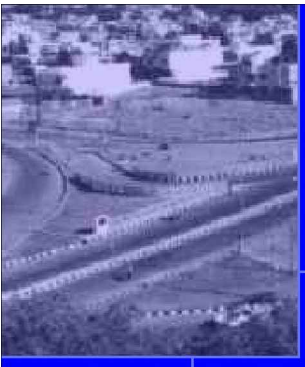


Eastern Economy Edition

PRINCIPLES OF TRANSPORTATION ENGINEERING

PARTHA CHAKROBORTY

ANIMESH DAS

PHI

Principles of Transportation Engineering

Partha Chakroborty
Professor
Department of Civil Engineering
Indian Institute of Technology Kanpur

Animesh Das
Professor
Department of Civil Engineering
Indian Institute of Technology Kanpur

PHI Learning Private Limited
New Delhi - 110 001
2012

PRINCIPLES OF TRANSPORTATION ENGINEERING
Partha Chakroborty and Animesh Das

© 2003 by PHI Learning Private Limited, New Delhi. All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher.

ISBN-978-81-203-2084-0

The export rights of this book are vested solely with the publisher.

Sixth Printing ... ... **July, 2011**

Published by Asoke K. Ghosh, PHI Learning Private Limited, M-97, Connaught Circus, New Delhi-110001 and Printed by Mudrak, 30-A, Patparganj, Delhi-110091.

To my father Durgadas Chakroborty
— Partha Chakroborty

To my mentor Prof. B.B. Pandey
— Animesh Das

Contents

Preface

India is on the threshold of a major forward thrust in the field of transportation infrastructure. Vehicular traffic, in both urban and rural areas, has in recent years increased manifold. The government has also realized that developing the transportation infrastructure is the key to overall development of the country. There is also a general increase in awareness about mobility. These factors, we believe, constitute a recipe for fast development in the area of transportation.

The need for professionals with specialization in transportation engineering, equipped with the knowledge of modern as well as traditional techniques, is bound to grow over the next few years. It is, therefore, not untimely to write a book on transportation engineering covering modern techniques without overlooking the traditional methods. Further, as of today, there is hardly any textbook on the market which covers the topic of transportation engineering from the perspective of Indian conditions. This book, we believe, not only fills this gap, but also presents the area of transportation engineering in a manner that will prepare students to tackle real-life problems.

The textbook is designed for the undergraduate as well as the first year master's students in civil engineering. It encompasses a wide range of topics from geometric design, to traffic engineering, to public transportation systems, to pavement design and construction, and many more. Figures and other explanation aids are used extensively to provide a proper grounding in the principles of transportation engineering. Each chapter contains an ample number of solved, illustrative examples, and exercise problems. Although the book primarily addresses the needs of students in civil engineering, it will be equally useful as a reference material for practising engineers and as a guide to those in urban planning, public administration, and management.

We hope that the instructors pursuing transportation engineering education at the various engineering colleges throughout the country will find this book stimulating. We also hope to get many constructive suggestions, criticisms, and corrections from our readers.

Any book is incomplete without a proper acknowledgment of the debt to many persons (other than the authors) who made it possible.

I, Partha Chakroborty, am deeply indebted to my mentor, Prof. Shinya Kikuchi for instilling in me the drive to work hard and for inculcating in me the discipline to think clearly. I wish to thank my student, Arijit Mandal, who helped me in solving some of the problems in the text. In all humility, I wish to express my gratitude towards my parents, Chhaya and Durgadas Chakroborty, who sacrificed a lot to give me a good education. I also thank my sister, Moushumi, for taking some of the photographs used here. Finally, I would like to thank my wife, Sharmistha, for the constant encouragement and support and my son, Promit, for bearing with a tired, and sometimes cranky, father for almost a year.

I, Animesh Das, wish to express my gratitude to my students, Pijush Ghosh and Prasenjit Basu, for many useful discussions during the preparation of this book. I am thankful to NHAI (Lucknow and Kanpur offices), ICT (Kanpur office), and Pinaki Roy Chowdhury of Lea Associates, New Delhi for providing me with some of the photographs used here. I am also thankful to J.C. Verma for diligently preparing many of the figures used in the text. Finally, I am grateful to my wife, Nibedita, for her active support and constant inspiration in this endeavour.

We wish to thank Kaushik Pahari for the secretarial help provided by him. Last but not the least, we also acknowledge admirable support provided by the Indian Institute of Technology Kanpur and, our publishers, Prentice-Hall of India, New Delhi.

PARTHA CHAKROBORTY
ANIMESH DAS

Introduction

1.1 TRANSPORTATION ENGINEERING

Mobility has always been important to human society. The Indians in Indus Valley Civilization built roads, the Egyptians built them, and the Romans built them as well. All civilizations of the past built roads because humans valued efficient mobility and because communication and trade were treated essential to the functioning of societies.

In our modern society, the need for efficient and safe transportation has increased so much so that the transportation facilities of a state are considered a mark of its progress, leading to a direct correlation between the two. This aspect is reflected in Table 1.1 which gives, for a sample of countries, the total length of paved roads, and a derived measure called the *road density* in metres of paved road per 1000 persons. The table also gives an independent and well accepted measure of development—*infant mortality rate*; the lesser the rate of infant mortality, the more developed the country.[1] As mentioned, the table shows a positive correlation between development (measured through infant mortality) and road density. In order to highlight this correlation further, Figure 1.1 plots road density versus the inverse of infant mortality rate (or the number of live births per infant death). The positive correlation established between development and paved road density is not incidental. Transportation does play a vital role in a country's development by facilitating trade between regions, reducing travel time costs, improving accessibility, etc. This text is concerned with the field of transportation engineering which shows how to build safe and efficient transportation systems.

Transportation engineering is the application of scientific processes (like observation, analysis, and deduction) to the planning, design, operation, and management of transportation facilities. There are various kinds of transportation facilities, some are used by human-powered vehicles (like cycles) while others are used by jet-powered

[1]Data on paved road length, population and infant mortality are taken from *2001 World Development Indicators* published by the World Bank [264].

Table 1.1 Road density and development

Country	*Total paved roads* (1000 km)	*Paved road density* (m/1000 *persons*)	*Infant mortality* (*deaths per* 1000 *live births*)
Argentina	63.35	1.71	18
Australia	353.33	18.41	5
Bulgaria	34.30	4.19	14
Egypt	49.98	0.78	47
Ethiopia	3.81	0.06	104
France	893.50	15.18	5
Greece	107.41	10.17	6
Hungary	81.68	8.15	8
India	1393.22	1.37	71
Italy	654.68	11.35	5
Mexico	109.40	1.12	29
Pakistan	109.40	0.79	90
Puerto Rico	14.40	3.67	10
Sierra Leone	0.90	0.18	168
South Africa	63.03	1.47	62
Thailand	62.99	1.04	28
U.S.A.	3732.76	13.26	7
Zimbabwe	8.69	0.72	70

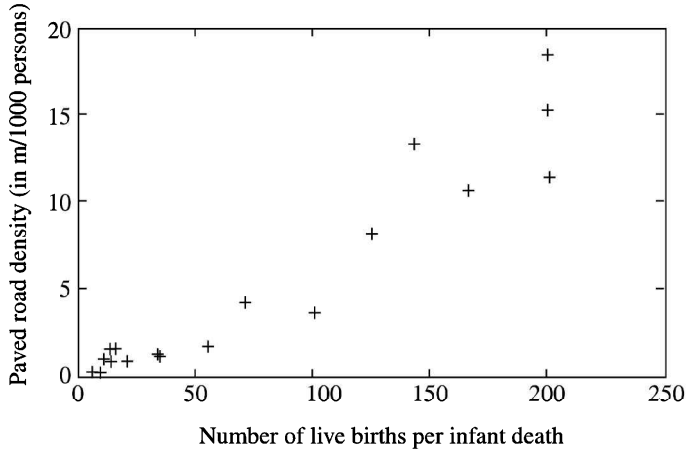


Figure 1.1 Road density and development.

vehicles (like aircraft); some are used for handling stationary vehicles (like parking lots or ports) while others are used for travelling at speeds in excess of 150 kmph (like high-speed trains); some are used by vehicles which move on land (like automobiles) while others are used by vehicles which move in air (like aeroplanes or helicopters); some are

used by drivers with little or no training (like the drivers of animal-drawn carts) while others are used by trained drivers (like pilots) with years of experience behind them. Transportation engineering is also multidisciplinary and requires knowledge from specialized fields such as psychology, economics, ecology and environment, sociology, management, optimization, graph theory, probability theory, statistics, computer simulation, and other areas of civil engineering (such as structural and geotechnical engineering).

This large breadth of transportation engineering presents a considerable challenge to developing an introductory text on the principles of transportation engineering. In order to meet this challenge, first, an attempt is made to rationally classify the field of transportation engineering. In Section 1.2, three schemes of classification of transportation studies are proposed and these classes have been developed. One of these schemes is used as the skeleton over which this book is built. Further, in order to keep the treatment focused, the emphasis in this book is on roadways (highways) based transportation systems.

1.2 CLASSIFICATION OF TRANSPORTATION STUDIES

In order to systematically study a vast field like transportation engineering, we need to classify it meaningfully and then study the various classes. This section makes an attempt to rationally classify this area of engineering.

