips thus have to be subtracted from the trips generated by the development in the TIA as they do not change the operation of the road network as result of its provision.

21.9.5 Assessment of traffic impact

The assessment of impact can be done only once the study area has been defined, growth rates defined and trips generated by the development have been calculated and assigned.

Calculation is in terms of the *Highway Capacity Manual* (Transportation Research Board, 2010) and based on the concept of level of service (LOS). This concept uses qualitative measures such as speed, volume of traffic, delays and freedom to manoeuvre to characterise conditions within a traffic stream. Six levels are defined from LOS A to LOS F, with LOS A representing the most favourable conditions and LOS F unstable or stop-and-go conditions.

In urban areas, LOS C or better are normally considered desirable and LOS D is acceptable. LOS E is defined as the capacity of the road and LOS F constitutes unstable flow. Both of these two LOSs are considered undesirable as design inputs.

21.10 ACCESS ARRANGEMENTS

21.10.1 Introduction

Ultimately the traffic impact of a development is determined by the number of accesses provided and their location relative to the surrounding road network.

Obviously, the impact is at its largest at the intersection(s) between the development and the surrounding network. Thereafter, the impact declines as the traffic to or from the development disperses across the network until a point is reached whereby differences in traffic volumes caused by the development are no longer discernible. In addition to the number and location of accesses, factors that need to be considered include the size of the accesses, the form of traffic control applied to them and their geometric layout.

These considerations are discussed in this section.

21.10.2 Access management

If a development is of any size, it is almost always preferable to have more than one access. It is then important to ensure, as discussed in Chapter 10, that the functional areas of these intersections do not overlap. In the case of an overlap occurring, the weaving between vehicles seeking entry to the development and those departing from the development could cause a total breakdown of traffic operation between the two intersections. This would be disadvantageous both to the local street network and to the commercial viability of the development if it were to be a retail centre as opposed to any other form of development.

Too close spacing of intersections could cause the linkage between a new residential area and the existing road network to be chaotic at peak hours. Access from a residential area, particularly if it is large, would usually be to a high-order road such as an urban arterial via signalised intersections. For these roads to function optimally, for preference under ‘green wave’ conditions, the spacing of successive intersections should be at intervals of the order of 500 to 600 metres or more dependent on traffic speeds permitted on the through road.

The number and size of accesses to be provided to a development are dependent on

- The number of trips generated or attracted by the development per time interval
- The size of the development
- The functional classification of the roads adjacent to the development
- The minimum access requirements
- Topographic restraints

21.10.3 Capacity of accesses

The capacity of an access is determined by the type of control applied to the intersection.

At signalised intersections, vehicles on the through road are brought to a halt periodically in favour of vehicles entering from side roads. It follows that the capacity of the intersection in terms of through traffic will be less than that of the road upstream of the intersection. If that road is running at a poor LOS, that is, a heavy traffic flow, the loss of green time at the intersection could cause it to become oversaturated with a growing queue of vehicles waiting to get through the intersection. This could back up to the preceding intersection and ultimately bring the local network to a complete standstill.

At priority controlled intersections, vehicles exiting from the development could be disadvantaged by inadequate gaps of a usable size in the traffic stream on the through road. A considerable queue could thus be generated within the confines of the development. The question could be asked: If the volume of traffic generated by or attracted to the development is so small that priority control will suffice, is it necessary to carry out a TIA? As a general rule, 150 trips per hour to and from the development would be sufficient to require a TIA. At between 50 and 150 trips per hour, a TIS would suffice and, at fewer than 50 trips per hour, it would not be necessary to have any form of traffic study except if the surrounding network is operating at or near capacity.

Roundabouts could resolve the problem of inadequate gaps as they have a capacity higher than that of priority controlled intersections.

The number of lanes provided also has a bearing on the capacity of the intersection. Reference is specifically to auxiliary lanes that could be aimed at removing turning traffic from the stream traffic, thus allowing them to move through the intersection area with a minimum of impedance caused by turning vehicles. Auxiliary through lanes would support matching the capacity and LOS of the intersection to the upstream uninterrupted flow of traffic.

21.10.4 Special cases of access control

Service stations do not normally require TIAs even though they are often located on roads with high traffic volumes. In fact, they do not generate traffic at all because the traffic drawn to them is already on the road network. All they do is displace a small amount of the traffic in the through stream through a time lapse of 5 to 10 minutes on average.

Would-be service station owners tend to favour sites at existing intersections because this enables them to capture trade from both roads. Unfortunately for them, these are not favoured by the road authorities because they generate turning movements in addition to those already in the intersection area. In effect, a service station converts a four-legged intersection into a multi-leg intersection with all the conflict points that this suggests. A further problem with corner sites is that motorists sometimes use the service station forecourt as a short cut bypassing the intersection, hence creating still more conflict points.

If a development is likely to attract a high volume of heavy goods vehicles, the desirability of having a separate service access should be considered. Because the dimensions and operating characteristics of trucks differ from those of passenger cars, they impose a significant burden on the efficiency of intersections. This is normally quantified by converting trucks into a number of equivalent passenger car units, suggesting that one truck would take the place of possibly several passenger cars. The layout of the development could be such that access by trucks to it could perhaps be via a minor street where the disruption caused by them would be reduced.

Chapter 22

Coordinate calculation

CONTENTS

22.1 INTRODUCTION

The designer works on survey plans in terms of a scale appropriate to the level of detail required. Route location could be on topographical sheets typically to a scale of 1:50,000 or, in the United Kingdom, 1 inch to the mile, which corresponds to a scale of 1:63,360. The horizontal alignment within the corridor defined during route location is normally done on survey plans to a scale of 1:10,000. Detailed design of features such as intersections could be to a larger scale again, such as 1:1000 or 1:5000. The fact of the matter is that these are all lines on paper. They need to be converted to lines on the ground. This is done by engineering surveyors who require information defining the road centreline and the various detail matters such as kerb lines at intersections. This information is provided in the form of coordinates. Armed with this information, the lines on paper can be set out to enable the construction of the road.

In this chapter, the calculation of coordinates is discussed.

22.2 THE NATURE OF COORDINATES

Any point on the earth's surface can be defined by three numbers: the x, y and z coordinates. The x value is given in the east–west direction and the y coordinate in the north–south direction. The z coordinate refers to height. They are all distances measured from some or other datum point which would have the value of zero for the x and y coordinates. This is very often the case for small local surveys where there isn't a need to tie them to the national grid. The datum for height could be a survey peg or beacon with a height arbitrarily set at zero or 100 metres or any other convenient level. For larger surveys, the datum in respect of height could be either mean sea level (MSL) or low water ordinary spring tide (LWOST).

Alternatively, the x and y coordinates may be measured in degrees (in Europe, gons, where there are 100 gons in a right angle). Reference then is to degrees longitude (Lo) and degrees latitude (Lat or Φ – phi). Longitude refers to the north–south or y direction and latitude to the east–west or x direction.

The equator is the datum for latitude, and lines of latitude of interest are the Tropics of Cancer and Capricorn. The word 'Tropic' stems from the Greek *tropos* meaning 'to turn back', which is what the sun apparently does at the summer and winter solstices. The Tropic of Cancer is the most northerly circle of latitude at which the sun may appear to be directly overhead at its zenith and lies at 23° 26′ 15.14″ north of the equator. This is marked by the longest days and shortest nights in the Northern Hemisphere because the Northern Hemisphere is tilted towards the sun to its maximum extent. In the Southern Hemisphere this is, of course, the winter solstice, when the sun is at its nadir or lowest, the days are at their shortest and the nights are longest. The summer solstice occurs on about 21 June for the Northern Hemisphere and 21 December for the Southern Hemisphere.

The other lines of latitude that are normally shown on maps are the Arctic and Antarctic Circles at 90° – 23° 26′ 21″ (approximately) or 66° 33′ 39″ north and south respectively.

The position of Lo 0° 0′ 0″ has, unlike latitude, no astronomical justification and has changed with time. In Grecian times, Lo 0° was located to the west of the western tip of North Africa because the concept of negative numbers had yet to be invented. The French, of course, had Lo 0° passing through Paris. Finally, at the International Meridian Conference held in Washington in October 1884, it was agreed that the Greenwich Meridian would be the international datum for longitude and measurements of time although France continued to use the Paris Meridian until 1911. Measurement of longitude runs from +180° east to –180° west and not from 0° to 360°.

Longitude and latitude are not particularly convenient for the geometrical calculation of directions and distances, and Cartesian coordinates are the preferred option. However, a coordinate of latitude measured from the equator to the north of Scotland or Canada would be an unwieldy number to use in calculation. Similarly, a coordinate of longitude measured in metres from Greenwich to Shanghai would be equally cumbersome. The problem is resolved by a translation of the datum point. For example, calculations carried out in Cape Town are vastly simplified by using, say, Lo 19°, Lat 33° as the origin of the Cartesian grid. The distances between the Lo 19° and Greenwich and Lat 33° and the equator could then simply be listed on the calculation sheet as constants if absolute coordinates are required.

22.3 ROUTE CALCULATIONS

22.3.1 Introduction

There are two basic and related forms of calculation that can be applied to the definition of a route. These are the join and the polar (also known as the traverse). The join uses the coordinates of two known points to calculate the distance and bearing between them. The polar calculates the coordinates of a point from those of a known point using the distance and bearing of the ray between them. All that the designer requires is a reasonable knowledge of geometry and trigonometry and the ability to solve an equation of the form

$$X_2 = X_1 + \delta X$$

The polar thus uses X_1 and δX to calculate X_2 whereas the join uses X_1 and X_2 to calculate δX.

22.3.2 The join

In the case of a road centreline, the designer would start off by scaling the coordinates of the points of intersection (PIs) between the various tangents comprising the horizontal alignment of the route. The calculation of the join between them will show a distance to several decimal places of a millimetre and a bearing to several decimal places of a second of an arc. It is customary to round off the distance to the nearest centimetre and the bearing to the nearest 5 seconds. This is a practical indication of the normal accuracy of setting out of the works. A polar is then carried out to recalculate the precise coordinates of the new point using the rounded off values of distance and bearing as described in the following section. The reason for this repetition of calculation is to avoid 'coordinate creep' as discussed in Section 22.3.3.

This process is continued through successive PIs to the end of the project. In consequence, all the PIs of the route with the exception of the first one will be calculated and precise. If the first PI is at the intersection of the route with an existing road, its coordinates will also be precisely known.

The format of the calculation of a join is shown in Table 22.1.

22.3.3 The polar (or traverse)

The polar is used to calculate the coordinates of a defined point on the basis of the known coordinates of a previous point and the distance and bearing between them. It is typically used in the calculation of the coordinates of the PIs along the centreline of a road as

Table 22.1 The format of the join calculation

Item	*Y*	*X*
PI 1	y_1	x_1
PI 2	y_2	x_2
Difference δ	$\delta y = y_2 - y_1$	$\delta x = x_2 - x_1$
tan θ	$\delta y/\delta x$	
θ	$\tan^{-1} \delta y/\delta x$	
L	$\delta y \sin \theta$	$\delta x \cos \theta$

Note: The calculation above is between two points of intersection, PI 1 and PI 2, defining a tangent and provides the bearing and distance from PI 1 to PI 2.

discussed previously. It is also used to establish the coordinates of a point on a traverse, for example, around the boundaries of a farm or quarry. If the traverse goes back to the starting point as in the case of the farm boundary, it is known as a closed traverse.

22.4 SETTING OUT THE WORKS

22.4.1 Introduction

Setting out of the centreline of the road is by use of a theodolite and, historically, a tape or chain. Calculation involved seven figure logarithms and trigonometrical ratios as defined in the seminal work by Baron von Vega.

In 1959, the Tellurometer was invented by Dr Wadley of the South African Council for Scientific and Industrial Research (CSIR). This device operates similarly to radar by emitting a wave that is returned by the remote station. The phase shift between the original and the reflected waves gave the distance travelled between the two points. Although the instrument and the remote station have to be intervisible, being a shortwave as opposed to an optical process, it is not affected by rain, mist or darkness and is highly accurate to about 50 kilometres. Except for small local surveys, the need for taping fell away.

During the 1970s, programmable hand held calculators became available and these had built-in trigonometrical ratios. It was no longer necessary laboriously to refer to Baron von Vega's tables.

Ongoing development of the Tellurometer and ultimately of microelectronics in general led to the birth of the total station. The total station added angular measurement to distance measurement for the first time. The first total station is attributed to Zeiss in 1971. Total stations have continued to improve and have a facility to download survey information to an external data storage device.

22.4.2 Setting out of a polar

The fieldwork involved in a polar involves the following:

1. A point with known coordinates is used for setting up.
2. A back sight to the previous known point is taken.
3. The bearing between these two points is already known so that that bearing +180° is set on the theodolite to orient it.
4. The bearing to the new point is set on the theodolite.
5. The location on the ground of the new point is identified by the length of the calculated ray at the bearing between the theodolite and the new point.

The format of the calculation of the polar is shown in Table 22.2.

Table 22.2 Calculation of the polar

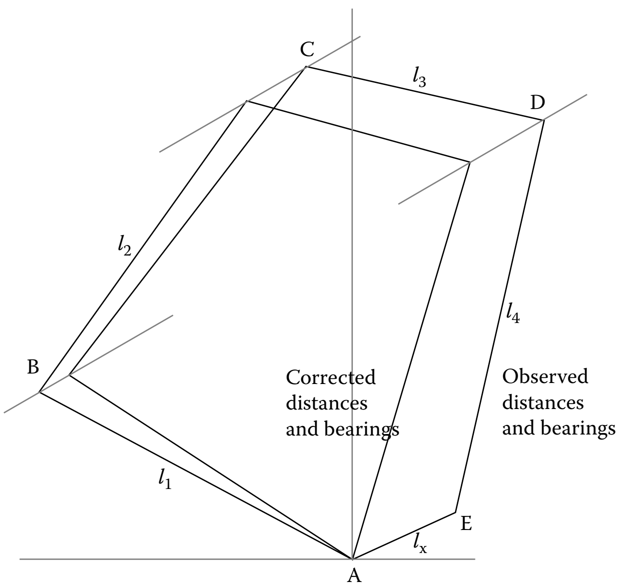
Item	*Y*	*X*	*Ray*
Coordinates of Pt 1	y_1	x_1	L
Difference	$L \sin \theta$	$L \cos \theta$	θ
Coordinates of Pt 2	$y_2 = y_1 + L \sin \theta$	$x_2 = x_1 + L \cos \theta$	

Note: The polar (also known as the traverse) is required for setting out a point, the coordinates of which are known, from a known point or beacon.

22.4.3 The Bowditch compass rule

Ideally, the coordinates of the last point of the traverse should be the same as those of the second known point or of the starting point in the case of a closed traverse. It is quite likely that errors, not to be confused with mistakes, in distance or bearing will result in the calculated coordinates of a point not matching its known values. An adjustment is required, usually in terms of the Bowditch compass rule. This states that the total linear error is proportional to the length of the traverse and that the linear error at the end of each of the successive legs of the traverse is in the same proportion as the length of the traverse to that point is to the total length of the traverse. This concept is illustrated in Figure 22.1.

The extent of the error in terms of distance and bearing between the known and the calculated coordinates on completion of the traverse is established by means of a join. A series of polars, one at each measured point, is then used to remove the error, thus ensuring agreement between the known and calculated coordinate values. The polars all use the bearing of the error and the lengths of the polars' rays are proportional to the ratio between the distance from the start of the traverse to the point of interest and the total length of traverse.



Length of traverse $l = i_1 + i_2 + i_3 + i_4$

$$\text{Correction B} = \frac{l_1}{l} \cdot l_x = \frac{97}{333} \times 32 = 9.3$$

$$C = \frac{l_1 + l_2}{l} \cdot l_x = \frac{181}{333} \times 32 = 17.4$$

$$D = \frac{l_1 + l_2 + l_3}{l} \cdot l_x = \frac{241}{333} \times 32 = 23.2$$

$$E = \frac{l_1 + l_2 + l_3 + l_4}{l} \cdot l_x = l_x = 32.0$$

Figure 22.1 The Bowditch compass rule.

22.5 SETTING OUT OF CURVES

22.5.1 Introduction

The exposition in the previous sections is of the setting out of the tangents comprising the alignment of the road. It is also necessary to set out the curves between these tangents. Horizontal curves are circular and have the parameters of radius and angle of deviation, also referred to as the deflection angle. The basic properties of the circular arc are illustrated in Figure 22.2.

The steps to be followed in the calculation of the coordinates and stake values of the various points of interest and their setting out are discussed in this section.