Any transportation system consists of various modes of travel ranging from walking to driving to use of crafts that fly. One way of classifying a transportation system would then be to form classes of different modes of travel. Such classification is termed *modal classification* and forms the subject matter of Section 1.2.1.

A little thought at this stage will show that all modes of travel consist of the same set of elements such as a person who drives, the vehicle which is driven, the path that is used, the user who uses the mode, and the like. Therefore, another way of looking at a transportation system could be to look at its elements. Such a classification is termed *elemental classification* and is described in detail in Section 1.2.2.

Another very different way of looking at a transportation system, is from the point of view of functions that a transportation engineer needs to carry out. For example, a transportation engineer may need to analyze and design facilities to be used by vehicles, or may need to determine routes of buses, or may even need to chalk out a plan of transportation-related activities. We can therefore classify the field of transportation engineering functionally. This *functional classification* of the area is described in Section 1.2.3.

Figure 1.2 shows a diagrammatic representation of the classification schemes discussed here. As can be seen from the figure, transportation engineering can be viewed as the engineering of transportation systems—a conglomeration of different modes of transport which in turn consists of various basic elements. Also, it can be seen that transportation engineering as a field requires various functions like planning, pavement engineering, and so forth.

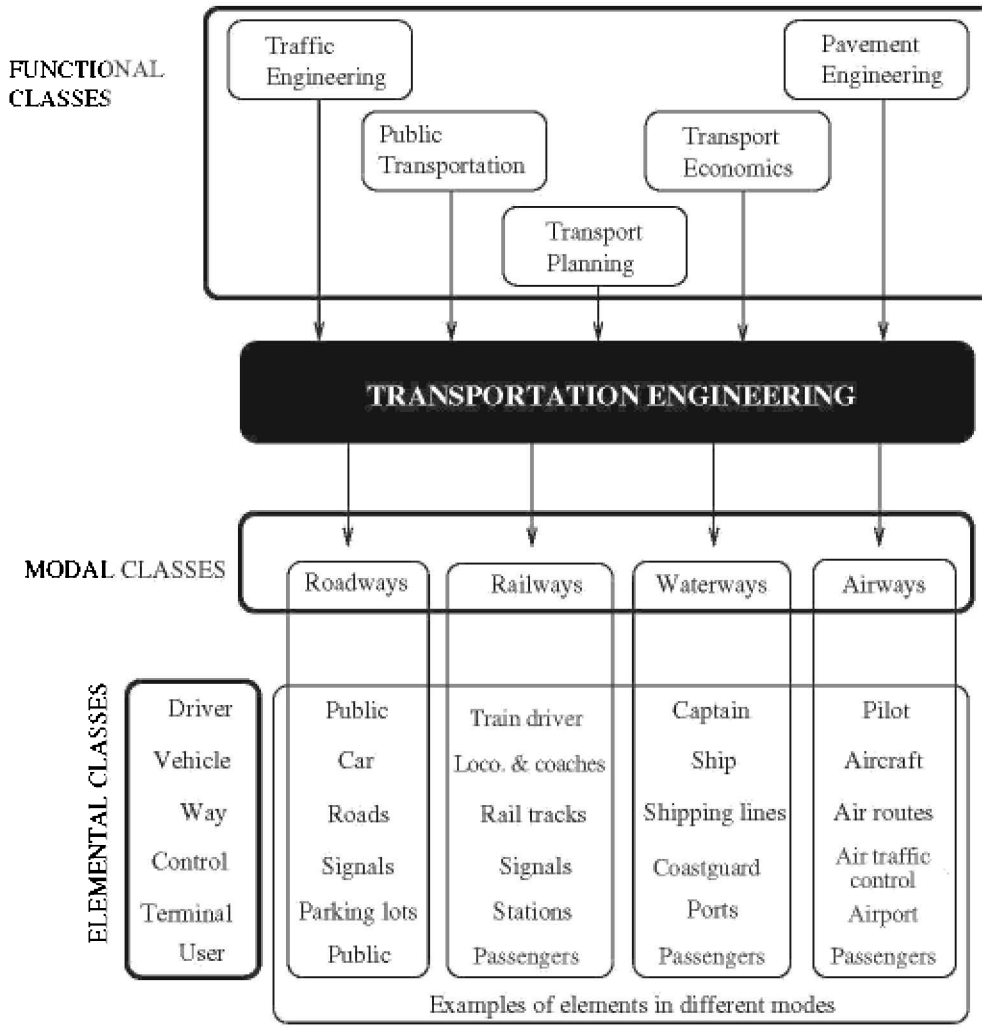


Figure 1.2 Classification schemes used for transportation engineering.

1.2.1 Modal Classification

A mode of transport may be simply defined as a means of transportation. A mode, according to this definition, includes walking, automobile, bus, train, ferry-boat, ship, aircraft, helicopter, and so on. However, all these modes use either the road, the rail tracks, the waterways (like rivers, travel paths in oceans), or the airways. Hence, a slightly different way of classifying the transportation systems according to modes would be to define classes as groups of modes which use a particular 'surface'. In this definition then, there are, in general, four classes—the roadways, the railways, the

waterways, and the airways (see Figure 1.2). These different modes are briefly described in the following text.

Roadways

In this mode of transport, all vehicles use the roadways to travel from one point to another. There are various kinds of vehicles in this class. Some are motorized (like automobiles, trucks, buses, etc.) and others are human- or animal-powered (like cycles, bullock carts, etc.). Some are private vehicles while others are meant to move either a large number of people (*public transport*) or goods over long and short distances. Figure 1.3 shows a typical road section being used by different types of vehicles. Figure 1.4 shows a tram (street-car) in Kolkata which runs on the roadway on special rails embedded onto the road surface.

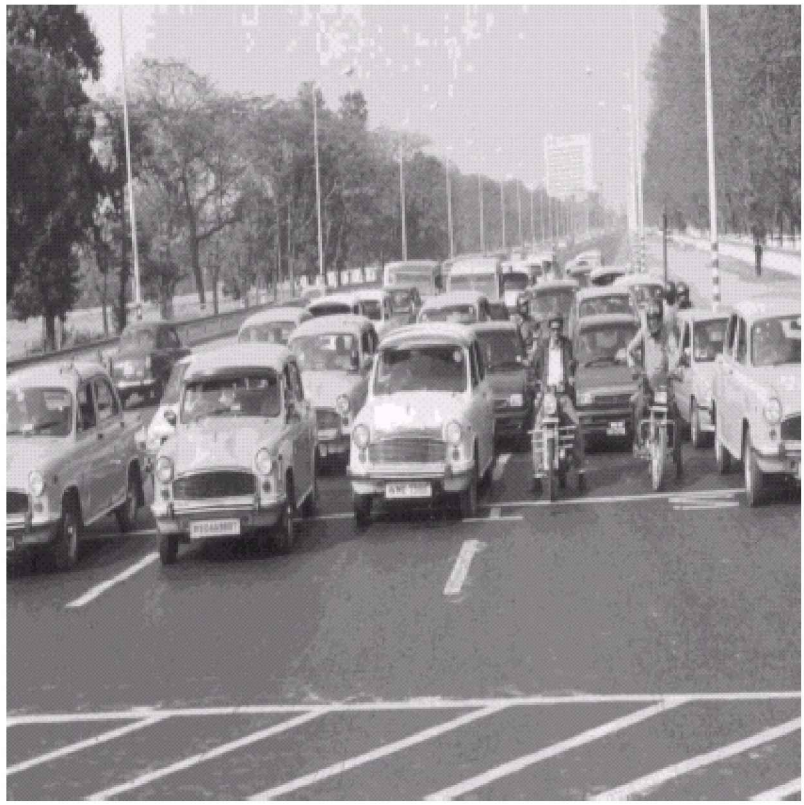


Figure 1.3 Roadway transport in Kolkata. (*Courtesy: M. Chakroborty*)

Roads are also of various kinds; some, like the intercity roads or by-passes, offer high speed of travel but have limited accessibility—that is, the road can be accessed only

Figure 1.4 A tram plying on the roads of Kolkata. (*Courtesy: M. Chakroborty*)

at a limited number of points. At the other end of the spectrum, there are local roads which provide very good accessibility (for example, this kind of road can be accessed by every home on the street) but offer low speed of travel. Figure 1.5 shows the different types of roads on a Cartesian space of speed-of-mobility versus accessibility. It may be pointed out that the different classes of roads have different design standards (for example, the Indian Codes of Practice IRC:86–1983 [80] and IRC:73–1980 [79]). The design features of roads are discussed in Chapters 3 and 12.