22.5.2 The coordinates of the BC and EC

The designer will have selected the radii for the various curves along the road. To set out each curve, its tangent length, the distance between the PI and the beginning (BC) and end of the curve (EC), must be determined from

$$T = R \tan \frac{\theta}{2}$$

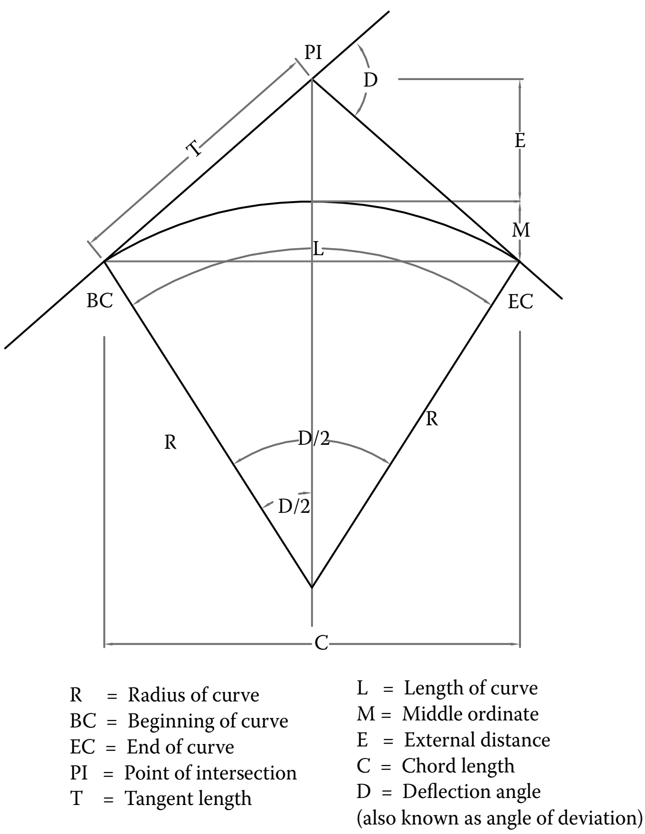


Figure 22.2 The elements of a circular arc. (From Caulfield B. *Topic 8: Horizontal curves*. Trinity College, Dublin, 2012.)

where

T = tangent length (m)
R = radius (m)
θ = angle of deviation (degrees)

The coordinates of the BC and EC are calculated by polars using the coordinate of the PI, the tangent length and the bearing from PI b to PI a in the case of the BC and PI b to PI c in the case of the EC. PI b is the PI of the curve to be set out, PI a is the preceding PI and PI c is the following PI.

With the ends of the curve fixed, the intermediate points along the curve need to be set out. A road centreline is normally defined by staking at 20-metre intervals. It seldom happens that the BC and EC fall naturally on a stake point, so there could be odd distances between the BC and EC and their closest stake points along the curve. As in the case of the adjustment of the PI coordinates for convenience, it is also possible to adjust the calculated tangent length to ensure that the BC or the EC, but seldom both, do, in fact, fall on full stake values. A calculation will show that the selected radius of, say, 2000 metres may now become 1995.326 … metres or some other unusual value but it is unlikely that anyone would object or even be aware of the change. In practice, this refinement is, in fact, not necessary.

22.5.3 Calculation of stake values

When using Imperial units, distances were measured in a variety of different units including feet, yards, or miles with rods, poles or perches all equal to 5½ yards and chains (typically a 66-foot or Gunter chain), where 10 chains are equal to a furlong and there are 8 furlongs to a mile. Three miles are equal to one league. In survey, specifically with respect to road works and railways, measurements were in chains and, even if the basic unit of measurement was the foot, reference was still made to a point along a road having a chainage of x feet, measured as a distance from the start of the project. With countries changing over to metric units, chainages were replaced by stake values (SVs) and this is the nomenclature used below.

The length of the curve is required to derive the SVs of the various points along the route. This length is calculated as

$$L = R\theta$$

with θ expressed in radians. A complete circle contains 2π radians or 360°, so the conversion of degrees to radians requires a multiplier of $\frac{2\neq}{360}$. The converse, conversion from radians to degrees, arises from $\frac{360}{2\pi} = 57.29578$, which is the number of degrees in a radian.

The distance from the start of the project, being PI 1 to the next PI, that is, its SV, is known from the joins discussed previously. The SV of BC 1 is thus the SV of PI 2 less the tangent length for curve 1. The SV of EC 1 is not determined from PI 2 and the distance to it is measured along the curve from BC 1. The SV of BC 1 plus the length of the curve gives the SV of EC 1. A new value of SV is now required for PI 2, specifically the SV of EC 1 minus the tangent length. With this new SV, the SV of PI 3 is determined from the previous join data. The process is repeated to the end of the project.

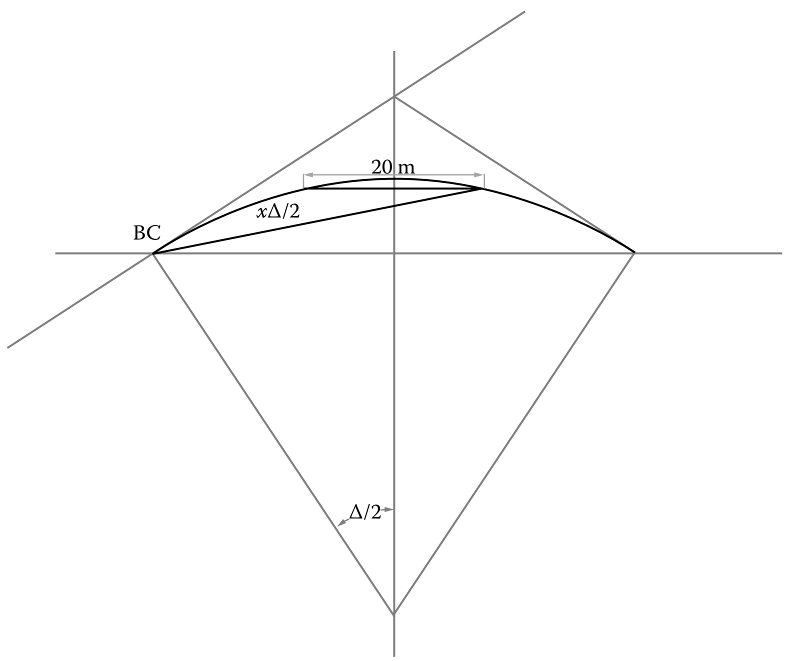
The surveyor is then able to proceed to the process of setting out the works. He or she will have already located and coordinated a number of survey beacons along the road. These are normally steel pegs embedded in concrete with the beacon number marked on the concrete. The setting out process is initiated by calculating a join between the survey beacon closest to the point to be set out and the point of interest. Setting up over the survey beacon and orientating the theodolite will enable taping in the correct direction through the calculated distance to the site of the required point which is then marked, usually with a wooden stake.

22.5.4 Setting out of intermediate points along the curve

Setting out of points along the curve can be by triangulation, which could be either calculated or physical. Triangulation is discussed in Section 22.6. In this case, calculated triangulation uses a baseline between two coordinated points, being the BC (or EC if this is more convenient) and the previous point on the curve, the angle between the base line and the line connecting the BC or EC to the unknown point and the distance from the preceding point on the curve, usually 20 metres. This is shown in Figure 22.3.

Physical triangulation is more convenient and involves taping on the ground along the curve. The line of sight from the BC or EC to the point to be set out is at an angle that is a fraction of half of the angle of deviation. This fraction is the ratio between the distance from the BC along the curve to the point of interest and the total length of the curve because a fundamental property of a circular curve is that it is a constant rate of change of bearing with length. This is illustrated in Figure 22.4.

Measurement of distance is never actually along a true curve but along a series of chords, which are relatively short in relation to the length of the curve. The error made is minuscule, being 0.3 mm over a distance of 20 metres for a radius of 1000 metres and 8 mm for a radius of 200 metres. The error is thus ignored.



x is the fraction of the deflection angle that addresses the location of the unknown point. The distance of 20 m is to the previously coordinated point.

Figure 22.3 Calculation of the coordinates of a point on a circle. (From Caulfield B. *Topic 6: Angle measurement: Intersection and resection.* Trinity College, Dublin, 2010a.)

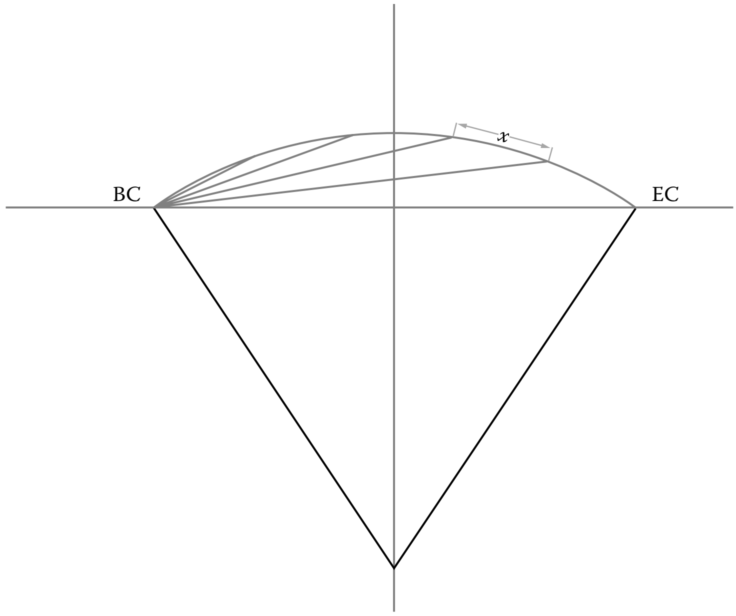


Figure 22.4 Setting out of points along a curve.

22.6 TRANSITION CURVES

22.6.1 Introduction

If a vehicle were to follow the road precisely when a tangent is followed immediately by a curve, the driver would have to turn the steering wheel instantaneously to match the changing direction. Obviously this is not possible and some slower rate of turning the steering wheel would be followed. In consequence, the route actually followed by the vehicle in changing from movement in a straight line to movement along a curve is a form of transition. If this were to be replicated along the centreline of the road, the alignment that would be provided would have to have the characteristics of

- Being tangential to the straight
- Curvature that is zero (i.e., infinite radius) on the tangent
- The radius decreasing in a linear fashion along their transition to the circular curve
- Joining the circular arc tangentially
- The radius at the end of the transition being that of the circular curve

The spiral that meets these requirements is variously referred to as the Euler spiral, the natural spiral or the clothoid in which, over some or other distance, the path of the vehicle changes from moving along a curve of infinite radius (otherwise the tangent) to moving along some lesser radius of curvature. On leaving the curve, the process is obviously reversed.

A transition curve differs from a circular curve because its main characteristic is that its radius is changing across its length. It follows that its shape is defined by a mathematical function more complex than that of a circular curve and, in consequence, it is also more difficult to set out. If the radius of the circular curve is sufficiently large, the transition path followed by the vehicle is easily contained within the width of the lane in which the vehicle

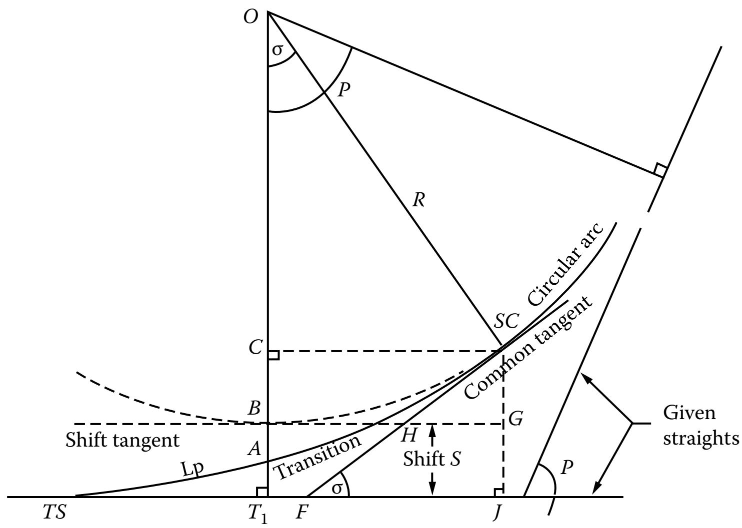


Figure 22.5 Geometry of the clothoid. (From Caulfield B. *Topic 9: Curves and transition curves*. Trinity College, Dublin, 2010b.)

is travelling. As the circular radius decreases, however, it becomes necessary to modify the shape of the road to more accurately accommodate the path being followed by the vehicle. A transition curve has to be provided.

The geometry of a clothoid is illustrated in Figure 22.5.

22.6.2 The radial forces involved

A vehicle travelling at constant speed along a curve is subjected to a centrifugal force, *F*, expressed as

$$F = \frac{mv^2}{R}$$

where
- m = the mass of the vehicle (kg)
- v = the speed of the vehicle (m/s)
- R = the radius of curvature (m)

It follows that the smaller the value of *R*, the higher is the force, *F*, acting on the vehicle and the faster the speed the force is also higher. To ensure the comfort of the occupants of the vehicle, the force should be allowed to increase gradually in a linear fashion from zero to the full value experienced on the circular arc. These requirements are expressed as

$$F \propto \frac{1}{R} \text{ and } F \propto L$$

and can be combined so that $\frac{L}{R} = K$ where K is a constant.

As discussed in Chapter 6, superelevation is used on circular curves to minimise the effects of the centrifugal force, thus allowing for higher speeds on shorter radius curves. Transition curves can be used to good effect in providing across their length for the increase in superelevation. The length of the transition curve is thus selected on the basis of the required length of the superelevation development as discussed in Chapter 6.

Two curves can be used to provide the transitional section of a curve. These are the clothoid and the cubic parabola. The clothoid is a true spiral and thus matches the requirement that RL should be a constant, K. The cubic parabola is not a true spiral because RL is not a constant. It can, however, be used over a certain range of circular curvature and has the advantage of being less complex that the clothoid.

22.6.3 The shift

In effect, the spiral curve winds in from the tangent to the circular curve so that, obviously, the circular curve has to be offset inward towards the centre of the curve from the tangent to accommodate the spiral. This is referred to as the shift of the transition curve and is calculated from

$$S = \frac{L^2}{24R}$$

where

S = shift (m)
L = length of transition curve (m)
R = radius of circular curve

The cardinal points along a compound curve comprising a circular curve bounded by two transition curves are

TS = tangent to the spiral
SC = spiral to the curve
CS = curve to the spiral
ST = spiral to the tangent

and this convention is used in Figure 22.5.

The length of the selected curve should not be so long that it creates the misleading impression of a circular curve that has a higher radius than that actually provided. Generous lengths of transition have been observed to result in an increased accident rate.

There are two bases from which the length of the transition curve can be calculated. These are to

- Establish a length of curve over which the radial force acting on the occupants of a vehicle increases from zero on the tangent to that experienced on the circular curve
- Match the length of the superelevation development

The expression used by some highway agencies for calculating the length of a spiral transition curve on the basis of passenger comfort is

$$L = \frac{0.0214V^2}{RC}$$

where
L = length of spiral (m)
V = speed (km/h)
R = curve radius (m)
C = rate of increase of lateral acceleration (m/s^3)

The factor C is an empirical value defining the level of comfort and safety levels of passengers and it has been found that above a value of 1.3 m/s^2 there is no longer any safety benefit in the use of a transition curve. In terms of comfort of the vehicle occupants, C is often given a value of 0.3 m/s^3.

The alternative is to set the length of the transition curve to equal the length of the super-elevation development and this is discussed in the following section of this chapter.

22.6.4 Setting out the transition

The starting point of the spiral, TS, is located at a distance, T, from the PI of the original tangents with

$$T = (R + S)\tan \Delta/2 + L/2$$

where
R = radius of circular curve
S = shift
Δ = deflection angle of the circular curve
L = selected length of transition curve

The deflection angle of the spiral is

$$\begin{aligned}\varnothing &= \frac{L}{2R} \quad \text{(in radians)}\\ &= \frac{L}{2R}\frac{180}{\pi} \quad \text{(in degrees)}\\ &= \frac{L}{2R}\frac{200}{\pi} \quad \text{(in gons)}\end{aligned}$$

It is to be noted that European practice is to use the gon in preference to the degree

(400 gons = 360°).

The clothoid is often defined in polar coordinates so that the chord length, $\delta l = r\delta\phi$ and $r = \dfrac{l}{K}$ so that

$$\frac{\delta l}{\delta\phi} = \frac{l}{K}$$

and hence

$$\phi = \frac{l^2}{2RL}$$

where

ϕ = deviation angle between the tangent and the transition curve at any point along the curve (radians)
l = distance along the transition curve to any point along the curve
R = the radius of the circular curve
L = length of the transition curve

This relationship between ∅ and l is the basic equation of a clothoid.

The transition can be set out by calculating the deviation from the tangent to the points along the curve, ϕ, and taping the distance from the TS along the curve to these points.

Alternatively, setting out could be by use of Cartesian coordinates. In this case

$$x = l - l^5/[40(RL)^2]$$

$$y = \frac{l^3}{6RL} - \frac{l^7}{336(RL)^3} + \cdots$$

where

R = radius of circular curve and
L = length of transition

For the purposes of setting out, it would be necessary to calculate the bearing and distance between the *TS* and the points along the transition curve by means of a series of polars between *TS* and the points along the transition curve, hence the comment that setting out of a spiral is more conveniently done by use of polar coordinates in the first instance.

The offset in the case of the cubic parabola is given by $\frac{l^3}{6K}$. As $K = rl$ it follows that the offset can be expressed as

$$y = \frac{x^3}{6rl}$$

22.7 TRIANGULATION

22.7.1 Introduction

In the calculations discussed in the previous section, only a single ray was involved and the available data were sufficient to fix either the distance and bearing between two known points or, given one set of coordinates and the bearing and distance to the unknown point, its coordinates were uniquely defined. In summary, from four items of information, two unknown items could be calculated. Consider the fact that the standard form of a linear equation is

$$y = ax + b$$

If the two parameters, a and b, of the relationship are known, the dependent variable can be calculated for any value of x. In coordinate calculation, two items of information have to be calculated, hence the need for two equations and four items of basic input information.

Triangulation is the name given to the process whereby the properties of the triangle are used to calculate the coordinates of an unknown point. Triangulation is brought into play when the available data are not so conveniently arranged that either a simple polar or join will suffice. For example, the bearing to the unknown point from one known point and the distance from another may be known. It is necessary to reorder this information so that the unknown coordinates can be calculated.

22.7.2 Proper identification of a triangle

The very name ‘triangle’ suggests that three angles are involved. It is, however, possible to have three connected rays, A, B and C referred to as ‘trilateration’.

Traversing a closed figure to end up pointing in the original direction obviously entails a course change of 360°. In short, it is axiomatic that the external angles of a closed figure sum to 360°.

Considering this in relation to a triangle, from basic geometry each external angle is also

$$\text{Ext} < A = 180 - \text{Int} < A;$$
$$\text{Ext} < B = 180 - \text{Int} < B;$$
$$\text{Ext} < C = 180 - \text{Int} < C;$$
$$360 = 540 - (\text{Int} < A + < B + < C)$$

from which it follows that the sum of the internal angles of a triangle is 180°.

This makes it possible, in terms of the fundamental linear relationship, to derive one new item of information, to whit the magnitude of the third angle of the triangle, from two items of information, being the magnitude of the other two angles. It is then possible to construct a triangle knowing only what the magnitude of two of the angles is. Unfortunately, the triangle is still not uniquely defined, because it could have London, San Francisco and Pretoria as the points of the triangle or the points could be metres from each other. Some indication of scale of the triangle is required such as its area or the length of one of the sides. Furthermore, the sequence of the angles hasn’t been defined. The two-dimensional triangle could be rotated in the third dimension so that, effectively, one could be looking at it from the front or the back.

With these further bits of information it is possible to uniquely identify and thus construct the triangle on paper. On the earth’s surface, the problem is still not adequately resolved because one further item of information is required and that is the orientation of the triangle. Is the base line of the triangle north–south or anything else through a range of 360°? Only with this last piece of information also in hand is it possible to uniquely solve the triangle.

22.7.3 Solving the triangle

All triangulation problems commence with knowledge of the coordinates of two points. This knowledge identifies the scale of the triangle and its orientation, for example, the baseline could be 5 kilometres long and have a bearing of 135° 27′ 46″.

The additional information required to geometrically construct the triangle could be a combination of

- The lengths of the other two sides of the triangle, whereby the position of the third point is fixed by the intersection of two arcs. Of course, the arcs intersect twice, that is, on both sides of the base line, so it is still required to resolve which of the two available answers the required one is.

- The length of one side and the opposite angle. The triangle is then defined by the intersection of a line at the given bearing from the appropriate point on the baseline and an arc around the other end of the base line.
- The two angles between the baseline and the unknown point which is then fixed by the intersection of the two rays.

Calculation of the triangle takes the form of deriving the information needed to be able to ultimately carry out a join or a polar from one or another of the two points defining the base line. For preference, calculation should be from both ends of the base line as a cross-check of the accuracy of observation and calculation.

Addressing the three possibilities listed above, the calculation of the coordinates of the unknown point can be derived as follows.

1. Two sides known (additional to the baseline length between A and B)
   From Figure 22.4, it can be seen that the height of the triangle, h, can be expressed as either

   $$h^2 = L_{AC}^2 - x^2 = L_{BC}^2 - (L_{AB} - x)^2$$

   from which it can be deduced that

   $$x = \frac{L_{BC}^2 - L_{AB}^2 - L_{AC}}{2L_{AB}}$$

   $$\cos < CAB = \frac{x}{L_{AB}}$$

   so that the coordinates of point C can be calculated by a polar from point A in terms of the distance L_{AB} and the bearing calculated from the bearing between points A and B plus or minus the angle at point A.
   The format of the calculations of the triangle is illustrated in Figure 22.6.
2. The length of one side known and the opposite angle (additional to the baseline length)
   In this case the well-known sine rule can be brought into play. This states that

   $$\frac{L_{AB}}{\sin < ACB} = \frac{L_{BC}}{\sin < BAC} = \frac{L_{AC}}{\sin < ABC}$$

   The known length in addition to that of the baseline, L_{AB}, is L_{AC} and the opposite angle is $<ABC$. Using the sine rule

   $$\sin < ACB = \frac{L_{AC}}{L_{AB}} \sin < ABC$$

   $$<BAC = 180 - <ABC - <ACB$$

   A polar from point A using the length L_{AC} and the bearing based on the bearing of AB plus or minus $<BAC$ will give the coordinates of point C. Alternatively, or as a

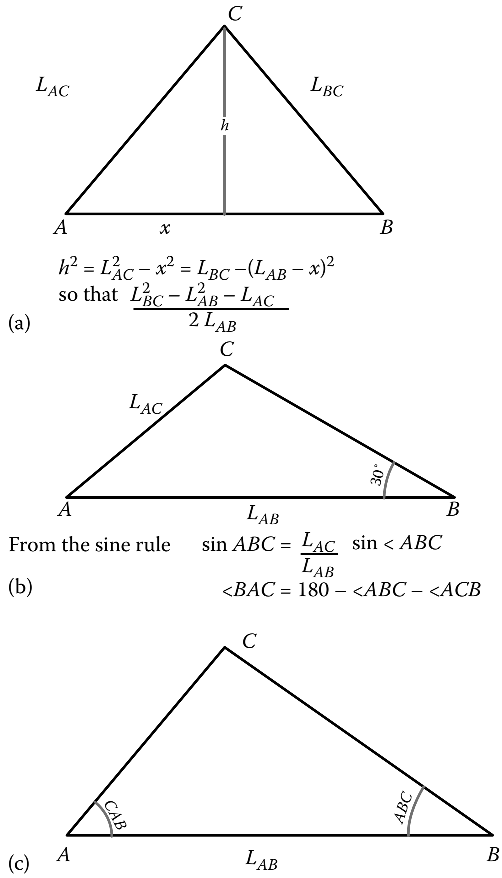


Figure 22.6 Calculation of the triangle. (a) Two sides known (additional to L_{AB}), (b) one side and the opposite angle known (additional to L_{AB}) and (c) two angles known.

cross-check, the length, L_{BC}, could be calculated and this used with the known $<ABC$ in a polar from point B to derive the coordinates of point C.

3. Two angles known

   This is also an application of the sine rule. The two known angles define the third so that both unknown sides can be calculated. The coordinates of the unknown point are once again established using a polar either from point A or point B with the other possibly serving as a cross-check.

22.8 RESECTION

The above calculations are based on the fact that it is possible to set up at a point for which the coordinates are known and, from a series of observations, derive the coordinates of unknown points. Resection is the process of setting up at a point for which the coordinates are unknown and then deriving these coordinates from observations to known points.

The usual application of resection is the calculation of the coordinates of survey beacons close to the route to be set out from observations to remote trigonometrical survey beacons. Joins between the coordinates of points along the road centreline and the survey beacons will provide the distances and bearings from the survey beacons to points along the centreline. These distances and bearings will then be used in the physical equivalent of the polar to locate the centreline of the road on the ground.

The process of resection can be geometrically compared to plotting, on a piece of tracing paper, rays from a point with the angles between the various rays to the remote trigonometrical stations, as observed in the field, between them. The tracing paper is then slid over a plan on which the trigonometrical stations have been plotted to scale until the rays pass through the known points. The point of intersection of the rays provides the location of the unknown point.

A minimum of three known stations is required to derive a unique answer, so that resection is also known as the three-point problem. If, in addition to the measured angles at the unknown point between the rays to the known points, the coordinates of only two points are known, the solution is not unique. To illustrate, if a circle is constructed such that its diameter is equal to the base line length, the angle at any point of the circumference of the circle is a right angle. Alternatively stated, if the angle between the two rays is a right angle, the unknown point could be located at any point along the circumference of a circle, the diameter of which is equal to the distance between the two known points.

The resection problem (Wolhuter, 2003b) is illustrated in Figure 22.7.

Measured horizontal angles, α and β, at the unknown point, P, are between the rays to the known points, A, B and C.

The coordinates of point P are expressed as (X_P, Y_P). The X-coordinate is measured parallel to the equator and is sometimes referred to as its easting. In the era of the sailing ships, tea clippers in the Roaring Forties were said to be 'running down their easting'. The Y-coordinate is measured along the line of longitude and is sometimes referred to as the northing of a point.

The X-coordinate is calculated from

$$X_P = \frac{k_1 X_A + k_2 X_B + k_3 X_C}{k_1 + k_2 + k_3}$$

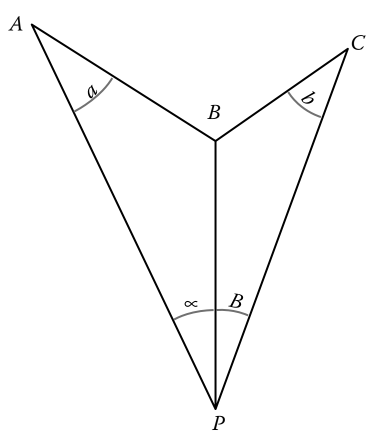


Figure 22.7 The resection.

and the *Y*-coordinate from

$$Y_P = \frac{k_1 Y_A + k_2 Y_B + k_3 Y_C}{k_1 + k_2 + k_3}$$

where

$\frac{1}{k_1} = \cot A - \cot \propto$ with $\propto$ = clockwise direction between directions *PB* and *PC*

$\frac{1}{k_2} = \cot B - \cot \beta$ with β = clockwise direction between directions *PC* and *PA*

$\frac{1}{k_3} = \cot C - \cot \gamma$ with γ = clockwise direction between directions *PA* and *PB*

Chapter 23

Environmental issues

CONTENTS

23.1 INTRODUCTION

Transportation is defined as the safe, convenient and accessible movement of people and goods with minimum environmental side effects. In this chapter the focus is the side effects, which, by definition, are always bad. Historically, geometric designers limited their interest to what lay inside the road reserve. Anything outside the road reserve was regarded with indifference except for the view that the environment was an endless supply of road building material and that swamps were things that had to be drained if they impacted on the road at all. More recently, designers have come to realise that there is a world beyond the boundary fences, that building materials are a finite resource, which, in many countries, have already been exhausted and that the previous stinking swamps are wetlands to be cherished.

The fact of the matter is that roads, as part of the transportation network, represent an economic benefit to a country, with many politicians echoing the International Road Federation in proclaiming that ‘prosperity follows the use of the road’. Unfortunately this benefit comes at a cost in terms of noise, air and water pollution, visual intrusion, dislocation of once cohesive communities and risk to the continued preservation of fauna and flora, not to mention risk to life and limb.

The life of any road comprises four components:

- Planning
- Design
- Construction and maintenance
- Usage

Construction has a short-lived nuisance value in terms of noise, dust and inconvenience to would-be road users. Its sphere of influence is limited to the immediate surroundings of the road and it is necessary to understand the impact of construction activities on plants and animals in close proximity to the road. Maintenance activities are usually not particularly intrusive.

Most of the problems arise after the road has been commissioned. By this time, these problems are set in concrete and remedial measures are usually expensive to implement. The only time at which it is really possible to minimise the side effects of the road is during the planning and design phases, with the major emphasis being on planning. It must be pointed out that determination of the nature and extent of the impact of the road on its environment is usually beyond the capabilities of the transportation engineer. Expertise in and input from other disciplines is essential if the goal of minimum side effects is to be achieved. In this chapter, these issues are discussed and measures aimed at minimisation of environmental impact suggested.

23.2 ECOSYSTEMS

23.2.1 Introduction

Long-range ecological surveys are needed to identify plants and wildlife that are located in the area where a road will be planned, designed and built. Sensitive ecosystems such as wetlands need to be identified and shown on maps of the area of interest. These surveys have to be updated on a regular basis to establish the effects of climate change and other environmental changes on the various eco-populations, that is, are they diminishing or increasing in the time prior to design and construction of the road. These surveys provide a baseline from which it is possible to predict what the impact of the road on the environment is likely to be. The long-time framework approach provides more opportunity than project level surveys to avoid, minimise or mitigate potential transportation problems (Cramer, 2009).

The Delaware Department of Transport has used a ranking system to identify the value of habitats. This is based primarily on the location of known threatened and endangered species of flora and fauna. It is possible that this ranking may be expanded to include other ecological values such as size and diversity of habitats, fragmentation of habitats and connectivity to other resources.

23.2.2 Effects of roads on the environment

The effect that roads can have on animals is generally negative. Various studies have shown that some species avoid roads because of increased noise, pollution, visual disturbance and predators using roads as corridors. These effects include

- Wildlife vehicle collisions
- Habitat loss
- Habitat fragmentation
  - Reduction in patch size
  - Edge effects
  - Barrier effects
  - Population fragmentation
- Altered habitat quality

These are discussed further in the in the following sections.

23.2.2.1 Wildlife-vehicle collisions

Wildlife collisions injure or kill millions of animals annually across the world. In the case of larger animals such as moose in North America, buck in Africa and kangaroo in Australia, property damage is significant and occupants of vehicles are often injured or killed in such crashes. The greatest impact is on animals that have large home ranges or need to migrate periodically either to complete their life cycle or in search of grazing. These animals often have to cross roads.

Populations with low reproductive rates or small populations are particularly vulnerable because of the relative importance of each adult in the population. It seems to be a common cause that increasing traffic volumes causes greater threats to animals up to the point where traffic volumes are so high, for example, 10,000 vehicles per day, that they create an impenetrable barrier deterring animals from crossing the road altogether.

23.2.2.2 Habitat loss

Habitat loss occurs when an area that previously provided food and shelter to animals falls prey to urban development. Species with large home ranges such as carnivores like lions are at greater risk than other animals.

23.2.2.3 Habitat fragmentation

As the road network develops and becomes more dense, the areas between the roads get cut into smaller and smaller bits. To illustrate: A farm being increasingly crisscrossed by road links will reach the stage when the remaining portions of the farm are no longer economically viable entities. A similar thing happens to undeveloped areas where animals lose access to food or water and biological diversity is impossible to maintain. Small populations could face extinction.

Wildebeest migrating across the Serengeti plains of Africa create herds comprising many hundreds of thousands of animals joined by many zebra and some Thompson's and Grant's gazelles, eland and impala. The Serengeti rainy season is short, typically starting in early November. The herds of the migration arrive in late November and December. Most hartebeest calves are born in February and, by May, the herds start moving back north. Migration continues until about August. The southward migration starts in October, arriving in the southern Serengeti in early November. And so the cycle repeats itself.

Seeing that the migration is aimed at the search for food, a road cutting across their migration path would have disastrous consequences. Ecologists suggest that it would cut the wildebeest herd to less than a third of its current size with a knock-on effect on the size of the population of carnivores, specifically lions. The Tanzanian government had intended to build such a road as a link to the relatively undeveloped north of the country. Fortunately, this intention was scrapped in 2011 as a result of the international outcry that greeted the proposal.

23.2.2.4 Altered habitat quality

Gravel roads create a nuisance outside the limits of the road reserve by coating vegetation with a layer of fine dust. This has economic implications where the road passes through areas where fruit and vegetables are grown, particularly for export. The dust coating causes animals to move away from the road in their search for food. In consequence, the availability of food diminishes. In effect, the deterioration in habitat quality is another form of habitat loss.

Noise scares birds and animals and the passage of a vehicle on any road surface, but particularly on gravel, is a sudden noise that will cause them to shun the vicinity of the road. Continuous noise, provided it is not too loud, seems to be less threatening and animals can become accustomed to the presence of vehicles and people in their immediate vicinity. The pigeons in London's Trafalgar Square or St Mark's Square in Venice are cases in point.

23.2.3 The accommodation of wildlife

Referring to the Serengeti wildebeest, it is not possible to build a road that would not be vastly damaging to their migration pattern. Providing several culverts for them to pass through is patently beyond the bounds of economic reality. Furthermore, whereas domestic animals such as cows on a dairy farm become accustomed to the twice-daily passage through an underpass, wild animals refuse to be constrained. If the migratory path is not to be cut, it would be necessary to provide continuity by putting the road into a series of underpasses. Each of these would have to be of a considerable length if the animal overpass is not to become a choke point, typically 50 metres or more.

The alternative is to leave the road unfenced, which could create a hazard for road users. In defence of this alternative, migrating animals do not move after sunset and would not, as a general rule, rest on the unyielding surface of a paved road. While crossing the road, the animals would therefore be visible to approaching motorists who would just have to stop and wait patiently until, possibly some days later, the herd had passed. If the road is to be left unfenced, temporary warning signage or speed restrictions could be provided as a safety measure. The preference for temporary signage is that migratory paths vary slightly from year to year and there are also times when no migration is taking place. Permanent signs become 'part of the furniture' and simply are not noticed.

Where smaller populations have to be accommodated, the rationale for providing an animal overpass remains. Small individual animals can, however, be accommodated by underpasses in the form of dry culverts. If these culverts serve to drain perennial flows of water such as streams or rivers, they should be wider than the stream to ensure dry areas on which land animals can pass through the culvert in safety. The inverts of such culverts should be constructed slightly deeper than the stream bed. In time, they silt up and the previous level and gradient of the stream bed will be reinstated.

Areas of high-value habitat should be connected. Areas that do not provide habitats on both sides of the proposed road should be identified as useful boundaries that could serve to identify routes with minimum ecological damage (Massachusetts Department of Transportation, 2006).

New Jersey barriers, typically provided on the medians of dual carriageways to separate the traffic on the opposing roadways, may trap animals and place them at risk of collision with vehicles. The barriers can be modified by providing holes in the barriers through which smaller animals can pass. These holes could also serve as scuppers improving storm water drainage.

23.3 POLLUTION

23.3.1 Introduction

Pollution is simply something that should not be there, with 'there' meaning the road reserve and the adjacent area. Pollution mainly takes the form of

- Air pollution
- Water pollution
- Soil contamination
- Noise

In addition, street lighting can also be a form of pollution, referred to as light trespass, by not being properly masked. Littering refers to the human propensity for throwing rubbish onto public and private property as opposed to placing it into garbage containers. Thermal pollution refers to temperature changes in natural water bodies, typically caused by power stations using them as coolants. The 20th century has seen a new form of pollution in radioactive contamination, where nuclear materials can remain active for several thousand years. This is principally from nuclear power stations and nuclear weapons.