From a transportation engineer's standpoint, the important aspects related to the roadways mode of transport include:

- Safe and efficient operation and control of road traffic
- Layout of roads
- Structural design of the roadway (pavement design)
- Roadway-based public transportation

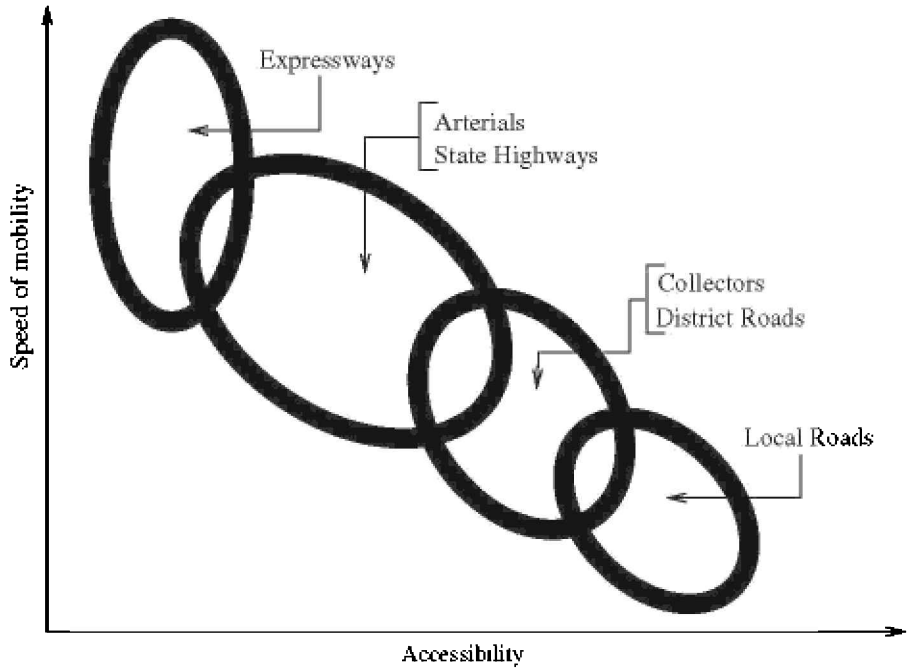


Figure 1.5 Different types of roads and their accessibility and speed characteristics.

Railways

In this mode of transport, all vehicles use rail tracks to move from one point to another. Vehicles in this class consist of a long train of coaches attached to a locomotive. The coaches and the locomotive have steel wheels which run on steel rail tracks. The locomotive either uses diesel or uses electricity as fuel (locomotives which run on coal are hardly used these days). These trains are used for transporting people as well as goods. There are long-distance trains as well as trains which run within an urban area (like the underground trains of Kolkata, London, Paris, etc.) or within a greater metropolitan area (like the local train services of Mumbai or Kolkata). Since the rail tracks provide a dedicated right-of-way for the train services, these are good for high speed mass transit facilities. Figures 1.6 and 1.7, respectively, show photographs of the underground train service and the greater metropolitan area train service which operate in Kolkata. Figure 1.8 shows a high-speed (operating at the average speed of approximately 220 kmph) long-distance train service operating in Japan.

Trains stop at pre-specified locations called stations. At stations, various activities take place like (i) boarding and alighting of passengers, (ii) loading and unloading of goods, (iii) regrouping of coaches or compartments, (iv) maintenance activities, (v) ticketing, etc. The movement of trains on rail tracks is controlled through signal systems, operated either manually or automatically.

Figure 1.6 Underground, within-the-urban-area, train service in Kolkata.

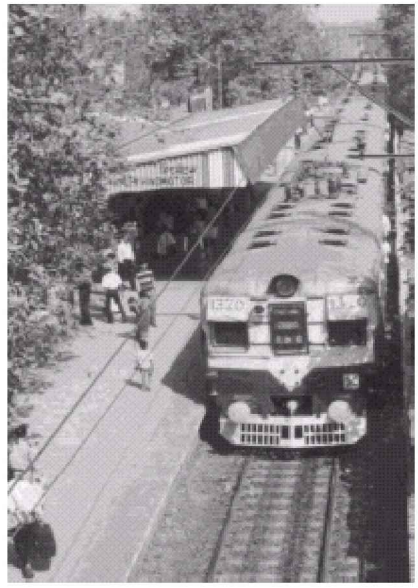


Figure 1.7 Greater metropolitan area train service in Kolkata. (*Courtesy: M. Chakroborty*)

From a transportation engineer's standpoint, the important aspects related to the railway mode of transport include:

- Safe and efficient operation and control of rail traffic
- Layout of rail tracks
- Structural design of the subgrade on which rail tracks run
- Planning of stations or terminals for railway vehicles

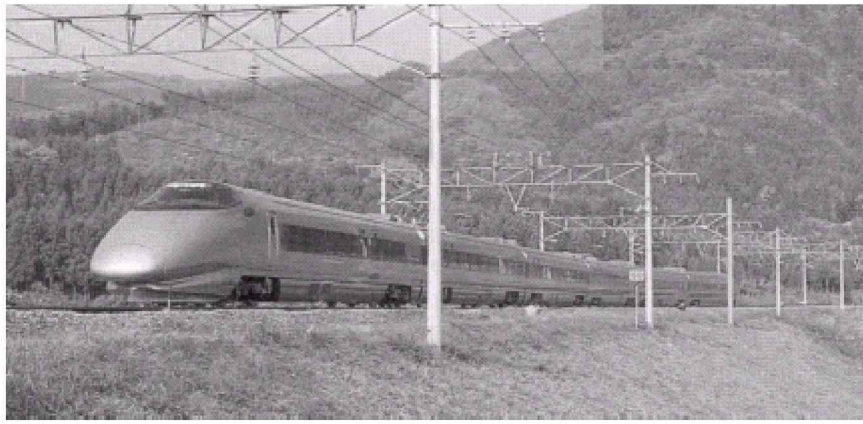


Figure 1.8 High-speed long-distance train service in Japan.

Although many of the principles discussed in this book (like geometric design discussed in Chapter 3 or routing and scheduling discussed in Chapter 6) are applicable to railways, there are certain issues which are not covered here. The interested reader may refer to Vuchic [256] or Mundrey [170] for various aspects of railway engineering.

Waterways

In this form of transport, all vehicles use channels in water bodies (like rivers, lakes, and oceans) to travel from one point to another. Simply stated, a channel is a charted corridor in an expanse of water which is safe and a reasonably direct connection between the origin and the destination. Various kinds of vessels operate on the waterways starting from the small ferry boats to the large ocean liners. Not all channels can be used by all types of ships, for example, the ocean liners generally cannot use the inland river channels.

Historically, waterways were the only connections between far-off places for transportation of both people and goods. However, due to the slow speed of vessels and the advent of airways, waterways are not used today for long distance voyages. Presently, waterways are used either to ferry people and goods over short distances (like across a river or lake or strait, for example, see Figure 1.9) or to transport goods over long distances (like between different countries).

Ships and ferry boats generally dock (or stop) at ports. At ports, various activities such as boarding and alighting of passengers, loading and unloading of cargo, refuelling, maintenance activities, ticketing, and customs, immigration, etc. are carried out.

From a transportation engineer's standpoint, the important aspects related to this mode of transport include:

- Safe and efficient operation and control of ship traffic especially at and near ports
- Planning and operation of ports

Figure 1.9 A ferry boat operating on the Hoogly river to ferry people between Kolkata and her sister city Howrah. (*Courtesy: M. Chakroborty*)

Airways

In this mode of transport, vehicles use air routes to travel from one point to another. Simply stated, air routes are charted paths in the air based on various characteristics like directness of connections, prevailing atmospheric and wind conditions, international agreements, and safety issues. Various kinds of aircraft use the airways—from small, single-engine planes to large jet aeroplanes.

Airways provide the fastest and one of the safest modes of transport. Their use ranges from small shuttle flights over distances of around two hundred kilometres to long transatlantic flights of well over three thousand kilometres. Airways also help in connecting remote places in difficult terrain. Presently, airways are extensively used to travel over long distances. Goods, perishable commodities in particular, are also transported in cargo planes over long distances.

Aircraft touch down and take off using long (sometimes around 4–5 km) straight pavements called *runways* constructed at airports. The runways are generally some distance away from the airport terminal buildings (where the land-side interfaces with the air-side of the transportation system) and are connected to them through taxiways and large open-paved surfaces called *aprons*. The activities at an airport include (i) boarding and alighting of passengers, (ii) loading and unloading of cargo, (iii) refuelling, (iv) maintenance, (v) air traffic control, (vi) ticketing, and (vii) customs, immigration, and so forth.

From a transportation engineer's standpoint, the important aspects related to this mode of transport include:

- Safe and efficient operation and control of air traffic especially at and near airports
- Planning and operation of airports
- Orientation of runways and layout of taxiways
- Planning the parking pattern of aircraft
- Structural design of the runways, taxiways, and aprons
- Planning and operation of vehicle circulation and parking on the land-side of airports.

In this text, much of the operational and planning aspects of airports have not been covered. The interested reader may refer to Horonjeff and Mckelvey [106], Wells [260] or Ashford et al. [6] for a detailed discussion on *airport engineering*.

1.2.2 Elemental Classification

As described earlier, a transportation system can be classified in terms of the basic elements which constitute any mode of transport (see Figure 1.2). These basic elements are identified and briefly described in the following.

Driver. Every mode of transportation has a driver who controls the vehicle used in that mode of transportation. The importance of this element in the analysis and design of the transportation facilities, however, varies from mode to mode. This element assumes maximum importance in the roadways where a tremendously large number of drivers interact with each other and the facilities. The characteristics of drivers are discussed in Chapter 2.

Vehicle. The vehicles which are used in transportation have certain characteristics (for example, turning radius, braking distance, accelerating capabilities, etc.) which influence the design and operation of the transportation facility. The characteristics of roadway vehicles are described in Chapter 2.