Visual pollution stems from large roadside billboards, open cast and strip mining and municipal landfill dumping of industrial and domestic waste. Roads can also be eyesores of note. If their visual intrusion is to be minimised, two aspects of geometric design require consideration. The first is internal harmony, which is the interaction between the elements of the horizontal and vertical alignment, the concept of the abstract ribbon in space. The second aspect is external harmony, which is the interaction between the road and the landscape. Ideally the road should blend into the landscape to the extent that the landscape would appear to be the poorer without it. Chapter 9 deals with this matter in depth.

23.3.2 Air pollution

Air pollution results from the release of chemicals and particulates into the atmosphere. Exhaust emissions from motorised vehicles include carbon monoxide, sulphur dioxide and various nitrogen oxides. The latter two combine with moisture in the atmosphere as sulphuric and nitric acid, resulting in acid rain which has been known to kill entire forests.

Animals' keen sense of smell results in their being aware of pollutants, particularly sulphur dioxide, much sooner than humans. They are therefore more inclined to move away from the road reserve, possibly to a considerable distance. On the downwind side of the road, this could be a kilometre or more. In effect, air pollution results in habitat loss.

23.3.3 Water pollution

Road design includes the disciplines of pavement design, which is design of the elements below the road surface, and geometric design, which is the design of the visible elements of the road. In the absence of a hydrologist, drainage is the purview of the geometric designer.

During dry spells, roads acquire a surface layer comprising dust, oil spills, rubber from vehicle tyres and other contaminants such as heavy metals. During the first rains thereafter, this material is washed off the road and into the storm water drainage system. It has been said that this water is far more polluted than anything found in a foul water sewer. It follows that the geometric designer must ensure that this contaminated water does not directly enter existing watercourses where it can end up in the potable water supply. This is best achieved by directing the storm water into retention or detention ponds, referred to in the European literature as storm basins.

The difference between the retention and the detention ponds is that, in the former, the water soaks into the ground and does not enter the storm water drainage system at all in the case of storms with a return period shorter than that designed for. Storms with a longer return period have to be accommodated by overflows that will direct excess storm water ultimately to the natural drainage systems of rivers and streams. Retention ponds are normally constructed in open areas such as parks and golf courses and can become large wetlands. Detention ponds are intended to regulate peak flows of storm water. They modify the time of concentration of the storm water, which is decreased by virtue of the lower roughness of the impermeable surfaces – buildings and paved roads. This results in the possibility of higher intensities of rainfall that have to be accommodated. In urban areas, the volume of runoff is increased by the spread of these impermeable surfaces. Detention ponds are intended to retard the flow of water and thus reinstate the hydrological conditions that prevailed prior to the development of the area. Because there is an outflow of water from a detention pond it can be smaller than the retention pond.

In both cases, the planting of vegetation such as grasses and shrubs that normally grow in or near to water can create a wetland that plays a role in water purification. Grasses help to trap suspended material. If the water can be retained in the detention pond for some time, its biological purification would be supported.

23.3.4 Noise

According to the World Health Organization (1999), an adverse effect of noise is defined as 'a change in the morphology and physiology of an organism that results in impairment of functional capacity; or an impairment of capacity to compensate for additional stress; or increases in the susceptibility of an organism to the harmful effects of other environmental influences'.

Noise can be defined as unwanted sound. It is an impediment to sleep and also makes conversation difficult, if not actually impossible. Little research seems to have been done on the effect of noise on animals but they have been observed to move away from noisy environments. Three solutions (Permanent International Association of Road Congresses [PIARC], 2000) to protect people living in close proximity to a road are

- Surfacing that reduces tyre/road contact noise
- Noise barriers
- The soundproofing of buildings

The noise levels created by various road surfaces at a speed of 90 km/h are 79 to 82 dB(A) by concrete paving, 74 to 80 dB(A) by bituminous paving and 71 to 78 dB(A) by premixed bitumen (also known as asphalt). The World Health Organization's guideline for community noise states that a noise level of 55 dB at the building facade of a dwelling is the maximum accepted level. The guideline also states that 45 dB should be the maximum noise level during the night.

Trucks comprise about 5 to 10 per cent of the traffic on a road but generate 50 per cent of the traffic noise. Above 50 km/h, the noise emanates from the tyre/road contact and, below this speed, by the engine and gear train.

Noise barriers can be earth berms, dense foliage or acoustic fencing. Earth berms are normally landscaped with plantings to make them aesthetically less intrusive. These plantings also support attenuation of noise. Acoustic fencing could be concrete or timber barriers and should be designed to blend with the environment.

Buildings where noise could be particularly distracting include schools, hospitals and churches. If the best efforts of the geometric designer have proven to be inadequate, the last resort is to soundproof the buildings themselves. In cold regions, windows are often double glazed. Double glazing substantially reduces the noise level inside buildings.

23.3.5 Visual pollution

Examples of visual pollution include large billboards alongside roads, scarring of the landscape from opencast and strip mining, overhead power lines, municipal tips of industrial and residential garbage and littering.

Roads can be and often are generators of significant levels of visual pollution. Borrow pits and spoil heaps outside the road reserve are invariably unsightly. Poor geometric design can result in the road itself being a visual intrusion of note. This topic is dealt with in depth in Chapter 9.

As discussed in Chapter 9, two aspects of aesthetic considerations require attention: internal and external harmony. Internal harmony refers to the interaction between the elements of the horizontal and vertical alignments, which are the horizontal curves and tangents and the vertical curves and the grades between them. This is often referred to as 'the abstract ribbon in space'. External harmony involves the way in which the road fits into the landscape. A road with a dead straight horizontal alignment ploughing through a series of gently rolling hills would almost certainly suggest a totally lack of artistic sensitivity on the part of the designer. A road running at right angles to a series of ridges creates the infamous roller coaster effect and should also be resisted. The guiding principle is to fit the road into the landscape to look as though it belongs there and has always been there.

23.4 COMMUNITY IMPACTS

23.4.1 Introduction

In addition to the impacts on the natural environment, there are also impacts on the social environment that have to be taken into account. These include

- Worsening of public health
- Loss of community cohesiveness
- Damage to community values
- Financial loss to local communities
- Injuries and death

In the past, geometric designers focussed their attention on the transportation imperatives of mobility and accessibility. The effect of their activities on local communities tended to be regarded with indifference until public outcry forced a change in attitude. The recent shift towards context-sensitive design has inevitably resulted in a greater level of environmental awareness than was previously the case. This has served to highlight the dichotomy of roads. They are useful to travellers but a profound nuisance to local communities. They provide an environment where people can engage socially and also an environment where they can die.

23.4.2 Public health

Public health is adversely affected by air pollution and it is addressed here in a slightly different context from the earlier discussion. Section 23.3.2 addresses air pollution in a rural context whereas this section focusses on the urban environment. The difference between the two stems from the greater presence of vehicles in the urban environment. In addition to the urban environment being noisier and more dangerous than the rural environment, the air is more difficult to breathe.

Particulates are very small particles, typically invisible to the naked eye. For convenience they are classified into two groups, a larger size group and a smaller one. The 'big' particles are between 2.5 and 10 micrometres in size (described as PM_{10} for particulate matter up to 10 micrometres in diameter). They are about 25 times thinner than a human hair. The 'small' particles are smaller than 2.5 micrometres, described as $PM_{2.5}$ and are about 100 times thinner than a human hair. Of the two, it is the small particulates that represent the greater danger because they can travel further into the lungs, where it is almost impossible to remove them by coughing. They comprise toxic organic compounds and heavy metals and are typically generated by motorised vehicles.

Health effects (PIARC) include

- Coughing and shortness of breath
- Asthma
- Lung damage such as emphysema, decreased lung function and chronic respiratory disease
- Premature death in individuals with existing heart and lung diseases

Tokyo probably has one of the higher levels of air pollution, as it has been recorded that traffic police on point duty can work only for fairly short periods before having to repair to small kiosks at their control points to recover by breathing pure oxygen.

23.4.3 Impact on communities

A sense of community derives from sporting activities, religious services and children playing at school. Driving a dual carriageway or, worse still, a freeway through the area will immediately result in a loss of contact between the two sides of the road and the town or village will become two smaller towns or villages. In time, there will be a duplication of facilities and the towns will become independent entities and the previous community cohesiveness will be irretrievably lost.

While this scenario could be avoided by locating the road to bypass the community, the bypass is a mixed blessing. A long-distance route passing through a town results in Main Street traffic becoming a mix of through and local traffic, a mix of mobility and accessibility, in fact. As all geometric designers are aware, this mix simply does not work because of the difference in the needs of the local and the through traffic. During holiday seasons, the Main Street could be jam-packed and brought to a standstill. This would result in the local community raising an outcry and demanding that the through traffic be taken away from their ‘beautiful little village’. Opening the bypass to traffic will result in an outcry from the local merchants who had previously relied on the passing traffic for their trade.

The author has personal knowledge of a village, Grabouw in the Western Cape Province of South Africa, that was brought to an economic standstill as a result of the N2 National Road being relocated to pass 4 kilometres to its west. With time, the bypass proved to be somewhat of a magnet drawing local development towards itself. It is still a bypass but it now closely skirts the edge of the development. Grabouw has changed from being a thriving village to becoming moribund and then a thriving small town. The metamorphosis took about 20 years, or an entire generation, to complete with much hardship and heartbreak along the way. There is always the possibility that the latter half of the metamorphosis will not take place and the village simply become an abandoned derelict.

Bypasses are useful in separating through traffic from local traffic, but should ideally not be too far from the village. If passers-by could easily see it, they may be more likely to stop for something to eat or drink or even to explore if on a holiday trip. This would be of benefit to the local community. More importantly, it could be an incentive to through drivers and their passengers to stop for a rest break, thus having a useful impact on road safety.

For preference, bypasses should be located to take advantage of the prevailing wind direction by being located where possible on the downwind side of urban developments.

23.4.4 Damage to community values

Most towns have landmarks, historical buildings and monuments that are important to them. Locating a road to pass through a building that had been erected during Norman times in England would certainly generate much resistance and force a relocation of the road, and rightly so. A restful park laid out with lawns, benches under large shade trees and possibly also catering for sporting activities such as swimming, tennis or bowling would be considered by the local community to be sacrosanct. Chopping it into smaller parks would not be acceptable.

Similarly, no community considers its cemetery to be a dumping ground for dead bodies. People who in their lifetimes made valuable contributions to their communities are buried there. And they are revered. Locating a road to pass through a cemetery is, because of the outcry it typically generates, virtually impossible. Furthermore, the need to exhume all the graves creates legal problems of note. For example, it is usually necessary to get the permission of the next of kin for the removal of a body.

23.4.5 Injuries and death

As quoted in Chapter 1, Mother Shipton (1488–1561) predicted that

> A carriage without horse will go,
> Disaster fill the world with woe.

In spite of the undoubted benefits that a road provides, people will driver faster than is wise, overtake others where there is inadequate sight distance, get behind the wheel of a car with an illegal level of alcohol in the bloodstream or behave in any of the other possible ways that could cause a crash. While injury or death of the driver often result in these cases, many innocent lives often also are lost and people injured or permanently maimed as a result of these crashes. As stated in Chapter 4:

The 'accident black spot' where numerous drivers make the same mistake is not unknown. It can reasonably be inferred that the road misinformed the driver in some or other way. While driver error has been identified as the cause for a high percentage of crashes, typically seventy per cent or more, it is suggested that some of these could have arisen from the road design seducing the driver into an incorrect response.

23.5 ENVIRONMENTAL IMPACT

23.5.1 Introduction

Environmental impact studies are a recent development in the growing concern regarding the nature of the impact of humankind on the natural environment.

In the United States, these studies are a two-step process. The first step is an environmental impact assessment (EIA). This is an assessment of the possible positive and negative impacts that a project may have on the environment. It is defined as 'the process of identifying, predicting, evaluating and mitigating the biophysical, social and other relevant effects of development prior to major decisions being taken and commitments made'. The intention of the assessment is to determine whether a project would significantly affect the environment in its broadest sense of air and water pollution, noise and risk to flora and fauna. The outcome of an EIA is either a finding of no significant impact (FONSI) or, alternatively, a requirement for a detailed environmental impact statement (EIS).

23.5.2 Legislation

Environmental law addresses two major topics: pollution and remedial actions. Broadly, legislation is intended to prevent pollution taking place and, should it occur, to mandate the actions that should be taken to remedy its consequences. Furthermore, the legislation imposes penalties on the parties responsible for the pollution. In general, these require the polluter to clean up the mess and also to pay fines to reimburse those who have suffered financial loss as result of the pollution.

The BP Deepwater Horizon oil spill in the Gulf of Mexico is a case in point. On 20 April 2010, high-pressure methane leaked from the well and rose into the drilling rig, where it exploded. This resulted in an oil leak, estimated to have a volume of about 5 million barrels (210 million US gallons or 800 million litres), that is widely regarded as the largest in the history of oil spills. In addition to having to bear the costs of remedial actions estimated at about $37 billion, penalties and fines served on BP amounted to $6 billion. Apart from the pollution of several miles of beaches, the loss of marine life is impossible to quantify.

The environment is no respecter of national boundaries and this finds expression in the more than a hundred international treaties and conventions in the field of environmental law arising from countries with common borders perceiving the need to jointly safeguard their shared environment. An outstanding example is the North American Commission for Environmental Cooperation This body, comprising Canada, the United States and Mexico, has the three countries jointly protecting their shared environment. In addition, each has legislation specific to its own concerns. A problem with international treaties and conventions is that, short of declaring war on the defaulter, they are unenforceable simply because no country is prepared to do anything that is not in its own best interests. For example, it is interesting to note that, as a major producer of air pollution, the United States has not ratified the Kyoto Protocol. All 190 members of the United Nations as well as the European Union, but with the notable exceptions of the United States and Canada, are parties to the Protocol. The United States signed but did not ratify the Protocol and Canada withdrew from it in 2011. The Protocol was adopted in 1997, and entered into force in 2005.

Many countries have enacted legislation aimed at protecting the natural environment from human depredation. China has 37 laws addressing environmental issues and Japan has 28. In Europe, 34 countries have enacted environmental legislation and Bulgaria alone has 19 such acts. In Africa, only two countries, Egypt and South Africa, have shown any interest in preserving the environment. Australia has 14 acts on its statute book and these are amplified by legislation by the individual states addressing issues specific to their own circumstances. New Zealand has 20 such acts.

23.5.3 The legislative requirements for environmental impact assessments

The legislation discussed in the preceding text generally requires that any project that has a prospect of damaging the environment must be the subject of an environmental impact assessment (EIA). The general structure of an EIA is

- Background
  - Description of the proposed project
  - The need for an environmental assessment
- The proposed remedial actions (if any)
- Comparison of these with possible alternative actions
- The setting up, conducting and outcome of a public involvement exercise
- The environmental consequences of the project and the recommended remedial action
- A conclusion and general summing up

An EIA is normally conducted by the government agency responsible for environmental matters. Having the EIA conducted by the implementing agency or the private body wishing to carry out the developmental work is an undesirable conflict of interest. There would be far too strong a temptation either to slant the assessment towards there being no significant side effects or to minimise whatever effects cannot be hidden from the public view.

In the United States, an EIA is required for all projects where there is a prospect of damage to the environment. The intention in conducting the assessment is to ensure that the authorities properly consider all the issues that have to be taken into account in the decision to proceed with the project or modify it in some way to minimise its side effects or to scrap it completely. In terms of the National Environmental Policy Act, EIAs require decision makers to justify their decisions on the basis of environmental studies and comment received from interested parties on the potential environmental impact of the proposed project. The usual outcome of an EIA is either to rule that there is no significant impact and the project can proceed as proposed or that a full impact assessment statement has to be prepared.

The structure of the European Union EIA is similar to that described above. A directive defining the execution of an EIA was introduced in 1985. Subsequently, it went through various amendments and the current version is Directive 1011/92/EU of 13 December 2011. The European Union requires the use of the Leopold matrix as a convenient way of describing the potential impact of a project on the environment. It is similar to utility analysis in its concept and is a table comprising columns describing the various activities of the project and rows defining the various environmental issues requiring consideration. The individual cells of the matrix are filled with values from –10 to +10 to define the magnitude of the impact of each activity on the environment and from 0 to 10 to define the importance of each to the environment.

23.5.4 Environmental impact statements

The two-step approach of an EIA being followed, if need be, by an environmental impact statement (EIS) in the consideration of the environmental effects of a project seems to be unique to the United States. Review of the processes followed in several other countries indicates that they limit themselves to a single process referred to universally as environmental impact assessment. It would appear that these are the equivalent of the American EIS.

The EIS has a structure similar to that of the EIA. The main difference between them is in the level of detail required. The EIS addresses the issues of

- A description of the affected environment
- Analysis of the impacts of the project on the environment, including on
  - Threatened or endangered species of flora and fauna
  - Air and water quality
  - Historic and cultural sites
  - Local communities in terms of the social and economic environment
- A range of alternative actions to mitigate the environmental impact of the project including a recommendation of the preferred action and also the nil or 'Do nothing' alternative; and sometimes
- A financial plan for the preferred option

It has to be accepted that, invariably, impacts will be negative. If these impacts are substantial, the proponent could be called on to provide an environmental mitigation plan.

Chapter 24

Storm water design

CONTENTS

24.1 INTRODUCTION

In First World countries, the geometric design team may include a hydrologist. This person would be responsible for determining the size of the catchment area for a drainage structure. From this and considerations of return period, time of concentration and intensity of precipitation, the magnitude of the flood to be accommodated could be determined and hence the drainage structure appropriately sized. In other countries where technical expertise is often in short supply, it is frequently necessary for the geometric designer to undertake the design of the storm water system without the benefit of this expert input. An ability to quantify the characteristics of the catchment area of a drainage structure, followed by determination of the quantity of storm water that needs to be accommodated by the structure, is thus necessary. This chapter serves as a brief introduction to hydrology.

There is a difference between urban and rural storm water drainage. In urban areas a distinction is drawn between major and minor storms and also between the means available to accommodate them. This distinction does not normally apply to rural areas. The designer needs to be aware of these differences and of how they can best be addressed. One of the differences is in the components of the drainage system that apply in the two environments. Broadly, urban drainage is by means of kerbs and channels connecting to an underground system in the case of minor storms whereas detention and retention ponds are often included in the design of storm water management systems for major storms. Rural storm water drainage is usually by means of open channels and includes mitre banks to discharge storm water from the road reserve into the surrounding environment. The methodology of storm water management is discussed.

This is followed by discussion of the hydraulics of open channel and piped flow. This relates to the drainage of the area traversed by a road insofar as the sizing of the required drainage structure is necessary. The geometric designer must be able to design a vertical alignment that will provide sufficient height above the stream bed to accommodate the selected structure.

It may be possible to select almost any height of structure if the height of the gradeline is such that it naturally allows this choice without being forced. In flat terrain, the gradeline will tend to be close to ground level. The choice of the dimensions of the required drainage structure will thus be constrained if a series of unsightly humps where the road is lifted over pipe or box culverts is to be avoided. A flow of storm water could be accommodated, for example, by a single large-diameter pipe. Equally, a number of pipes of a smaller diameter could accommodate the same flow. It may even be necessary to select the dimensions of a series of low-height box culverts to the same end. In short, although the hydrology of the catchment area will indicate the magnitude of the flow to be accommodated, the designer does have a choice regarding the means whereby it can best be accommodated. To exercise this choice properly, the geometric designer would need to have some knowledge of hydraulics.

Hydraulics is also brought to bear on the drainage of the road itself. Rain falls on the road itself and this water can have dire consequences with respect to road safety and can also harm the road structure. The geometric designer will be directly involved in the development of the storm water reticulation system to ensure the integrity of the road structure and the safety of the road user as well as that of people living in close proximity to the road. A knowledge of the tools at his or her disposal and how they can best be utilised to this end is thus essential. This is also discussed in this chapter.

Finally, protection of the road itself against the ravages of storm water flow is essential. In addition to the differences between urban and rural drainage, there is also the difference between the philosophies of First World and Third World drainage that has to be borne in mind. First World drainage is aimed at preventing storm water from reaching the road, that is, upstream protection, by means, *inter alia*, of appropriate clearances above the level of flow. Third World drainage is predicated on water being allowed to flow over the road, that is, downstream protection, and then ensuring that protection of the downstream embankment slope against scour is provided. The benefit of this system of protection is that it lends itself to labour intensive methods of construction. Third World drainage will thus also enjoy attention in this chapter.

To summarise, the content of this chapter includes discussion of

- Hydrology and the calculation of the various characteristics of floods
- Drainage of the road surface
- Rural drainage
- Urban drainage
- Drainage systems for low-volume roads

24.2 FLOOD CALCULATIONS

24.2.1 Introduction

The principal concerns regarding considerations of storm water drainage are to protect the road from harm and also to ensure that damage to the surrounding environment caused by the road is kept to a minimum. The last-mentioned arises from the fact that storm water that had previously flowed freely across the landscape becomes concentrated to flow through culverts. In general, once storm water is concentrated, it is difficult to disperse the flow back to its original drainage pattern and, in consequence, erosion results.

The damage that can result from storm water is related to five characteristics (South African National Road Agency Limited (SANRAL), 2013):

- The maximum flood level that can be reached at a given point, measured in metres
- The maximum flow rate during the flood, measured in cubic metres per second (cumecs)
- The maximum flow speed measured in metres per second
- Flood duration being measured in hours from the time when the flow depth starts to increase above its normal level until it returns to the normal level

These, with the exception of the maximum flow speed, are typically illustrated by a hydrograph. A typical flood hydrograph is shown in Figure 24.1.

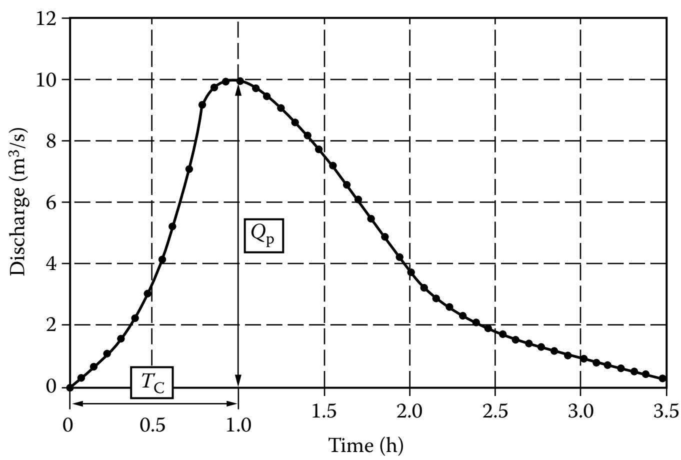


Figure 24.1 Typical flood hydrograph.

The maximum flow rate, known as the peak discharge (Q_P), is necessary to determine the culvert or bridge size required to convey the flood across the road reserve. It is also used to determine the back water curve resulting from the damming effect of the constriction provided by the bridge or culvert. With the peak discharge determined, the high flood level (HFL) can be calculated using hydraulic calculations as described in Section 24.3.

24.2.2 Methods of estimation of peak discharges

Hydrological modelling started with the development of the rational method by TJ Mulvaney in 1850. Since then, numerous hydrological models have emerged worldwide. These fall into two broad categories: statistical models and physically based models. They have also become significantly complex as the ecological and environmental scientists have sought to expand them to become hydrological transport models whereby the silt load carried by a flood is of interest, as are the various pollutants such as nitrogen-based and other fertilizers that can have a deleterious effect on plant and aquatic life (Todini, 2007).

24.2.2.1 Statistical models

Statistical methods rely on historical events to predict the likelihood of future events. These events comprise data gathered on the catchment area of interest or on a catchment with similar characteristics in the same geographical region. The closer the similarity between the two catchment areas, the more reliable the final determination of the flooding is going to be. Should there be any change in the physical characteristics of the catchment area, there could be a change in the flood causing factors and these would need to be taken into account.

An example is where what was previously pasturage becomes developed into a residential township. The roofs of houses and the surfaced road network would have a 100 per cent runoff compared to the previous 5 or 10 per cent runoff. The road network typically comprises about 20 to 30 per cent of the area of a township and houses comprise about a further 10 to 15 per cent. It follows that the portion of the catchment area that has been developed will show a higher runoff than was previously the case. Furthermore, because the overland flow will take place at a higher speed over the new impervious layer, the time of concentration will be shorter and hence the rainfall intensity higher than was previously the case.

Various statistical distributions can be used to model the flow characteristics of a catchment area. Typical examples include the binomial, the normal and the log-normal distributions. These distributions are discussed in Chapter 20. In the United States, the log-Pearson type III distribution is often used and, in the United Kingdom, the general extreme value (GEV) is normally used.

Statistical methods are generally geared to the estimation of the flood size occurring with a selected frequency referred to as the return period. Small culverts are typically those with a span of less than 2 metres and are often designed to accommodate the worst flood in 5 to 10 years. Culverts with a total span of between 2 and 6 metres would normally be required to accommodate a return period of about 20 years. With a total span of more than 6 metres, reference is usually to a bridge and the return period could be 50 or even 100 years. The difference in return period is related to the extent of the risk of damage caused by flooding. Replacing a pipe or small box culvert is not particularly expensive and it would not be sensible to construct a vast and expensive edifice to accommodate the worst storm of all time when this amounts to only a few cubic metres per second (cumecs). Replacing a large bridge, on the other hand, is a costly exercise and the risk of its being overtopped would have be related to a flood so severe that it could cause widespread damage of the road network and also other infrastructure and property. Hence, consideration of the longer return periods is required.

24.2.2.2 Physically based models

Geometric designers are principally interested in the volume of water that has to be accommodated by their drainage structures, but also in

- The possibility of scour under and downstream of the culverts and bridges forming part of their drainage system
- Erosion caused by the discharge of storm water from the road reserve onto the surrounding fields
- Silting of their culverts and side drains which can cause water to flow back to or be trapped on the road surface

As such the rational method is usually adequate for their purpose, which primarily has to do with the extent of runoff.

Runoff has four components:

- Surface runoff
- Interflow
- Groundwater flow
- Channel precipitation

24.2.2.3 Surface runoff

Surface runoff is water that reaches the streams by travelling over the soil surface. Streams include, in addition to permanent streams, tiny rills and rivulets that carry water only during or immediately after rain storms (Linsley et al., 1949). In consequence, surface runoff includes short distances of overland flow to the nearest minor channels, in contrast to long distances of overland flow. Some water will soak into the ground and some may be trapped as surface storage in local depressions. Runoff can occur only after infiltration and surface storage have taken place. Because the bulk of the surface runoff reaches the streams during the period of rain with the rest of the flow reaching the streams much later and over a longer period of time, it is the major contributor to flood peaks.

24.2.2.4 Interflow

Interflow is that portion of the precipitation that penetrates the soil surface and then moves laterally until its course is intercepted by a stream channel or it returns to the surface as seepage or springs. An impervious material layer close to the surface would increase the interflow portion of the total runoff. Like surface flow, interflow is a residual as it follows on percolation through the soil possibly to great depths as well as recharge of soil moisture.

24.2.2.5 Groundwater flow

Some storm water will flow downwards because of gravity and cause a rise in the level of the ground water table. This will occur only after a considerable period of precipitation because there is little likelihood of the groundwater table being added to in the case of a light rain on dry soil. It is possible that groundwater will not contribute to runoff at all and, if it does, it certainly will not have an impact on the peak flow. One of the problems that can be caused by excessive ground water is that it could lubricate a potential slip circle, hence causing a catastrophic landslide. This is particularly the case where the road is built on a fairly steep side slope. It would be necessary to ensure that, where the road is on embankment, it would be properly benched into the hillside. Similarly, a cut slope above the road should also be benched to reduce the possibility of its slipping onto the road surface.

24.2.2.6 Channel precipitation

Channel precipitation is rainfall directly onto the surfaces of streams, rivers and lakes. It can be taken into account by multiplying the surface area of all lakes, rivers and streams by the duration and intensity of the rainfall. While the drier states of the United States have less than 0.5 per cent of their surfaces covered by water, Maryland and Delaware both have more than 20 per cent of water surface area (Linsley et al., 1949). The weighted average for the United States is 4.3 per cent.

24.2.3 The rational method

It is the peak runoff rate that has to be accommodated by whatever drainage structure is provided. This rate is calculated as (Bengston, 2011)

$Q = CiA$ in Imperial units or $Q = 0.0028\ CiA$ in SI units

where

Q = peak runoff rate (cusecs; cumecs) due to a rainfall of intensity, i, over a catchment area, A
C = catchment coefficient (dimensionless)
i = rainfall intensity (in/h; mm/h)
A = area of catchment area draining to the point where the peak runoff rate is needed

24.2.3.1 The runoff coefficient

The runoff coefficient, C, is the fraction of the rainfall that becomes surface runoff and thus has a value of between 1 and 0. Its value is a function of the soil type, the land use and the slope of the catchment area.

The soil type has been classified by the United States Soil Conservation Service (SCS) into four groups as shown in Table 24.1.

The land use that affects the extent of runoff from a catchment area includes the extent of the area with impervious surfaces such as street, parking areas and buildings where the runoff would 100 per cent of the rainfall falling on them. Vegetative cover intercepts surface runoff which can then penetrate the soil instead of running off. The extent and nature of the vegetation varies from forest and bush to grazing and cultivated farm land. In urban areas, land use varies from industrial, commercial and residential uses to streets and parking. The runoff from residential properties is a function of the size of the individual stands. The smaller the stand, the more intensive usually is its development and the house on the stand usually covers a greater proportion of the area of the stand.

Table 24.1 US Soil Conservation Service soil groups

Soil group	*Description*
A	Deep sand or loess
B	Shallow loess or sandy loam
C	Clay loams, clayey soil or shallow sandy loam
D	Expansive soils, heavy plastic clays

The slope of the catchment area dictates the speed with which the water drains out of the catchment area. This reduces the time of concentration and hence increases the maximum possible intensity of the rainfall on the catchment. It also reduces the amount of time in which the rain can soak into the soil. In consequence, a steeply sloping catchment area has a higher runoff than a flat catchment. The slope of the catchment is simply the height difference between the highest point of the catchment area and the culvert or bridge that transports the flood water across the road reserve. In short, it is simply the average slope of the catchment area from water shed to culvert.

The runoff from the four soil groups for the different land uses and average slopes is shown in Table 24.2.

24.2.3.2 The rainfall intensity, i

The third term in the rational formula is derived from intensity–duration–frequency data. These are normally provided by the agency responsible for hydrological information pertaining to the area for which these calculations are to be carried out. These data may be provided in the form of equations, graphs or tables. If provided as equations, they can be incorporated into spreadsheets for ease of calculation. When in the form of graphs, a family of curves for different storm durations ranging from as little as 5 minutes up to several days are plotted against recurrence intervals, usually on a logarithmic scale from 0.1 hour up to 100 years, to read off the dependent variable of precipitation, which may be quoted either in inches or millimetres. The recurrence interval, or return period, is usually a policy decision by the responsible road agency, for example, as briefly discussed earlier regarding the return periods for different sizes of drainage structures from small culverts to bridges.

24.2.3.3 The time of concentration

The rainfall intensity is a function of the duration of the storm and the duration curve is selected on the basis of the time of concentration of the storm. The time of concentration is the length of time it takes a drop of water to move to the proposed culvert site from the point of the catchment area furthest away from it. This means that the entire catchment area contributes to the flood reaching the structure. This is the maximum amount of water that can flow from the catchment area because, if the storm has a duration longer than the time of concentration, the intensity of rainfall will be lower. Similarly, if the storm duration is shorter than the time of concentration, the contribution to the flood at the drainage structure will not be from the entire catchment area so that flood reaching the structure will once again be lower than the maximum.

The time of concentration, T_C, is the sum of three flow components:

- Sheet flow
- Shallow concentrated flow
- Channel flow

so that

$$T_C = T_{Sheet} + T_{Shallow} + T_{Channel}$$

Sheet flow occurs in the upper reaches of a watershed and is seldom longer than about 100 metres. This is followed by shallow concentrated flow in small runnels or rills cut into the

Table 24.2 Runoff coefficient, C

	Soil group A			*Soil group B*			*Soil group C*			*Soil group D*		
Slope	<2%	2%–6%	>6%	<2%	2%–6%	>6%	<2%	2%–6%	>6%	<2%	2%–6%	>6%
Land use												
Forest	0.08	0.11	0.14	0.10	0.14	0.18	0.12	0.16	0.20	0.15	0.20	0.25
Bush	0.14	0.22	0.30	0.20	0.28	0.37	0.26	0.35	0.44	0.30	0.40	0.50
Pasture	0.15	0.25	0.37	0.23	0.34	0.45	0.30	0.42	0.52	0.37	0.50	0.62
Farmland	0.14	0.18	0.22	0.16	0.21	0.28	0.20	0.25	0.34	0.24	0.29	0.41
Industrial	0.85	0.85	0.86	0.85	0.86	0.86	0.86	0.86	0.87	0.86	0.86	0.88
Commercial	0.88	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.90	0.90	0.89	0.90
Streets, road reserve	0.76	0.77	0.79	0.80	0.82	0.84	0.84	0.85	0.89	0.95	0.91	0.95
Parking areas	0.95	0.96	0.67	0.95	0.96	0.97	0.95	0.96	0.97	0.97	0.96	0.97
Residential												
0.4 ha	0.22	0.26	0.29	0.24	0.28	0.34	0.28	0.32	0.40	0.46	0.35	0.46
0.2 ha	0.25	0.29	0.32	0.28	0.32	0.36	0.31	0.35	0.42	0.46	0.38	0.46
0.1 ha	0.30	0.34	0.37	0.33	0.37	0.42	0.36	0.40	0.47	0.52	0.42	0.52
0.05 ha	0.33	0.37	0.40	0.35	0.39	0.44	0.38	0.42	0.49	0.54	0.45	0.54

Source: Bengston HH. *Rational method hydrologic calculations with Excel*. Continuing Education and Development, Inc., New York, 2011.

Note: 0.4 hectares (ha) = 1 acre.

soil by running water. Following these, the storm water ultimately discharges into natural or manmade channels or an underground drainage system as discussed in the following section. These components of the time of concentration are expressed as

$$\text{Sheet flow: } T_{\text{Sheet}} = \frac{0.007(nL)^{0.8}}{(P_2)^{0.5}S^{0.4}}$$

$$\text{Shallow concentrated flow: } T_{\text{Shallow}} = \frac{L}{3600\,V}$$

where the flow speed has a typical value of $V = 16.13S^{0.5}$ and

$$\text{Channel flow: } T_{\text{Channel}} = \frac{L}{3600\,V}$$

where

L = length of flow measured along the centreline of flow (m)
n = Manning's roughness coefficient
P_2 = 2-year return period 24-hour duration precipitation for the relevant geographical region (mm)
S = average ground slope (m vertical/m horizontal)
T = time of concentration (h)
V = average flow speed for each flow regime (m/s)

The roughness coefficient for overland flow is shown in Table 24.3.

Table 24.3 Manning's roughness coefficient for overland flow

Surface description	*Roughness coefficient,* n
Smooth asphalt	0.011
Smooth concrete	0.012
Ordinary concrete lining	0.013
Good wood	0.014
Brick with cement water	0.014
Vitrified clay	0.015
Cast iron	0.015
Corrugated metal pipe	0.024
Cement rubble surface	0.024
Fallow	0.050
Cultivated soil	
Cover <20%	0.060
Cover >20%	0.170
Range (natural)	0.130
Grass	
Short	0.150
Dense	0.240
Bermuda	0.41
Woods[a]	
Light underbrush	0.400
Dense underbrush	0.800

[a] When selecting *n*, it is only cover up to a height of about 30 millimetres that would obstruct sheet flow.