Way. Every mode of transportation uses a specified path which is either constructed or charted. For example, in the case of roadways and railways the way (road or rail track) has to be laid out and constructed while in the case of waterways and airways, the ways used are only charted paths on water bodies or in atmosphere. The design and construction of roadways is described in detail in the section on pavement engineering (see Chapters 12 and 13).

Control. In order to ensure safety and efficiency of operation, there are system level controls which are imposed on the movement of the vehicles. These controls could be static (in the form of rules or road signs like "No U Turn" or "One-Way") or dynamic

(in the form of road or rail signals, or instructions from air traffic controllers, etc.). Indepth discussion on the analysis and design of control mechanisms for roadway traffic is provided in Chapters 4 and 5.

Terminal. This is a location where the vehicles of a mode stop for various reasons including (i) boarding (loading) and alighting (unloading) of passengers (goods), (ii) resting when not in use, (iii) refuelling, (iv) maintenance, etc. The terminal facilities for roadways (parking lots) are discussed in Chapter 5.

User. Obviously any transportation system runs to provide service to its users. The users are (i) the public at large for transportation modes which cater to passenger transport or (ii) organizations for transportation modes which cater to goods transport. The transportation system must be sensitive to the needs of its users. Some of the characteristics of human beings as users of the roadway transportation system are discussed in Chapters 2 and 9. The effect of concern for users on operation of certain transportation systems also becomes evident in Chapter 6.

1.2.3 Functional Classification

This scheme of classification of transportation engineering divides this discipline in terms of the different functions required of an engineer working in this field. Functionally, transportation engineering can be divided into the following primary classes: (i) traffic engineering, (ii) pavement engineering, (iii) public transportation, (iv) transport planning, and (v) transport economics. There are certain other functions which a transportation engineer may need to perform, like providing specialized transport for the elderly or the handicapped, logistics planning, etc. In the following, the primary divisions enumerated above are briefly described.

Traffic Engineering

This area of transportation engineering deals with the analysis, design, and operation of transportation facilities used by vehicles of various transportation modes. Such a study assumes utmost importance in the case of roadways as the number of vehicles using the transportation facilities are the highest as well as the most varied both in terms of their type, their origins and destinations, their purposes, etc. Generally therefore, the scope of traffic engineering, in the most commonly used meaning of the term, is limited to roadways traffic. For example, the USA based Institute of Transportation Engineers, ITE, defines traffic engineering as (see [130]) "that phase of transportation engineering which deals with planning, geometric design and traffic operations of roads, streets, and highways, their networks, terminals, abutting lands, and relationship with other modes of transportation."[2]

[2]As quoted in McShane and Roess [157].

Pavement Engineering

This area of transportation engineering deals with the structural analysis and design of the way used by different modes of transportation. Specifically, pavement engineering is concerned with (i) the analysis, structural design, construction, and maintenance of roadway pavements, runways, taxiways, and rail tracks and their drainage and other associated structures, and (ii) the materials used in the construction of all such structures.

Public Transportation

The area of public transportation is concerned with the analysis, design, and operation of public transportation systems. A public transportation system is a transportation system which operates to move the general public from one point to another. It includes, at one end of the spectrum, para-transit systems like share-taxis (which operate on fixed routes but not according to any fixed schedule) to rapid-transit systems like greater metropolitan area train services, at the other end of the spectrum. The design of a public transportation system includes the design of routes (including stop locations), design of schedules, determination of fare structures, and crew scheduling.

Transport Planning

Transport planning deals with planning transportation facilities which will be able to meet the present and future needs in a sustainable manner. This field focuses on issues like estimation of future demands, needs and problems; generation of alternative transportation solutions; studying the financial, economic, and technological implications of these alternatives; and analyzing their impact on the environment, land-use and demograph trends of an area. Transport planners are also entrusted with the task of choosing the right alternative and preparing a plan for its implementation.

Transport Economics

This area studies the various economic costs and benefits of building and operating different transportation facilities. The area focuses on (i) identifying the economic costs and benefits and their incident sectors, (ii) studying the numerous techniques available and formulating new techniques to estimate these costs and benefits, (iii) analyzing the financing and cost recovery aspects of transportation projects, and (iv) suggesting economic ways of solving certain transportation problems.

1.3 ORGANIZATION OF THE BOOK

The book divides the presentation of the principles of transportation engineering according to the functional classification scheme described in this chapter. The book thus has five parts covering (i) Traffic Engineering, (ii) Public Transportation, (iii) Transport Planning, (iv) Pavement Engineering, and (v) Transport Economics.

In Part I, there are four chapters. The first chapter (Chapter 2) describes the basic elements of traffic engineering. Chapter 3 presents the principles of geometric design with an emphasis on roadways. Chapter 4 dwells on the theory of traffic flow for highways while the last chapter in this part (Chapter 5) discusses the design process of roadway traffic facilities.

The Part II has Chapters 6 and 7. Chapter 6 is devoted to operations of public transportation systems. The capacity of such systems is analyzed in Chapter 7.

The Part III of the book is on transport planning. It consists of Chapters 8 and 9. Chapter 8 looks into the transportation planning process, reasonably in detail, while Chapter 9 explores in detail, one of the important aspects of the planning process, namely demand forecasting.

The Part IV of the book deals with the pavement engineering aspects of transportation engineering. This part covers materials used in pavement construction (Chapter 10), structural analysis of pavements (Chapter 11), pavement design (Chapter 12), construction procedures for different types of pavements (Chapter 13), and pavement maintenance strategies (Chapter 14).

Finally, Part V of the book briefly presents the different aspects of highway economics and finance.

PART I

TRAFFIC ENGINEERING

Properties of Traffic Engineering Elements

2.1 INTRODUCTION

Traffic engineering is mainly concerned with the flow of vehicular traffic on roadways. This chapter discusses the basic properties (or characteristics) of the transportation elements (described in Chapter 1) as they relate to traffic engineering. Specifically, the chapter describes the relevant features of (i) vehicles, (ii) drivers and users, (iii) roads, (iv) control mechanisms, and (v) terminal or parking facilities.

2.2 VEHICLE CHARACTERISTICS

The importance of characteristics of vehicles to traffic engineering is self-evident. Among the different features which characterize a vehicle, the ones which are of importance to a transportation engineer are: (i) size, (ii) weight and axle configuration, (iii) power-to-weight ratio, (iv) turning radius, (v) turning path, and (vi) pollution creation. In the following three subsections, these factors are discussed in detail.

2.2.1 Size, Weight, Axle Configurations and Power-to-Weight Ratio

A vehicle has three dimensions, the length, the width, and the height. All the three dimensions are required in the design of different transportation facilities. For example, when designing open air on-street or off-street parking facilities the length and width of vehicles are important input parameters. The height of vehicles (especially those of trucks and buses) are important considerations when placing signs and designing overpasses and underpasses.

The weight of vehicles, especially of heavy vehicles, plays an important role in the design of both flexible and rigid pavements. Hence, knowledge of vehicle weights is important for transportation engineers. Since the weight of a vehicle is transferred to the pavement layer through the axles, the wheel and axle configuration of vehicles also plays an important role in the design of pavements. In fact, as will be discussed in the

chapters on pavement analysis and design, it is the number of axles (and not vehicles) of a standard weight which is considered as a variable in the design of pavements.

The power-to-weight ratio of a vehicle is a parameter which characterizes the ease with which a vehicle can move. For example, human-powered vehicles like cycles or rickshaws have very low power-to-weight ratio and their operating characteristics (like acceleration capability, sustainable speeds on slopes, etc.) are thus very poor. Motorized vehicles, like automobiles, motor bikes, etc. have high power-to-weight ratios and hence have good operating characteristics. Heavy vehicles, on the other hand, though motorized, have poorer power-to-weight ratios than those of other motorized vehicles owing to the heavy weight of the vehicles (especially when full). This ratio is important to transportation engineers as it relates to the operating efficiency of vehicles on roads and especially on positive gradient road sections. For example, the length for which a positive gradient can be maintained on a road is often limited by its effects on the operation of heavy vehicles.

2.2.2 Turning Radius and Turning Path

Every vehicle has a minimum turning radius which is the radius of the circle that will be traced out by the front wheels if the vehicle moved with its steering turned to the maximum extent possible. This radius is dependent on the design and class of the vehicle. For example, a big vehicle like a bus has a much larger turning radius than that of a smaller vehicle like an automobile.

Another important feature related to vehicles is the turning path traced by the vehicle. Since only the front wheels turn and (i) the rear wheels are fixed, and (ii) the vehicle's body extends beyond the tyres, different points of the vehicle trace out different paths as shown in Figure 2.1. Due to this, as can be seen from the figure, the

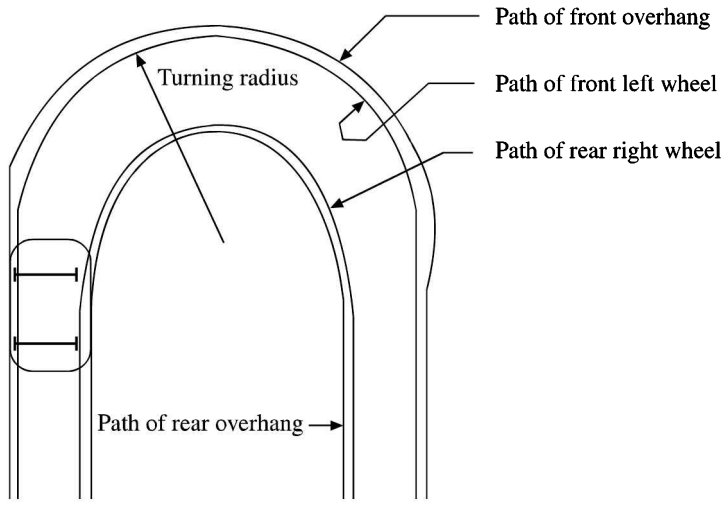


Figure 2.1 Typical turning path of vehicles.

effective width of a vehicle on a turn increases. This fact is taken into account while designing turns at intersections of roads (where the corners are arched in order to increase the space for turning) and at sharp horizontal curves (where the width of the road on the curve is increased in order to accommodate the increased effective width of vehicles).