24.3 DRAINAGE OF THE ROAD SURFACE

The quantity of water that flows off a road surface during a storm is not trivial. In the interests of road safety, specifically the avoidance of hydroplaning and wheels locking in an emergency braking situation, water must not be allowed to accumulate on the road surface. Spray generated by a leading or passing vehicles temporarily blinds the driver and the greater the depth of water, the greater the extent of spray. Water should thus be removed from the road surface as expeditiously as possible (Figure 24.2).

In rural areas, it is possible to discharge storm water directly onto the surrounding terrain with due cognisance of the need to avoid erosion. If embankment material is erodible, the water flow should be constrained by berms and discharged down the embankment periodically via paved chutes. In urban areas, the road reserve serves as the conduit draining water from the surrounding area so that this option is not available. The water must be conducted by kerbs and channels to points where it can be discharged into an underground piped system and thence to natural watercourses.

According to the *South African Drainage Manual* (2013), the depth of water likely to develop on a paved surface can be calculated by following a three-step process consisting of the determination of

1. The direction of flow, otherwise known as the energy slope, given by

   $$S_f = \sqrt{n_1^2 + n_2^2}$$

2. The flow path length expressed as $L_f = W \;\; 1 + \frac{n_2^2}{n_1^2}$ and combining these into
3. Calculation of the depth of flow $d = (4.6 \times 10^{-2})(L_f I)^{0.5} \left(S_f^{-0.2}\right)$

where

S_f = direction of flow
n_1 = road crossfall per cent
n_2 = gradient per cent
W = road width (m)
I = intensity of rainfall (mm/h)

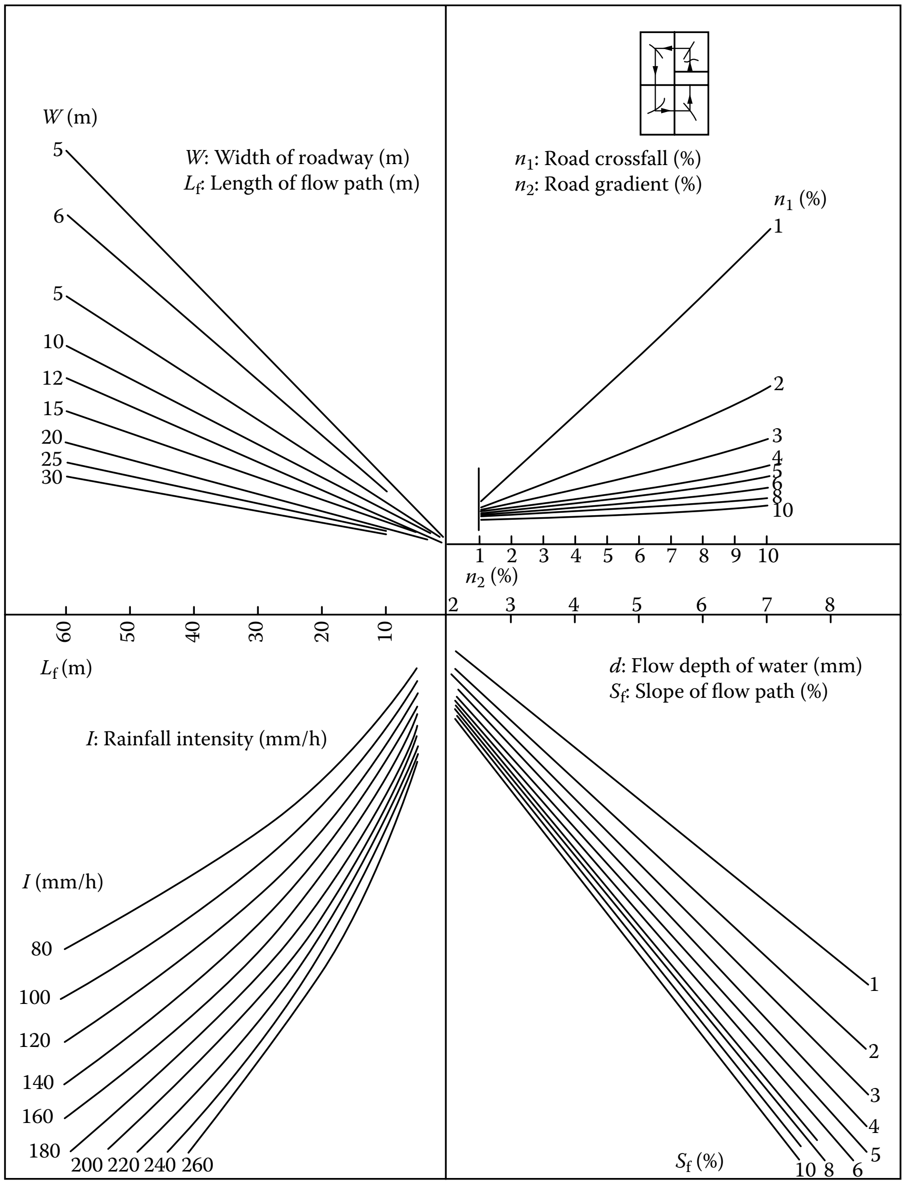
As an example: the depth of the water resulting from a storm with a flow intensity of 100 mm/h on a 10 metre wide road surface where the gradient is 6 per cent and the crossfall is 2 per cent is to be calculated.

$$S_f = \sqrt{2^2 + 6^2} = 6.32 \text{ per cent}$$

$$L_f = 10 \;\; 1 + \frac{6^2}{2^2} \;\; ^{\frac{1}{2}} = 31.62$$

$$d = 4.6 \times 10^{-2}(31.62 \times 100)^{0.5}(0.0632)^{-0.2} = 4.49 \text{ millimetres}$$

The flow depth generated by a rainstorm should not exceed 6 millimetres if hydroplaning is to be avoided. The minimum slope of the road surface along the line of flow should, at least, be 2.0 per cent and, preferably, 2.5 per cent to ensure, if at all possible, that this depth



$L_f = W(1+(\frac{n_2}{n_1})^2)^{0.5}$

$d = 4.6 \times 10^{-2}\,(L_f I)^{0.5}\,(S_f)^{-0.2}$

$W\ L_f$ in m: d in mm
n_1, n_2 in %: I in mm/h
S_f in m/m

Figure 24.2 Depth of water on a road surface. (From South African National Road Agency Limited [SANRAL]. *Drainage manual*, 6th ed. Pretoria, 2013.)

is not exceeded. Where the gradient of the road is 0 per cent, it follows that the camber or crossfall should at least have the selected minimum value. The development of superelevation will cause a flat spot on the road surface but this is unavoidable. Similarly, the contours of the road surface in intersection areas should also be carefully checked to avoid flat spots as far as possible.

24.4 RURAL DRAINAGE

24.4.1 Introduction

Rural drainage is normally by means of unpaved open drains and is focussed on the removal of storm water from the road surface in the first instance followed by drainage of this water plus the water that fell in the verges from the road reserve onto the adjacent land. Alternatively, the water could be contained in a drain within the road reserve and transported to a natural water course.

24.4.2 Scour and silting

Drains may either silt or scour depending on their gradient and hence the speed of flow within them. As a rough guide, silting is likely to occur at flow speeds of less than 0.6 m/s. Unpaved drains should thus not be flatter than about 0.5 per cent because, at such gradients, self-cleansing flow speeds are likely not to be achieved.

Paved drains have a lower coefficient of friction than unpaved drains so that self-cleansing speeds can be achieved at gradients that are almost flat. It is recommended that minimum slopes should be not less than 0.3 per cent. Practical experience suggests that it is difficult to construct a drain accurately to the tolerance required by a gradient this flat. Local imperfections may thus cause silting of an otherwise adequate drain. The problem with silting is that water already captured by the drain could fill to a greater depth and ultimately flow back onto the road surface.

Scour can be equally inimical to road safety. A badly scoured side drain could become so deep that an out-of-control passenger car crossing it could dig into the far bank of the drain and somersault, with possibly fatal consequences. Scour results from a combination of flow speed and material characteristics. Maximum permissible flow speeds are shown in Table 24.4 for various material types.

If the flow speed is likely to cause scour, the problem can be resolved by resorting to paving of the drain. Not being susceptible to scour, paved drains can have a V-profile. Self-cleansing speeds are thus achieved at relatively small flows and the need for maintenance reduced.

An alternative to paving is to reduce the flow speed by the construction of a series of weirs across an unpaved drain. In effect, the drain becomes a series of stilling basins. This does have the disadvantage that water flowing over the weir could result in a deep area of local erosion immediately downstream of the weir. This option should thus be employed with care and limited to low-volume roads where possible. Furthermore, the spacing between

Table 24.4 Scour speeds for various materials

Material	*Maximum permissible flow speed (m/s)*
Fine sand	0.6
Loam	0.9
Clay	1.2
Gravel	1.5
Soft shale	1.8
Hard shale	2.4
Hard rock	4.5

Source: South African National Road Agency Limited (SANRAL). *Geometric design guidelines*. Pretoria, 2002.

successive weirs should be as long as possible, meanwhile ensuring that scour speeds do not occur between weirs.

24.4.3 Selection of drainage structures

It may be possible to select almost any height and type of structure if the height of the gradeline could be such that it naturally allows this choice without being forced. In flat terrain, the gradeline would tend to be close to ground level. The choice of the dimensions of the required drainage structure would then be constrained if a series of unsightly humps where the road is lifted over pipe or box culverts is to be avoided. While a flow of storm water could be accommodated by a single large diameter pipe, a number of pipes of a smaller diameter could accommodate the same flow. It may even be necessary to select the dimensions of a series of low-height box culverts to the same end, as these could be even lower than a small diameter pipe. In short, while the hydrology of the catchment area will indicate the magnitude of the flow to be accommodated, the designer does have a choice regarding the means whereby it can best be accommodated. To exercise this choice properly, the geometric designer would need to have some knowledge of hydraulics to be able to calculate the sizes of the drainage structures appropriate to the circumstances.

The flow speed through a drainage structure can be calculated by use of the Manning formula:

$$V = \frac{1}{n} R^{0.67} S^{0.5}$$

where

V = cross-sectional average flow speed (m/s)
n = Manning coefficient of roughness
R = hydraulic radius (m)
S = slope of channel or pipe

Values of the Manning coefficient of roughness for channels and pipes are offered in Table 24.5. The hydraulic radius is the quotient of the cross-sectional area of the flow and the wetted perimeter of the channel or pipe. The flow in cumecs is then calculated as the product of cross-sectional area and the flow speed.

24.4.4 Roadside safety

Previously, storm water management was focussed on protection of the design layers of the road. In consequence, a *minimum* depth of drainage channel was specified, typically of the order of 600 millimetres or more. With the greater emphasis on roadside safety now

Table 24.5 Maximum suggested spacing of mitre banks

Slope of side drain (%)	*Spacing (m)*
4	200
6	150
8	120
10	80
12	40

universally adopted, it was realised that these depths were potentially lethal to road users and a *maximum* depth of about 300 millimetres has become the norm. It follows that the width of the channel has to be increased to accommodate an unaltered flow of water because the need to keep the design layers unsaturated has not changed. In consequence, the channel generally has a trapezoidal profile that also supports ease of maintenance. Also in the interests of road safety, the sides of the side drain should have slopes of 1*V*:4*H* or flatter. This will enable the passage of an out-of-control vehicle through the drain in relative safety provided the angle at which the drain is being traversed is of the order of 10 to 12 degrees, it being noted that the angle at which vehicles that are out of control leave the road is normally of this magnitude.

24.4.5 The components of rural drainage systems

Longitudinal channels that are located outside the earthworks prism, that is, at the bottom of an embankment slope, for example, are referred to as side drains and, when inside the edge of the shoulder, are referred to as edge drains.

Provided the material from which the embankment is constructed is not erodible, water can simply be allowed to flow over the shoulder edge and down the embankment slope. In cuts, the side drain would be located outside the shoulder breakpoint and be constructed to a depth not greater than 300 millimetres below the level of the shoulder breakpoint.

Edge drains are usually used only in combination with guardrails. This is because guardrails serve as points of concentration of water, with these concentrations being prone to create gullies down a fill slope. The edge drains are normally raised rather than depressed to ensure that passenger car wheels don't get snagged under the guardrail.

Edge drains may either be kerbs or berms, with the latter being a small ridge constructed either of concrete or of bitumen. Kerbs are seldom used on freeways because they may be a safety hazard if not built in conjunction with guardrails and are, in any event, expensive.

Side drains and edge drains are illustrated in Figure 24.3.

Cut faces are protected by catchwater drains. These drains take the form of raised berms and protect the cut face from overland flow from outside the road reserve, which could cause the face to erode. The raised structure of the catchwater drain is because, if the *in situ* material is retained in its natural condition, that is, grassed, the possibility of the catchwater drain itself being eroded would be minimised. Even if the cut is through material that is not likely to erode, the catchwater drain would reduce the quantity of water that would have to be removed by the side drain at the bottom of the cut slope.

The profile of a typical catchwater drain is shown in Figure 24.4.

The medians of rural dual carriageways are normally depressed with gently rounded bottoms. In fact, the rounding could be virtually from shoulder breakpoint to shoulder breakpoint. These median drains feed into underground pipes that transport water away from the road.

Chutes are used to convey storm water runoff down an embankment that would otherwise erode without this protection. They are normally constructed of half round precast concrete channels. If the embankment is of any height, say higher than about 2 metres, the flow reaching the bottom of the embankment would be highly energised and thus would certainly create a gully of monstrous proportion, possibly leading to the destruction of the embankment itself. A chute should thus debouch into a stilling basin that would serve two functions. It would destroy the kinematic energy of the streams of water leaving the chute and would also spread the water over a wider area. The latter function would cause the flow of water to be shallower and thus slower flowing, reducing the probability of erosion still further.

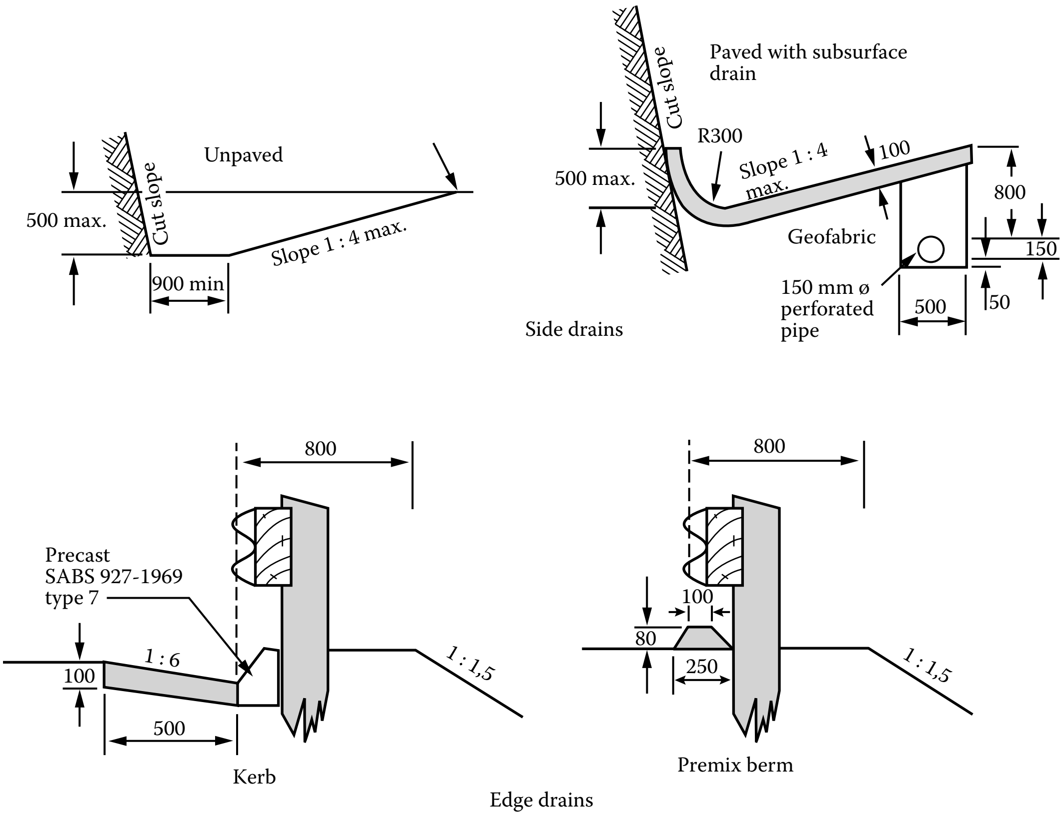


Figure 24.3 Typical profiles of side and edge drains.

Mitre banks are used to deflect storm water out of a side drain into the surrounding land use. They are thus built at an angle normally of about 45° or less to the centreline of the road. The selected angle is dependent on the contours of the area into which the water flow is directed. If the selected angle causes the mitre bank to be parallel to the contours, the flow volume would obviously arise only from the depth of flow where the side drain meets

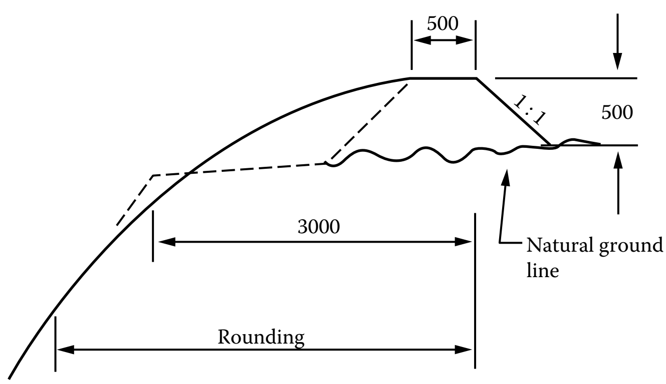
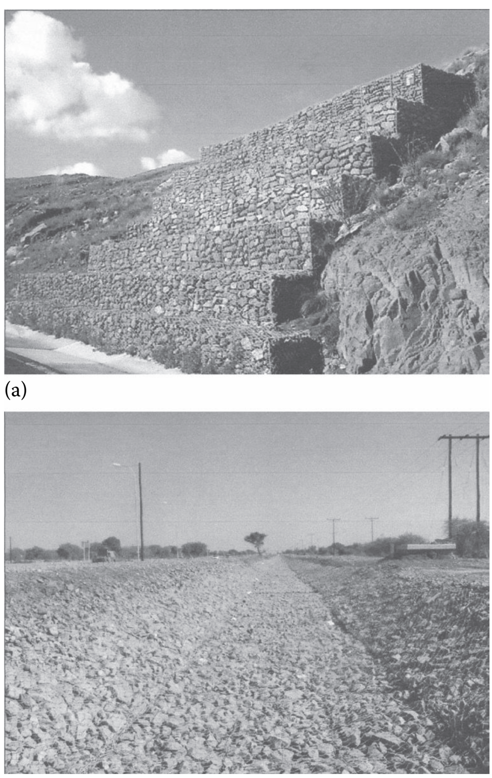


Figure 24.4 Typical profile of a catchwater drain.

Table 24.6 Manning's coefficient of roughness for pipes and channels

Conduit material	*Manning's coefficient of roughness*
Closed conduits	
Concrete pipes	0.010–0 015
Corrugated metal pipes	0.011–0.037
Plastic pipes (smooth)	0.009–0.015
Plastic pipes (corrugated)	0.018–0.025
Open channels	
Kerb and channel	0.012–0.016
Concrete	0.011–0.015
Rubble or rip rap	0.020–0.036
Vegetation	0.020–0.150
Bare soil	0.016–0.025
Rock cut	0.025–0.045
Natural channels, top width <30 m at flood stage	
Fairly regular section	0.025–0.060
Irregular section with pools	0.040–0.150

Note: The lower of the two values for every conduit material apply to well-constructed and smoother pipes and channels.



(a)

(b)

Figure 24.5 Typical gabions. (a) Gabion for bank protection and (b) gabion for erosion protection.

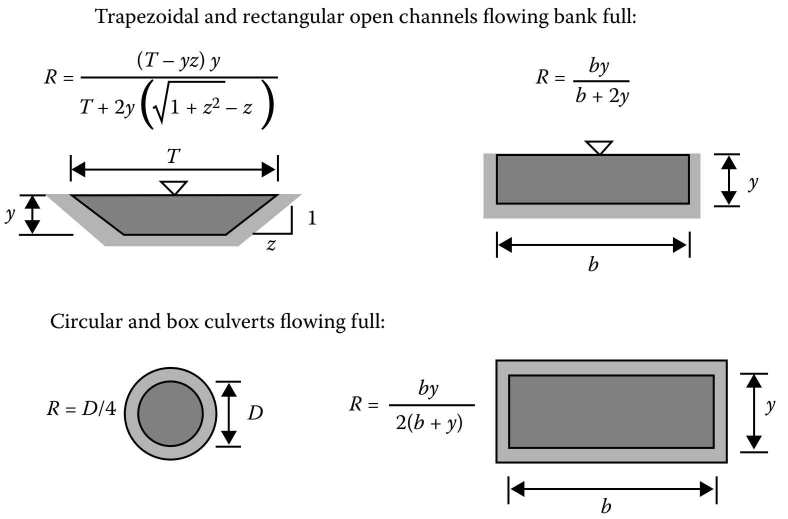


Figure 24.6 Hydraulic radius.

the mitre bank. The mitre bank should thus be angled across the contours to have a slope that would cause a flow speed of less than the scour velocities reflected in Table 24.3. The desirable gradient of the mitre bank should be at about 2 per cent, as a slope flatter than this would probably become silted up whereas, at a gradient in excess of 5 per cent, erosion becomes a possibility. The mitre bank should be so constructed that it becomes shallower and shallower until its invert is at ground level.

Water should be drained out of the road reserve as frequently as possible and the spacing of the mitre banks should be as suggested in Table 24.6 (Ethiopian Roads Authority, 2011).

Seeing that mitre banks force a change of direction on the flow of water, they would scour if constructed of soil. It is thus necessary to protect them by means, for example, of stone pitching. The stone pitching would have to extend below the ground surface to ensure that it is not undercut and the mitre bank destroyed.

Gabions are a form of stone pitching in which the stones are packed into wire cages as shown in Figure 24.5 (see also Figure 24.6).

24.5 URBAN DRAINAGE SYSTEMS

24.5.1 Introduction

Drainage of urban areas is towards rather than away from the road reserve. This is because paved road surfaces have a runoff of 100 per cent and the land usage in urban areas is characterised by intensive development typically also having a high runoff. This water would have to be transported to a channelised drainage system as soon as possible and the obvious location of this system would be in the road reserve. Urban drainage entails the protection of road users against the effects of large quantities of water flowing into the road reserve from adjacent properties.

24.5.2 Major and minor storms

A distinction is drawn between accommodation of major and minor storms (SANRAL, 2002). Historically, the focus of design of storm water management systems was limited to the minor storm but this is no longer acceptable. Major storms generate significant volumes of runoff and the damage they can cause is considerable. This arises from the depth and speed of flow. The basic requirements differ between major and minor storm water management systems. Statistically, the longer the duration of a storm the lower is the likely intensity of rainfall and hence the extent of the runoff. Major storms are thus dealt with by increasing the time of concentration and slowing the resulting flow whereas minor storms are best accommodated by rapid removal of the flood water. This dichotomy suggests that two separate but allied drainage systems should be provided.

24.5.3 The components of urban drainage systems

Storm water is drained by kerbs and channels discharging into underground connector pipes that ultimately connect to sewer mains. The major storms dictate the spacing of inlets into the connector pipes, with the spacing of these outlets being such that prescribed flow depths, as illustrated in Table 24.7, are not exceeded (see also Figure 24.7).

24.5.4 Grid inlets

Water flowing alongside a kerb is captured by grid inlets, also referred to sometimes as drop inlets. On steep gradients, a considerable amount of water flows past the grid inlet until the point is reached where the kerb is overtopped.

24.5.5 Discharge pipes

Discharge pipes serve as connectors from grid inlets to storm water mains. Where connectors are parallel to the road centreline, they should not be placed under the travelled way because this creates problems of maintenance, principally placing workmen at risk from passing traffic. Every time the road is resealed, it follows that the manholes would have to be raised to, once again, be flush with the road surface.

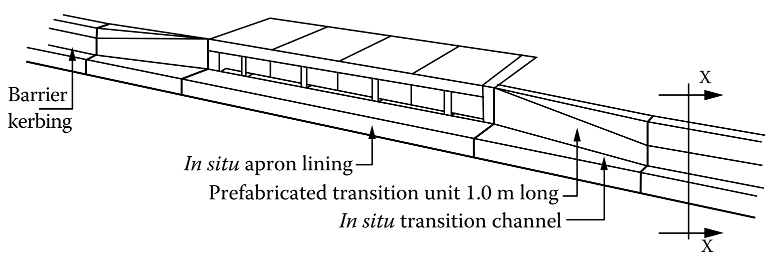
Manholes should be located at a spacing of 200 to 350 metres for pipes that have a diameter of 1200 millimetres or more and at 100 to 200 metre spacing for pipes with a diameter of less than 1200 millimetres. For ease of maintenance, manholes should also be provided at

- Changes of directions that are greater than about 10°
- Changes in the size of pipes
- All junctions between pipes
- Slope reductions

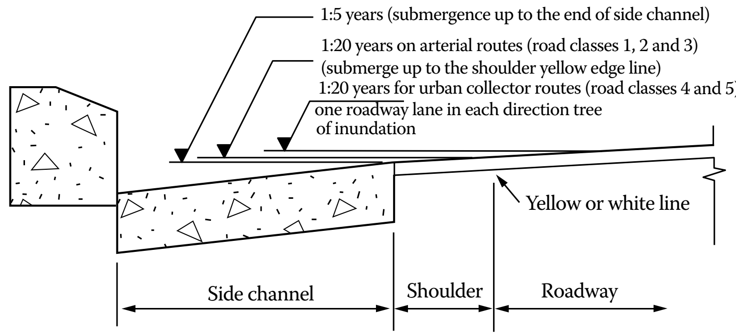
Table 24.7 Maximum permissible depths of flow alongside surfaced roads

Type of road	*Maximum encroachment*
Residential and lower order streets	No kerb overtopping[a] Flow may spread to top of camber
Residential access collector	No kerb overtopping[a] Flow spread must leave at least one traffic lane width free of water
Local distributor	No kerb overtopping[a] Flow spread must leave at least one lane in each direction free of water
Higher order roads	No encroachment is permissible on any traffic lane

[a] Encroachment onto sidewalks is not acceptable.



General layout of a side channel and side outlet



*Note:
Maximum depth of rainwater
flow on roadway outside
side drain is 6 mm

Cross-section XX through side drain

Figure 24.7 Maximum permissible depths of flow alongside surfaced roads.

Slope reductions should be avoided if at all possible. Reductions in slope can result in reductions in flow speed. In consequence, material particles of a size that could be transported along the steeper slope would settle to the invert of the pipe at the flatter slope. This silting could result in the flatter pipe being completely blocked.

24.5.6 Retention and detention ponds

Major storms dictate the size of the sewer mains and have an added component – retention or detention ponds. The difference between these ponds is essentially a matter of degree. Retention ponds are intended to capture all the runoff from the design storm but with an overflow to accommodate storms of a still greater severity. Detention ponds are limited to capturing sufficient storm water to increase the time of concentration of the catchment area at least to that that prevailed prior to its urbanisation.

Retention ponds require significantly sized areas such as parks or golf courses. Where the development of the urban area is intensive, such as in commercial areas, the space required by these ponds would, in all probability, not be available. The only option then is to make use of detention ponds.

The runoff from streets and parking areas includes suspended solids, heavy metals, nutrients and organic material and, as suggested above, water running off a road surface after a long dry period is more polluted than anything found in a foul water sewer. This quality of water may have a significant impact on the quality of water in the streams into which the runoff from the road system is ultimately discharged. Drainage system components that can be used to improve the quality of the runoff water include detention ponds, wetlands, infiltration basins, grassing of side drains and detention ponds and sand filters.

24.6 DRAINAGE SYSTEMS FOR LOW-VOLUME ROADS

Most roads in Third World countries fall into the category of low-volume roads, that is, with traffic counts of fewer than 400 vehicles per day. Very often, they are typically built using labour-intensive methods as a means of job creation. Excavation is by picks and shovels, material transport by wheel barrows and compaction by hand stampers with further compaction by means of passing traffic. For this reason it is sensible to reduce earthworks volumes as far as possible and gradeline levels are often just about at ground level.

The benefit of having this low a gradeline is that storm water does not get concentrated to flow through a culvert and remains spread. Flow speeds are low and water flowing across the road is not likely to cause damage except where it flows over the downhill road edge. This road edge needs to be protected either by having a slope of about $1V{:}6H$ or by means of stone pitching.

Culverts are replaced by drifts or fords and low level causeways. Drifts are designed for water to flow over the road surface and it is not expected that vehicles can use the roads at all times. They will simply have to wait at the side of the river until the flow depth has reduced to the extent when the drift becomes navigable once more. The level of the drift should be as close as possible to the river bed level and the normal depth of the water should not exceed 150 millimetres. Some cars will float in water with a depth of only 300 millimetres. The maximum 5-year flow should not exceed 6 cumecs, as this would be sufficient for a passenger car to be washed downstream. Approaches to a drift should have a roughened concrete surface at a gradient not exceeding 10 per cent extending from the river bank to the level of the drift, which should also be constructed of concrete. If there are many heavy vehicles in the traffic stream the gradient of the approaches should be 7 per cent or less.

Low-level causeways normally comprise a battery of several low-height box culverts intended to accommodate only the perennial flow of water in the river or stream. Storm water will always overtop these causeways.

References

American Association of State Highway and Transportation Officials (AASHTO). *A policy on geometric design of highways and streets*. Washington, DC, 2001.

American Association of State Highway and Transportation Officials (AASHTO). *A policy on geometric design of highways and streets*. Washington, DC, 2004.

American Association of State Highway and Transportation Officials (AASHTO). Barrier design and placement. In *Overview of AASHTO roadside design guide design process*. Washington, DC, 2005.

American Association of State Highway and Transportation Officials (AASHTO). *Manual for assessing safety hardware*. Washington, DC, 2009.

American Association of State Highway and Transportation Officials (AASHTO). *A policy on geometric design of highways and streets*. Washington, DC, 2011a.

American Association of State Highway and Transportation Officials (AASHTO). *Roadside design guide*, 4th ed. Washington, DC, 2011b.

American Association of State Highway and Transportation Officials (AASHTO). *Guide for the development of bicycle facilities*, 4th ed. Washington, DC, 2012.

Ang AHS and Tang WH. *Probability concepts in engineering planning and design*. John Wiley & Sons, New York, 1975.

Austroads. *Guide to road design, Part 3: Geometric design*. Sydney, 2009a.

Austroads. *Guide to road safety, Part 6: Road safety audit*. Sydney, 2009b.

Bared JA. *Drivers' evaluation of the diverging diamond interchange*. Techbrief FHWA-HRT-07-048, Federal Highway Administration, Washington, DC, 2007.

Barton D. *Guide to road design, Part 3: Geometric design.* Report AGRD03/09, Austroads, Sydney, 2009.

Bengston HH. *Rational method hydrologic calculations with Excel*. Continuing Education and Development, Inc., New York, 2011.

Bligh R, Miaou S-P, Lord D and Cooner S. *Median barrier guidelines for Texas*. Texas Department of Transportation, Austin, 2006.

Boodlal L. *Accessible sidewalks and street crossings*. Federal Highway Administration, Washington, DC, 2003.

Botswana Department of Transport. *Road design manual*. Gaborone, 2012.

Box PC and Oppenlander JC. *Manual of traffic engineering studies*, 4th ed. Institution of Transportation Engineers, Arlington, VA, 1976.

Burrell RC, Mitchell MF and Wolhuter KM. *Geometric design guidelines*. South African National Road Agency Limited (SANRAL), Pretoria, 2002.

California Department of Transportation (Caltrans). *Traffic safety systems*. Sacramento, 2012.

Campbell JL, Richard CM and Graham J. *Human factors guidelines for road systems*. NCHRP Report 600, Transportation Research Board, Washington, DC, 2008.

Caulfield B. *Topic 6: Angle measurement: Intersection and resection.* Trinity College, Dublin, 2010a.

Caulfield B. *Topic 9: Curves and transition curves*. Trinity College, Dublin, 2010b.

Caulfield B. *Topic 8: Horizontal curves*. Trinity College, Dublin, 2012.

Churchward C, Palmer JF, Nasasauer JI and Swanwick CA. *Evaluation of visual impact methodologies*. NCHRP Report 741, Transportation Research Board, Washington, DC, 2013.

Collins HJ and Hart CA. *Principles of road engineering*. Arnold Publishers, London, 1935.
Cox RL and Arndt O. *Using an extended design domain concept for road restoration projects*. 3rd International Symposium on Highway Geometric Design. Transportation Research Board, Chicago, 2005.
Cramer PC. *New approaches to ecological surveys: A synthesis of highway practice*. NCHRP Synthesis 400, Transportation Research Board, Washington, DC, 2009.
Daniello A and Gabler HC. *Effect of barrier type on injury severity in motor cycle-to-barrier collisions in North Carolina, Texas and New Jersey*. Transportation Research Record 2262, Transportation Research Board, Washington, DC, 2011.
Department of Infrastructure Energy and Resources, Transport Infrastructure Branch. *Road safety barriers design guide: Part A*. Hobart, Tasmania, 2006.
Egebjerg U et al. *Beautiful roads – A handbook of road architecture*. Danish Roads Directorate, Copenhagen, 2002.
Ethiopian Roads Authority. *Design manual for low volume roads*. Addis Ababa, 2011.
Fambro DB, Fitzpatrick K and Koppa RJ. *Determination of stopping sight distances*. NCHRP Report 400, Transportation Research Board, Washington, DC, 1997.
Federal Highway Administration. *Road safety audit manual*. Washington, DC, 2006.
Federal Highway Administration. *Managed highways: A primer*. Washington, DC, 2008.
Federal Highway Administration. *Manual on uniform traffic control devices for streets and highways*. Washington, DC, 2009.
Fitzgerald WJ. *W-Beam guardrail repair: A guide for highway and street maintenance personnel*. Federal Highway Administration, Washington, DC, 2008.
Fitzpatrick K, Carlson P, Brewer MA Wooldridge MD and Miaou S. *Design speed, operating speed and posted speed practices*. NCHRP Report 504, Transportation Research Board, Washington, DC, 2003.
Fitzpatrick K and Eun Sug Park. *Safety effectiveness of the HAWK pedestrian crossing treatment*. Federal Highway Administration, Washington, DC, 2010.
Fitzpatrick K and Wolff T. *Left-turn lane installation guidelines*. 2nd Urban Street Symposium, Transportation Research Board, Washington, DC, 2003.
Fitzpatrick K, Wooldridge MD and Blaschke JD. *Urban intersection design guide*, Vol. 1: *Guidelines*. Texas Transportation Institute, College Station, 2005.
Florida Department of Transportation. *Florida intersection design guide*. Tallahassee, 2007.
Fosgerau M, Hjorth K and Lyk-Jensen S. *The Danish value of time study*. Danish Transport Research Institute, Copenhagen, 2007.
Freund JE. *Mathematical statistics*, 2nd ed. Prentice/Hall International, London, 1972.
Frith B. *Economic valuation of the impact of safe speeds: Literature review*. New Zealand Transport Agency, Wellington, 2012.
Gartner NH, Messer CJ and Rathi AK. *Traffic flow theory: A monograph*. Transportation Research Board, Washington, DC, 2001.
Gattis JL, Bryant M and Duncan LK. *Acceleration lane design for higher truck volumes*. Mack-Blackwell Transportation Centre, University of Arkansas, Fayetteville, 2008.
Gauteng Department of Public Transport, Roads and Works. *Road design manual*. Volume 1: Geometrics. Johannesburg, 2007.
Gerlough DL and Huber MJ. *Statistics with application to highway traffic analyses*, 2nd ed. Eno Foundation for Transportation, Westport, CT, 1978.
Giuffrè O, Guerrieri M and Granà A. *Turbo roundabout general design criteria and functional principles: Case studies from the real world*. 88th Annual Meeting, Transportation Research Board, Washington, DC, 2009.
Grant EL, Ireson WG and Leavenworth RS. *Principles of engineering economy*, 7th ed. John Wiley & Sons, New York, 1982.
Hakamies-Blomqvist L, Sirén A and Davidse R. *Older drivers – A review*. VTI report 497A – 2004. Swedish National Road and Transport Institute, Linköping, 2004.
Hall LE, Powers RD, Turner DS, Brilon W and Hall JW. *Overview of cross-section design elements*. Transportation Research Board, International Symposium on Highway Geometric Design Practices, Boston, 1995.