2.2.3 Vehicle as a Source of Pollution

Vehicles (which operate on fossil fuel) emit pollutants into the atmosphere. Vehicles release hydrocarbons (due to unburnt fuel and fuel evaporation) which react with nitrogen oxides in the presence of sunlight to produce ground level ozone which irritates the eyes and aggravates respiratory problems. Vehicles also release various oxides of nitrogen, a major contributor to the formation of acid rain. Carbon monoxide—which reduces flow of oxygen in the bloodstream—also forms a part of the automobile emission. Carbon dioxide, a greenhouse gas, is also emitted by automobiles. The rate of release of these pollutants is affected by the operating speed, frequent acceleration/deceleration, type and age of vehicle, air temperature, etc. A traffic engineer can aid in the process of reducing pollutant emissions by designing roads (especially ones with frequent signalized intersections) where drivers can maintain a reasonable cruising speed and do not have to accelerate and decelerate frequently. For example, signals could be coordinated such that vehicles in different movements do not have to stop at all the signals. Driver education and compliance with traffic regulations (like stopping engines at red lights) also aid in reducing pollutant emissions.

Moving vehicles also transmit noise, caused by engine functions, action of tyres on roads, aerodynamics of moving vehicles, horns, the squeals of brakes, etc. to the surrounding area. Noise pollution, other than hurting the auditory faculty of people, also causes psychological and other physiological problems by interferring with sleep, concentration, and certain daily activities of human beings. Noise emission is also dependent on age and type of vehicle, type of engine, driving pattern, etc. Traffic engineers, often have to mitigate the effect of noise on habitations by building noise barriers along heavily travelled roads. Sometimes though, especially in urban areas, nothing can be done, except enforcing rules on the use and type of horns, to reduce the effect of noise.

2.2.4 Design Vehicle

On any given road, vehicles of different classes move. On Indian roads, the vehicle classes include: (i) motorized two-wheelers, (ii) motorized three-wheelers, (iii) passenger cars or automobiles, (iv) buses, (v) single-unit trucks, (vi) semi-trailers, (vii) truck-trailer combination, and (viii) slow non-motorized vehicles like cycles, rickshaws, and animal-drawn carts. Each class of vehicle differs from the other in most of the characteristics mentioned above.

For the purposes of design, an engineer needs to choose a type of vehicle based on the characteristics of which the road design is done. Such a vehicle is referred to as a *design vehicle*. The Indian Roads Congress, in IRC:3–1984 [54], provides some of the characteristics of some of the types of vehicles listed here. The AASHTO [3] also gives the characteristics of many of the vehicle classes listed here.

2.3 HUMAN FACTORS AND DRIVER CHARACTERISTICS

Two of the important constituents of a transportation system are drivers and transport users (or passengers). An understanding of some of the human characteristics of drivers and passengers is therefore essential for proper design of transportation facilities. In the following, some of these characteristics are described.

2.3.1 Perception–Reaction Process

Human beings react in different situations by first perceiving the scenario, then inferring a suitable course of action, and finally implementing that action. This entire process is referred to as the perception–reaction process. The time required to complete this process is referred to as the *perception–reaction time*.

Various studies have been carried out to determine the perception–reaction time of drivers under different situations which arise in transportation engineering. It is found that this reaction time not only increases with (i) age and (ii) intoxication level, but is also affected by features like (i) expectancy (where drivers have learnt to anticipate certain stimulus) and (ii) complexity of the scenario (where the information to be processed for a response is large). Obviously, for different situations the reaction time is different. Situations where the perception–reaction time plays a very important role are emergencies where drivers need to brake in order to be safe. Hence, a lot of work has gone into determining the perception–reaction time in these cases. Based on these results, both the AASHTO [3] and the IRC (for example, see IRC:73–1980 [79] suggest a value of 2.5 seconds as the perception–reaction time for braking.

2.3.2 Psychological Characteristics

Certain human factors, psychological in origin, play an important role in transportation engineering. That is, these factors are not due to any physiological characteristics or functions of the human body. In the following, two such factors are discussed.

Value of time

Human beings value time and the way that time is spent. This value for time has implications when designing public transport systems (specifically routes and schedules) and various other traffic facilities like signalized intersections. For example, less number

of people will use a route which has a longer travel time compared to a route which has a shorter travel time; yet on the other hand the longer route may be preferred if the journey on this route is a lot more comfortable (either because the road is less congested or because the buses running on the route are less crowded).

Safety considerations

Safety requirements play a major role in traffic engineering. For example, a longer distance between vehicles is required to be maintained while going at higher speeds. As will be discussed in Chapter 4, this parameter ultimately leads to an upper bound on the maximum number of vehicles that can flow on a road. Driving needs to be much more cautious on narrow roads (though wide enough to accommodate one vehicle comfortably) leading to smaller capacities for such roads.

The concept of maintaining a safe buffer distance also plays an important role when drivers choose gaps in the opposing adjacent lane in order to overtake a slow moving vehicle. It is seen that although, in order to complete the overtaking manoeuvre, a vehicle needs a certain obstacle-free distance, say D, in the opposing lane, in reality a vehicle initiates an overtaking manoeuvre only when the distance gap between the vehicle and an oncoming vehicle is much greater than D. The reason for this is that drivers are extra cautious when travelling in the opposing lane for overtaking purposes, and want to have a buffer distance with the oncoming vehicle even after completing the overtaking manoeuvre. Various studies have been done to understand the overtaking behaviour. Based on these studies, the distance which the drivers look for in order to initiate and complete an overtaking manoeuvre is generally described as shown in Figure 2.2. In the figure, d_1 may be thought of as the 'perception–reaction' distance—this is the distance between the point at which the driver perceives a possibility for overtaking and actually starts initiating the overtaking manoeuvre. Distance d_2 is the distance physically required to complete the overtaking manoeuvre, distance d_4 is the distance the opposing vehicle travels during the overtaking manoeuvre, and distance d_3 is the buffer distance arising out of safety considerations of drivers. It must be noted that all these distances increase with the speed of the traffic streams. As shown in the figure, the sum $d_1 + d_2 + d_3 + d_4$ is referred to as the *overtaking distance*. The overtaking distance requirements at various speeds are generally provided in the codes and may have been derived based on a simplified representation of the above description. AASHTO [3] and IRC:73–1980 [79] are two such codes which provide the overtaking distance requirements at various speeds.

At unsignalized intersections, drivers sometimes have to choose gaps in the opposing stream of traffic in order to complete their turning manoeuvre. Here again, the drivers choose gaps which are much larger than those actually required for the manoeuvre. The minimum gap size which a driver chooses, referred to as *critical gap*, generally increases with the speed of the opposing stream, the number of opposing lanes, and the age of the driver, and reduces with the amount of time a driver spends in

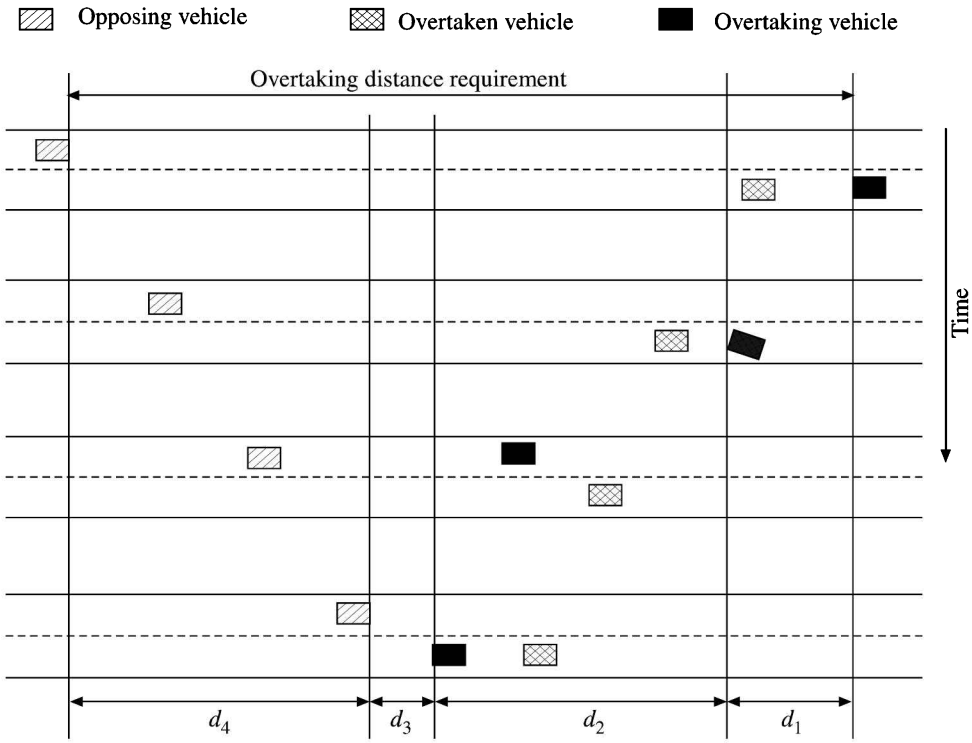


Figure 2.2 Components of overtaking distance.

waiting for a gap. In general, the critical gap varies between 4 and 7.5 seconds depending on the situation. The Highway Capacity Manual [103] provides a table to determine design values of critical gaps in various situations.

The gap acceptance behaviour of drivers has direct implications on the performance of unsignalized intersections (as discussed in Chapter 4) and a somewhat indirect implication on the performance of signalized intersections.

2.3.3 Comfort

Drivers as individuals, care about the comfort of the drive. As engineers, the design of various traffic facilities should be such that drivers do not face any discomfort. Among the common causes of discomfort to drivers, some are (i) excessive deceleration rates, (ii) excessive jerk, and (iii) glare.

If drivers are required to decelerate beyond a certain rate, they feel uncomfortable. The maximum value of the comfortable deceleration rate is around 3 m/s^2, and it obviously varies from person to person. The designer must therefore be aware of the comfortable deceleration rates preferred by drivers. A good example of how this value affects design can be seen in the dilemma zone analysis and amber time fixation discussed in Chapter 5.

High jerk, generally greater than 0.7 m/s^3, causes considerable discomfort to drivers. Drivers always experience jerk while negotiating curves on roads. Sometimes the curvature is large enough to cause uncomfortable jerk due to changes in centrifugal acceleration. The level of discomfort caused by jerk has a direct bearing on the geometric design of curves as will be seen in Chapter 3.

Glare is defined as an intense, blinding light. It causes extreme discomfort to the driver as the driver is effectively blinded for a very small period of time, posing a big safety hazard. The problem, in the context of design of traffic engineering, can occur from bright and ill-placed street lights or signboard lights. But mostly it occurs from headlights of oncoming vehicles, which use 'high-beam'. The problem of glare has direct bearings on the design of street lighting and street sign lighting. Sometimes, median barriers have to be constructed on high-speed two-way roads to eliminate the problem caused by glare.

2.3.4 Vision

The aspects of human vision which are important for a traffic engineer are (i) visual acuity, (ii) field of vision, and (iii) colour perception. In the following, each of these is explained in detail.

Visual acuity

Visual acuity refers to how well a person can see. Normal vision is defined as the ability of a person to recognize a letter (or an object) of approximately 8.5 mm size from a distance of nearly 6 m. A person with normal vision is said to have 6/6 vision. As per this notation, a person with 6/9 vision has poorer than normal vision because he/she can read (or recognize) from a distance of 6 m what a normal person can read (or recognize) from a distance of 9 m. Alternatively, a person with 6/9 vision can read (or recognize) from a given distance letters (or objects) which are 9/6 (=1.5) times bigger than those which a normal vision person can read (or recognize) from the same distance. In general, therefore, a 6/x vision person will have to be $6 \div x$ times closer than a normal vision person to be able to recognize the same letter (or object), or the letter (or object) has to be $x \div 6$ times larger for a 6/x vision person to be able to recognize it from the same distance as a 6/6 vision person. Further, visual acuity is affected by contrast and brightness of the object and the relative speed between the object and the driver.

Knowledge of visual acuity of drivers is necessary while designing road signs. The chapter on design of traffic facilities (see Chapter 5) shows how knowledge of visual acuity is used in the design of traffic signs.

Field of vision

Visual acuity reduces with the angle of vision. Persons can see most clearly within a 3-degree cone (see Figure 2.3). Clarity of vision is reasonable within a 10-degree cone.

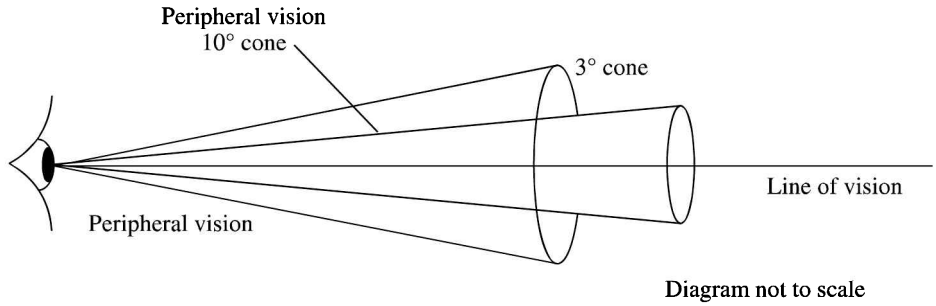


Figure 2.3 Field of vision for humans.

However, beyond that till about a 160 degree cone everything is peripheral vision. Information on field of vision is important while deciding the placement of road signs. For example, signs should be placed within the 10-degree cone of vision of drivers. Sometimes on wide two-way roads this is not possible and in such cases, signs are placed in overhead positions.

Colour perception

It must be understood that all types of colours and colour combinations are not equally discernible and only the most visible of these should be used in traffic facilities. Light colours on dark backgrounds or dark colours on light backgrounds are most easily discernible by the human eye. Based on discernibility considerations (and considerations of classification of signs) codes in various countries specify the colour combination for various types of signs. In India, for example, IRC:67–1977 [40] provides, among other things, the colour combination to be used on various traffic signs.

2.3.5 Design Driver

Drivers are different from one another in all of the characteristics mentioned above. Some drivers may have a low perception–reaction time, but very good visual acuity, and so on. Hence, for design purposes, the designer must choose those characteristics which make the design safe for most drivers. A driver, albeit fictitious, who has the characteristics chosen by the designer is referred to as the *design driver*. In general, a design driver is assumed to have a perception–reaction time of 2.5 seconds, comfortable deceleration rate of 3 m/s^2, allowable jerk of about 0.7 m/s^3, a 6/7.5 visual acuity (see Bell et al. [11]), and a critical gap value of between 4 and 7.5 seconds (depending on the complexity of the manoeuvre and the number and speed of opposing streams).

2.4 ROAD CHARACTERISTICS

Various road characteristics affect the flow of traffic. The most important among them are (i) width, (ii) presence or absence of shoulders, (iii) surface conditions, (iv) slopes, and (v) curves. Among these, the first two are discussed in latter chapters; in this chapter the other three features and their effects are described.

2.4.1 Surface Conditions

The surface conditions of a road may be described through two parameters: (i) the frequency and/or extent of the distressed sections (like, potholes, depressions, stripped sections) and (ii) the friction offered by the road surface. If a road surface is severely distressed, it will cause considerable hindrance to smooth flow of traffic due to frequent slowing down of vehicles and changing their paths to avoid potholes and the like.

Road surfaces, however, should provide sufficient friction to enable vehicles to move and stop effectively. A very smooth road surface (like when there is snow on the road or when it is wet) causes hindrance to flow of traffic and gives rise to possible safety hazards. The coefficient of rolling friction offered by dry paved surfaces should be around 0.5; this reduces to about 0.3 if the surface is wet. In the presence of snow, this coefficient is even lower.

Another friction coefficient of road surfaces, known as the coefficient of side friction, is also important to traffic engineers. This friction coefficient, which is a measure of the resistance offered by the road surface to movements orthogonal to the direction of motion, comes into play while designing horizontal curves and superelevations. This aspect of road surface friction and its importance in design will be discussed in detail in Chapter 3.

2.4.2 Slopes

Moving on uphill roads, or positive slopes, requires an additional effort from a vehicle. As indicated earlier, vehicles with low power-to-weight ratio have problems moving on sustained positive grades or slopes; the steeper the grade the shorter is the extent over which these vehicles can move at any reasonable speed. While designing roads, therefore, attention must be paid to the slope and length of the uphill section.

Another matter related to slopes and friction of roads that may be mentioned here is the braking distance requirement of vehicles. Figure 2.4 shows a vehicle moving at a speed of v_i m/s on a road inclined at an angle θ with the horizontal. The rolling friction coefficient of the road surface is f_r. Further the minimum distance x, along the incline which the vehicle will have to travel in order to reduce its speed from v_i to v_f is given

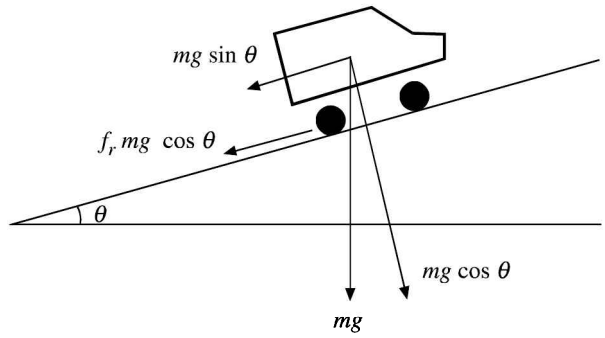


Figure 2.4 Force diagram for determining the braking distance.

by (where g is the acceleration due to gravity)

$$x = \frac{v_i^2 - v_f^2}{2g(\sin\theta + f_r \cos\theta)}$$

$$= \frac{v_i^2 - v_f^2}{2g\cos\theta(\tan\theta + f_r)}$$

It is customary, however, to represent the braking distance along the horizontal. If d_b is the braking distance and G the slope of the incline in fractions, then

$$d_b = x\cos\theta = \frac{v_i^2 - v_f^2}{2g(G + f_r)} \tag{2.1}$$

It may be noted that the above relation also works if the incline is downhill; in this case, however, G is negative. As will be seen later, this relation is utilized to determine the minimum distance required for a vehicle to come to a stop (i.e. $v_f = 0$). Another related concept is that of *stopping distance*. Stopping distance is the sum of the braking distance required to come to a stop and the distance travelled during the perception–reaction time.