Harmelink M. *Volume warrants for left-turn storage lanes at unsignalised at-grade intersections*. Highway Research Record 211, Highway Research Board, Washington, DC, 1967.

Harwood DW, Gilmore DK, Richard KR, Dunn JM and Sun C. *Passing sight distance criteria*. NCHRP Report 605, Transportation Research Board, Washington, DC, 2007.

Hauer E. *Safety in geometric design standards II: Rifts, roots and reform*. 2nd International Symposium on Highway Geometric Design. Transportation Research Board, Mainz, 2000.

Highways Agency et al. *Design manual for roads and bridges*, Vol. 6 Sect. 1 Part 1. *Highway link design*. London, 2002.

Hughes W, Jaganathan R, Sengupta C and Hummer J. *Alternative intersections/interchanges: Informational report*. Report FHWA-HRT-09-060, Federal Highway Administration, Washington, DC, 2010.

Hummer JE. *Introduction to and update on unconventional interchanges*. Raleigh, NC, 2010.

Illinois Bureau of Public Roads and Streets. *Geometric design manual*. Springfield, 2009.

Illinois Department of Transportation, Bureau of Design and Environment. Interchange type and design studies. *Design and environmental manual*. Springfield, 2010.

Institute of Transportation Engineers. *Trip generation manual*. Washington, DC, 2012.

Iowa Department of Transport. *Geometric design manual*. Des Moines, 2004.

Irish National Roads Authority. *Design manual for roads and bridges*. Part 2, Vol. 6, Section 1, NRA TD 27/07. Dublin, 2007.

Jones TE and Petts RC. *Maintenance of minor roads using the length-man system*. Fifth International Conference on Low-Volume Roads, Raleigh, NC, 1991.

Jordaan PW. *Peaking characteristics of rural road traffic*. Doctoral dissertation, Pretoria University, 1985.

Kikuchi S, Felsen M, Mangalpally S and Gupta A. *Trip attraction rates of shopping centres in Northern New Castle County, Delaware*. Delaware Centre for Transportation, Newark, 2004.

Klein LA, Mills MK and Gibson DRP. *Traffic detector handbook: Third edition – Volume I*. Federal Highway Administration, Washington, DC, 2006.

Kühn RD. *Greenshield's legacy: Highway traffic*. Transportation Research Circular E-C140, Transportation Research Board, Washington, DC, 2011.

Kuzmyak JR and Dill J. *Walking and bicycling in the United States: The who, what, where and why*. TR News 280 May–June 2012, Washington, DC, 2012.

Lamm R, Psarianos B and Mailaender T. *Highway design and traffic safety engineering handbook*. McGraw-Hill, New York, 1999.

Lamm R, Wolhuter KM, Beck A and Ruscher T. *Introduction of a new approach to geometric design and safety*. 20th South African Transport Conference, CSIR, Pretoria, 2001.

Lei Y, Yi Q, Mehdi A, Lei G and Chenyan G. *Left-turn lane design and operation*. Texas Department of Transportation, Austin, 2007.

Leisch JP. *Service interchange forms: Adaptability, design features and operational characteristics*. Transportation System Planning, Boise, 2002.

Levine J and Garb Y. *Congestion pricing's condition promise: Promotion of accessibility or mobility*. Pergamon Press, Oxford, 2002.

Linsley RK Jr, Kohler MA and Paulhus JL. *Applied hydrology*. McGraw-Hill, New York, 1949.

Lunenfeld H and Alexander GJ. *A user's guide to positive guidance*. Federal Highway Administration, Washington, DC, 1990.

Markow MJ. *Engineering economic analysis practices for highway investment*. NCHRP Synthesis 424, Transportation Research Board, Washington, DC, 2012.

Massachusetts Department of Transportation. Wild life accommodation. In *Project development and design guide*. Boston, 2006.

Mauroit C, Ghaman R, Kamali F, Kurosaka T, Talens H, Gerlach J, Bezak B and Prat D. *Human powered transport*. World Road Association (PIARC), Paris, 2008.

Maxwell A and Wood K. *Review of traffic signals on high-speed roads*. European Transport Conference, Strasbourg, 2006.

McMahon K and Dahdah S. *The true cost of road crashes: Valuing life and the cost of serious injury*. International Road Assessment Programme (iRAP), Basingstoke, UK, 2008.

Michigan Department of Transportation. 2004.

Miller MA. Bus lanes/bus rapid transit systems on highways: Review of the literature. California Path Program, Berkeley, 2009.
Milton J and St Martin A (eds). *Understanding flexibility in transportation design*. Washington State Department of Transportation, Olympia, 2005.
Ministry of Roads and Bridges, South Sudan. *Low-volume roads design manual.* Juba, 2013.
Moon J-P, Hummer JE and Van Duyn MH. *Introduction of the new nano interchange design as a directional freeway to freeway interchange.* 4th International Symposium on Highway Geometric Design, Transportation Research Board, Valencia, 2010.
National Heart, Lung and Blood Institute. *Body mass index*. Bethesda, MD, undated.
Neumann TR. *Design guidelines for very low-volume local roads*. American Association of State Highway and Transportation Officials (AASHTO), Washington, DC, 2001.
Neumann TR et al. *Thinking beyond the pavement*. Maryland State Highway Administration Workshop, Annapolis, 1998.
New Jersey Department of Transportation. *Roadway design manual.* Trenton, 2011.
New South Wales Road and Traffic Authority. *Traffic signal design*. Sydney, 2011.
Nezamuddin N, Crunkleton J, Tarnoff PJ and Waller ST. *Speed distribution profile of traffic data and sample size estimation*. Traffic and Engineering Control (TEC) Magazine, pp. 147–152, London, 2010.
Office of Economic Co-operation and Development. *Obesity update, June 2014*. Paris, 2014.
Ogden BD. *Railroad-highway grade crossing handbook* – Revised second edition. Office of Safety Design, Federal Highway Administration, Washington, DC, 2007.
Operational Effects of Geometrics Committee. *Modelling operating speed: Synthesis report*. Transportation Research Board, Washington, DC, 2011.
Oregon Department of Transportation. Facility analysis and simulation team. *Modelling performance indicators on two-lane rural highways.* Salem, OR, 2010.
Permanent International Association of Road Congresses (PIARC). *Environmental impact of existing roads*. Montreal, 2000.
Pinard MI, Ellis C-H, Johansen R et al. *Guideline: Low-volume paved roads*. Southern African Development Community, Maputo, 2003.
Plumert J, Kearney J and Cremer J. *How does traffic density influence cyclists' gap choices?* International Conference on Road Safety and Simulation, Rome, 2007.
Polus A, Matter-Habib C, Pollatschek M and Jarroush J. *Comprehensive consistency model and its impact on safety.* 3rd International Symposium on Highway Geometric Design, Transportation Research Board, Chicago, 2005.
Potts IB and Harwood DW. *Benefits and design/location criteria for passing lanes*. Missouri Department of Transportation, Kansas City, 2004.
PWV Consortium. *Roundabouts on provincial roads in Gauteng: Proposed warrants*. Department of Public Transport, Roads and Public Works, Johannesburg, 2002.
Queensland Department of Transport and Main Roads. Chapter 12, *Road planning and design manual.* Brisbane, 2002.
Queensland Department of Transport and Main Roads. Chapter 5, *Road planning and design manual.* Brisbane, 2004.
Queensland Department of Transport and Main Roads. Chapter 8, *Road planning and design manual.* Brisbane, 2005.
Road Traffic Management Corporation. *South African road safety audit manual*. Pretoria, 2012.
Rodegerdts L, Bansen J, Tiesler C et al. *Roundabouts: An informational guide*. National Cooperative Highway Research Program (NCHRP) Report 672, Transportation Research Board, Washington, DC, 2010.
Rodrigues J-P. *The geography of transport systems*. Department of Global Studies and Geography, Hofstra University, New York, 2013.
Roess RD and Ulerio JM. *Capacity of ramp-freeway junctions*. NCHRP Project 3-37, Polytechnic University, Brooklyn, 1993.
Ross HE Jr, Sicking DL and Zimmer RA. *NCHRP Report 350: Recommended procedures for the safety performance evaluation of highway features.* Transportation Research Board, Washington, DC, 1993.

Sampson J. Lagging versus leading green right turn arrows. *Civil Engineering*, vol. 19 (8). South African Institution of Civil Engineering, 19 p., Johannesburg, 2011.

Sampson JD. Queue-based traffic signal warrants: The 4Q/6Q warrant. *Journal of the Institute of Traffic Engineers*, Washington, DC, 1999.

Sampson JD, van As SC, Pauw I and Joubert HS. *Manual on services in road reserves*. South African Committee of Transport Officials (COTO), Pretoria, 2011.

Savolainen PT and Tarko AP. *Safety of intersections on high-speed road segments with superelevation*. Indiana Department of Transportation and Purdue University, West Lafayette, 2004.

Schermers G and Kraay JH. *Road safety audits – Towards a safer road infrastructure*. Workshop 'Road traffic safety: Policy, research implementation and co-operation between South Africa and the Netherlands'. Pretoria, 1999.

Schutt JR, Phillips KL and Landphair HC. *Guidelines for aesthetic design in highway corridors: Tools and treatments for Texas highways*. Texas Department of Transportation, Austin, 2001.

Smiley A, Smahel T and Page L. *Senior drivers and highway design*. Canadian Automobile Association, Ottawa, 2008.

Snare MC. *Dynamic model for predicting maximum and typical acceleration rates of passenger vehicles*. MSc thesis, Virginia Polytechnic Institute, Blacksburg, 2002.

Snell M. *Cost Benefit analysis: A practical guide*, 2nd ed. Thomas Telford, London, 2011.

South African Department of Transport. *Pedestrian and bicycle facility guidelines Draft 1.0*. Pretoria, 2002.

South African Department of Transport. *Pedestrian and bicycle facility guidelines*. Pretoria, 2003.

South African National Road Agency Limited (SANRAL). *Geometric design guidelines*. Pretoria, 2002.

South African National Road Agency Limited (SANRAL). *Drainage manual*, 6th ed. Pretoria, 2013.

Southern African Development Community. *Road traffic signs manual*. Part 3: *Traffic signals design*. Maputo, 2001.

Southern African Transport and Communications Commission (SATCC). *SATCC road and bridge design standards: Geometric design of trunk roads*. Maputo, 1997.

Spath SB. *A comparative analysis of diverging diamond interchanges*. 2008.

Stantec Consulting Limited. *Traffic impact assessment guideline*. Alberta Infrastructure and Transportation Division, Edmonton, 2005.

Staplin L, Lococo K, Byington S and Harkey D. *Guidelines and recommendations to accommodate older drivers and pedestrians*. Report FHWA-RD-01-051, Federal Highway Administration, Washington, DC, 2001.

Stephens LB. *Barrier guide for low volume and low speed roads*. Federal Highway Administration, Washington, DC, 2005.

Steyn AGW, Smit CF and du Toit SHC. *Moderne statistiek vir die praktyk*. JL van Schaik, Pretoria, 1984.

Texas Department of Transportation. *Roadway design manual*. Austin, 2010.

Theeuwes J and Godthelp J. *Self-explaining roads* (IZF 1992 C-8 [in Dutch]). TNO Human Factors Research Institute. Soesterberg, The Netherlands, 1992.

Todini E. *Hydrological catchment modelling – Past, present and future*. Hydrological and Earth Systems Sciences, Göttingen, Germany, 2007.

Transit New Zealand. *State highway geometric design manual*. Wellington, 2002.

Transportation Association of Canada (TAC). *Geometric design guide for Canadian roads*. Ottawa, 1999.

Transportation Research Board. *Highway geometric design*. Mainz, 2000.

Transportation Research Board. *Highway capacity manual 2010*. Washington, DC, 2010.

Tunnard C and Pushkarev B. *Man-made America – Chaos or control?* Yale University Press, New Haven, CT, 1964.

van As SC and Joubert HS. *Traffic flow theory*. University of Pretoria, 1993.

Vuchic VR. *Urban public transportation: Systems and Technology*. Prentice Hall, Upper Saddle River, NJ, 1981.

Wachs M. A very brief history of why Americans hate their commutes. The Atlantic Monthly Group, 2014.

Wardrop JG. *Some theoretical aspects of road traffic research*. Road Paper 36, Part 1, Proceedings of the Institution of Civil Engineering, London, 1952.

Washington State Department of Transportation. *Design manual*. Olympia, 2014.

Watermeyer RB. *Labour-based construction, the development of emerging contractors and the Reconstruction and Development Programme.* Soderland and Schutte, Johannesburg, 2003.

Wepener DA, Engelbrecht RJ and Kruger P. *Manual for traffic impact studies*. Department of Transport, Pretoria, 1995.

Wilson EM and Lipinski ME. *Road safety audits – A synthesis of highway practice.* NCHRP Synthesis 336, Transportation Research Board, Washington, DC, 2004.

Wolhuter KM. *TRH 17: Geometric design of rural roads*. Department of Transport, Pretoria, South Africa, 1988.

Wolhuter KM. *Warrants for climbing lanes*. M.Eng. thesis, University of Pretoria, Pretoria, South Africa, 1990.

Wolhuter KM. Geometric design and layout planning. In *Guidelines for human settlement planning and design*. Department of Housing, Pretoria, South Africa, 2000.

Wolhuter KM. *Lecture T1: Introduction. Geometric Design and Course*. South African Road Federation, Johannesburg, South Africa, 2003a.

Wolhuter KM. *Lecture T6: Survey*. South African Road Federation, Johannesburg, South Africa, 2003b.

Wolhuter KM. *Lecture T7: Economic considerations*. *Geometric Design Course*. South African Road Federation, Johannesburg, South Africa, 2003c.

Wolhuter KM and Garner D. *Spacing of interchanges in Gauteng*. Department of Public Transport, Roads and Works, Johannesburg, 2003.

Wolhuter KM and Garner D. *Elderly drivers: The problem and its solution*. Department of Public Transport, Roads and Works, Johannesburg, 2004.

Wolhuter KM and Garner D. *Interchange spacing in Gauteng*, 3rd International Symposium on Highway Geometric Design, Transportation Research Board, Chicago, 2005.

Wolhuter KM, Garner D and Schmid KG. *Interchange spacing in Gauteng*. Third International Symposium on Highway Geometrics, Transportation Research Board, Chicago, 2005.

Wooldridge M, Fitzpatrick K, Harwood DW, Potts IB, Elefteriadou L and Torbic DJ. *Geometric design consistency on high speed rural two-lane roadways.* NCHRP Report 502, Transportation Research Board, Washington, DC, 2003.

World Health Organization. *Guidelines for community noise.* Geneva, 1999.

World Road Association. *Human factors guideline for safer road infrastructure*. Permanent International Association of Road Congresses (PIARC) Technical Committee, Paris, 2008.

Zhou H and Rouhlamin MP. *Guidelines for reducing wrong-way crashes on freeways.* Illinois Centre for Transportation, Auburn, 2014.

Transport Engineering

Geometric design covers the design of the visible elements of the road—its horizontal and vertical alignments, the cross-section, intersections and interchanges. Good practice allows the smooth and safe flow of traffic as well as easy maintenance. Geometric design is covered in depth. The book also addresses the underpinning disciplines of statistics, traffic flow theory, economic and utility analysis, systems analysis, hydraulics and drainage, capacity analysis, coordinate calculation, environmental issues and public transport.

A key principle is that we need to recognise what the driver wishes to do rather than what the vehicle can do. The book takes a human factors approach to design, drawing on the concept of the 'self-explaining road'. It also emphasises the need for consistency of design and shows how this can be quantified, and sets out the issues of the design domain context, the extended design domain concept and the design exception. The book is not simply an engineering manual, but properly explores context-sensitive design.

Changes in geometric design over the last few years have been dramatic and far-reaching and this is the first book to draw these together into a practical guide which presents a proper and overriding philosophy of design for road and highway designers and students.

Keith Wolhuter is a Fellow of the South African Institution of Civil Engineers and has fifty years of experience in the field, producing national guidelines on geometric design, notably in Southern Africa.

Cover image by Rodney Burrell

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

Y135499
ISBN: 978-0-415-52172-7
90000
9 780415 521727

www.sponpress.com