EXAMPLE 2.1

Determine the braking distance for the following situations: (i) a vehicle moving on a positive 3 per cent grade at an initial speed of 50 km/h, final speed 20 km/h; (ii) a vehicle moving on a 3 per cent downhill grade at an initial speed of 50 km/h, final speed 20 km/h; and (iii) a vehicle moving on a level road at an initial speed of 50 km/h, final speed 0 km/h. Also determine the stopping distance for (iii). Assume the coefficient of rolling friction to be 0.5 and the perception–reaction time to be 2.5 s.

Solution

(i) Initial speed, v_i = 50 km/h = 13.89 m/s
Final speed, v_f = 20 km/h = 5.56 m/s
g = 9.81 m/s^2 and G = 3/100 = 0.03

$$d_b = \frac{13.89^2 - 5.56^2}{2\times 9.81(0.03 + 0.5)} = 15.58 \text{ m}$$

(ii) Initial speed, v_i = 50 km/h = 13.89 m/s
Final speed, v_f = 20 km/h = 5.56 m/s
g = 9.81 m/s^2 and G = − 3/100 = − 0.03

$$d_b = \frac{13.89^2 - 5.56^2}{2\times 9.81(-0.03 + 0.5)} = 17.57 \text{ m}$$

(iii) Initial speed, v_i = 50 km/h = 13.89 m/s
Final speed, v_f = 0 km/h = 0 m/s
g = 9.81 m/s^2 and G = 0/100 = 0.0

$$d_b = \frac{13.89^2}{2\times 9.81(0.0 + 0.5)} = 19.67 \text{ m}$$

Stopping distance for (iii)

Initial speed, v_i = 50 km/h = 13.89 m/s
Final speed, v_f = 0 km/h = 0 m/s
g = 9.81 m/s^2, G = 0/100 = 0.0, perception–reaction time = 2.5 s

$$\text{Stopping distance} = 13.89 \times 2.5 + \frac{13.89^2}{2\times 9.81(0.0 + 0.5)} = 54.39 \text{ m}$$

2.4.3 Curves

Roads often have curves either in the horizontal plane (for example, when the road turns) or in the vertical plane (for example, when the road has varying gradients). Curves always pose a restriction on the distance over which the driver can see the road. This factor is taken into account in the design of curves. Detailed description of highway curves and their design is presented later in Chapter 3.

Curves, especially those in the horizontal plane, are often such that they create considerable centrifugal force on the vehicles moving along them. This factor also needs to be kept in mind while designing curves. Again, a detailed description of this aspect is given in Chapter 3.

2.5 CONTROL MECHANISMS

Any traffic system has controls which are either static or dynamic. Static controls broadly comprise: (i) rules of driving, (ii) road signs such as STOP, NO U TURN, and the like, (iii) raised islands (which in effect delineate preferred paths of vehicles), and (iv) road markings like dashed or solid lines dividing lanes, and so on. Dynamic controls, on the other hand, are the signals at signalized intersections which determine the right-of-way for the different movements at an intersection at any given time.

Control mechanisms have, as expected, a considerable effect on traffic flow. Properly designed controls improve the efficiency of flow while poorly designed control mechanisms have the opposite effect. A detailed discussion on the effect of control mechanisms and their proper design procedures can be found in Chapters 4 and 5.

2.6 TERMINAL FACILITIES

For a traffic engineering system, terminal facilities include home garages, on-street parking areas, and off-street parking lots. To a traffic engineer, the design of on-street parking facilities and off-street parking lots are of interest. In general, on-street parking does not require additional infrastructure; however, it has considerable impact on the flow of traffic on the road. Off-street parking lots, on the other hand, require additional infrastructure. Chapter 5 provides a good discussion on the design of parking facilities.

EXERCISES

1. Perform some background study and write short notes on:
   (a) Lead and its relation to vehicular pollution
   (b) CNG and its role in reduction of pollutant emission from vehicles
   (c) Technologies to reduce evaporative emissions
   (d) Effect of frequent accelerations and decelerations on emission rates
2. Perform some background study and write short notes on:
   (a) Pollutants emitted by gasoline driven vehicles
   (b) Pollutants emitted by diesel driven vehicles
   (c) Pollutants emitted by CNG driven vehicles
   (d) Pollutants emitted by LPG driven vehicles
3. Undertake a survey of various classes of vehicles plying on the roads of your city and find out the various characteristics of these vehicle classes.
4. In your laboratory, perform an experiment to measure the reaction time of your fellow students for a simple stimulus (like a light turning red from green). Define a certain class interval (like 0 to 0.2 s, 0.21 to 0.4 s, etc.) and draw a histogram for

the reaction times noticed. Also determine the mean reaction time, the median reaction time, and the 85th percentile reaction time.

5. Take a small wooden block and glue a piece of cycle tyre tube on one of its surfaces. Locate a few road sections with varying slopes (you may get the steepest slopes on the sides of a speed bump). Place the block with the rubber side touching the road surface. Determine, at least a lower bound on the friction coefficient of a road surface from this experiment.
6. Determine the braking distance for the following situations: (i) a vehicle moving on a positive 2 per cent grade at an initial speed of 30 km/h, final speed 0 km/h; (ii) a vehicle moving on a –3 per cent grade at an initial speed of 50 km/h, final speed 0 km/h; and (iii) a vehicle moving on a level road at an initial speed of 40 km/h, final speed 0 km/h. Assume the coefficient of rolling friction to be 0.3.
7. Determine the stopping distances for the cases given in Exercise 6.

Highway Geometric Design

3.1 INTRODUCTION

This chapter presents the principles of designing the layout of roads. Such designing is commonly referred to as *geometric design of roads*. Proper designing of the layout of a road is important from two aspects: (i) it facilitates smooth flow of traffic and (ii) it improves safety. These improvements are derived from (i) good geometric design of direction changes in roads, (ii) good geometric design of slope changes in roads, and (iii) good delineation of desirable vehicular paths at confusing locations such as intersections.

Direction changes in roads are achieved by providing curves between two straight stretches in two different directions. Similarly, slope changes in roads are achieved by providing curves between two straight stretches at different gradients. Although, it is conceivable that two straight stretches which need to be joined have different directions as well as slopes, the basic principles of geometric design are generally illustrated in the case where roads with zero gradient change directions or where straight stretches change gradients. Layout design of road sections joining two roads with different directions is referred to as *geometric design of horizontal curves*. Layout design of road sections joining two roads with different gradients (or slopes) is referred to as *geometric design of vertical curves*. Layout design of road sections for the purpose of proper delineation of vehicular paths is referred to as *channelization design*.

In the following sections, the principles of each of the above three types of geometric designs are discussed. The details of laying of these geometric features on the road are not discussed here as they may be readily obtained in any introductory text on surveying. However, a brief description of the geometry of the cross-section of a typical road is provided.

3.2 TYPICAL ROAD CROSS-SECTION

In this section, the geometric features of a typical vertical cross-section of a road are

described. The description is primarily based on the general practice followed in India and around the world.

Figure 3.1 shows the various different space requirements for constructing a road. The ideal values for the different widths (indicated in the figure) for various road classes may be found in IRC:73–1980 [79]. The space requirements are stated bearing in mind the requirement in the future for upgradation of roads and possibilities of encroachments.

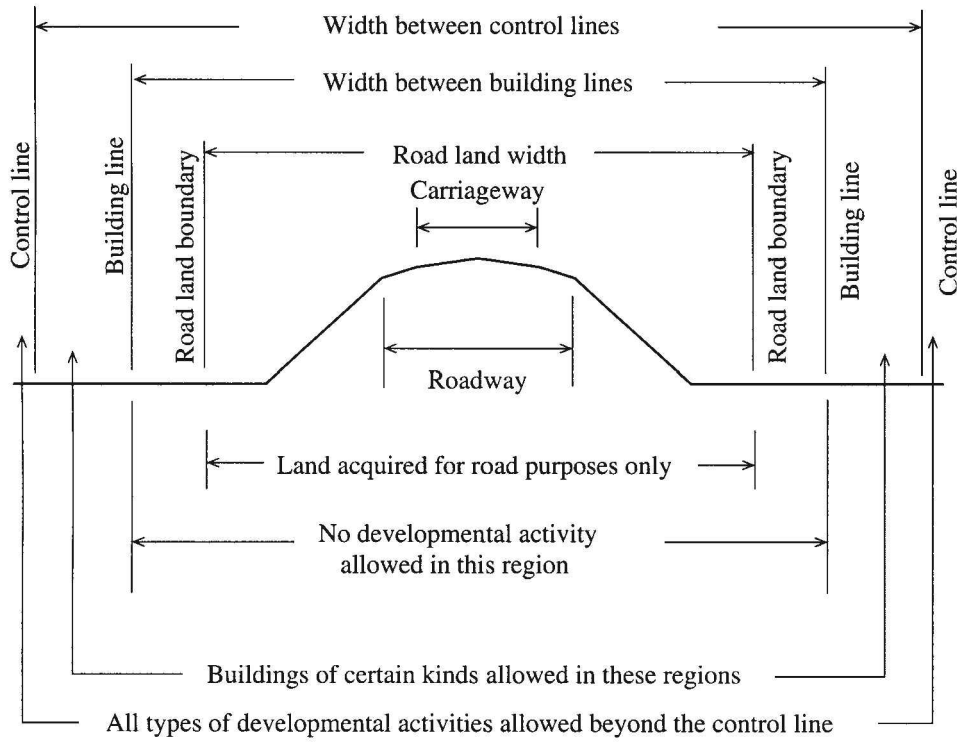


Figure 3.1 Land requirements for constructing a road.

Figure 3.2 shows two typical cross-sections of roads. The outward slope in the main carriageway is called *camber* and is provided to aid surface drainage. The ideal value for the camber, depending on the top surface of the road, generally ranges from 1.7% for good bituminous surfacing to about 4% for earth roads. The slope of the shoulders should be steeper than that of the camber. Details of the recommended slopes under Indian conditions may be found in IRC:73–1980 [79].

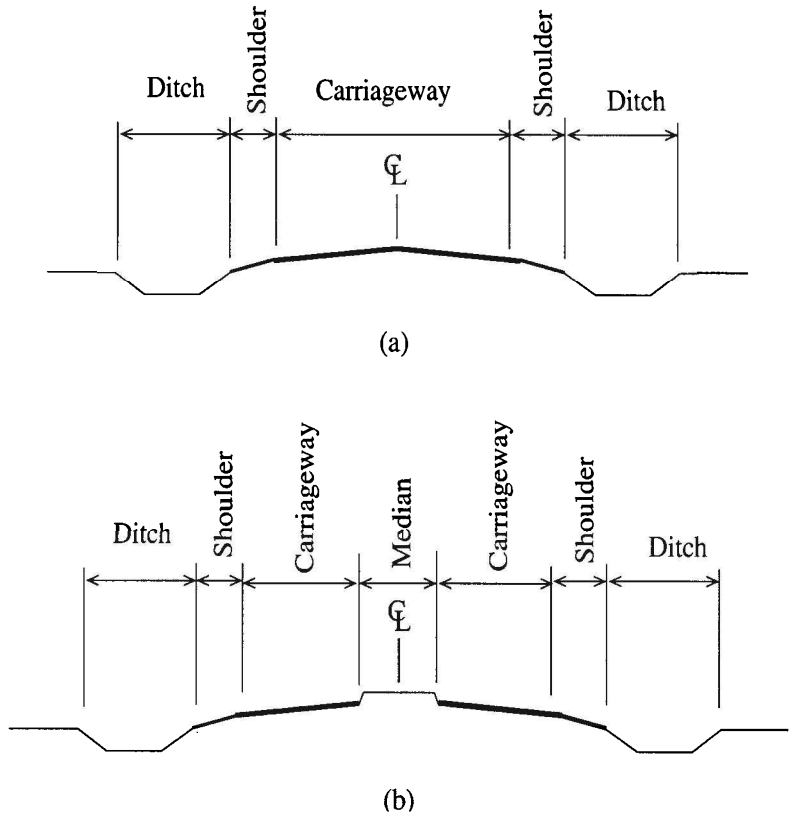


Figure 3.2 Typical cross-sections of (a) a two-lane undivided roadway and (b) a two-lane divided roadway with raised median.

3.3 HORIZONTAL CURVES

When a road changes its direction, generally a circular curve is provided to join the two straight stretches. The primary issue in the design of such a curve is to determine a safe radius for the curve. It should be noted that the circular motion of vehicles along the curve creates centrifugal accelerations. Such an acceleration (or a force) can slide a vehicle outwards or even overturn it if the radius of the curve is too small for the design speed of the road. For a curve of given radius, the speed at which it can be safely negotiated, can be increased by raising the outer edge of the road with respect to the inner edge thereby tilting the vehicle while it negotiates the curve. In this case, a component of the vehicle's weight helps in offsetting the centrifugal force. The raising of the outer edge is termed *superelevation*. Thus, simplistically speaking, there are two variables which need to be designed for a horizontal curve, namely the radius and the extent of superelevation. This is the topic of discussion in the next section.

Later, other issues like available sight distance on the curve and transition curves are also discussed. These issues are introduced as and when the need for understanding them arises.

3.3.1 Radius and Superelevation

The issue here is to determine the safe radius of the curve and its superelevation for the given design speed of the road. In order to understand how this can be determined, consider the cross-section of the curve shown in Figure 3.3. In this figure, the vehicle is moving perpendicular to the plane of the paper.

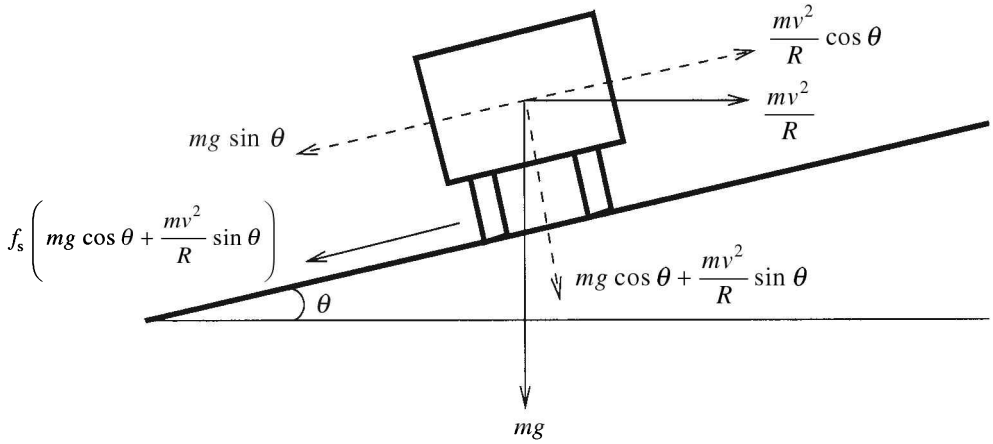


Figure 3.3 Free body diagram of a vehicle negotiating a circular horizontal curve.

From Figure 3.3, it can be seen that for the vehicle to be steady (i.e. the vehicle does not slide outwards) the following relation should hold:

$$\frac{mv^2}{R}\cos\theta = mg\sin\theta + f_s\left(mg\cos\theta + \frac{mv^2}{R}\sin\theta\right) \tag{3.1}$$

where v is the design speed of the road, R is the radius of the curve, θ is the angle by which the curve is tilted, i.e. it is the extent of superelevation (usually, tan θ is referred to as e, the superelevation rate), f_s is the coefficient of side friction offerred by the road surface, g is the acceleration due to gravity, and m is the mass of the vehicle.

On simplification, Eq. (3.1) gives the relation

$$\frac{v^2}{R}(1 - ef_s) = g(e + f_s)$$

Hence,

$$R = \frac{v^2}{g}\frac{1 - ef_s}{e + f_s} \tag{3.2}$$

In geometric design, however, the term ef_s is often ignored as it is generally very close to zero for practical values of e and f_s. Next, the issue of practical values (or at least limits) of e and f_s is discussed.

Practical values of the coefficient of side friction

The coefficient of side friction is a property of the tyre material, the road surface condition and the speed. However, the question to be answered here is not what the coefficient of side friction is in a particular condition but what is the value we should use for the purposes of design. The IRC codes (see for example, IRC:73–1980 [79] and IRC: 86–1983 [80]) suggest a single value of 0.15. However, AASHTO [3] gives a more detailed suggestion on the value of f_s based on many empirical studies quoted by AASHTO. According to AASHTO recommendation, depending on the design speed of the curve, we shall use the value of f_s obtained from the following equation

$$f_s = \begin{cases} 0.19 - 0.0006v & 30 \text{ kmph} \le v \le 80 \text{ kmph} \\ 0.24 - 0.0012v & 80 \text{ kmph} \le v \le 110 \text{ kmph} \end{cases} \tag{3.3}$$

AASHTO modifies these f_s values for low-speed urban roads and states that values as high as 0.3 can also be used for such roads.

Practical values of the superelevation rate

The value of the superelevation rate that can be used is dependent on many factors such as the frequency of snowfall (in cold countries), the type of terrain, the type of area (urban or rural), and the frequency of slow-moving vehicles. Depending on these conditions, various codes suggest different maximum levels of superelevation rates that can be used. However, before proceeding to look at the suggestions of different codes, a simple analysis to study the effect of superelevation on slow-moving or static vehicles is undertaken.

Consider a stopped vehicle on the curve as shown in Figure 3.4. It is clear from the figure that if e is very high, then the vehicle will slide inwards—an undesirable situation. To prevent such a situation,

$$mg \sin \theta \le f_{s,\max} mg \cos \theta$$

where $f_{s,max}$ is the maximum coefficient of side friction that can be achieved. The above equation means that the value of e that can be used should satisfy the inequality

$$e < f_{s,\max} \tag{3.4}$$

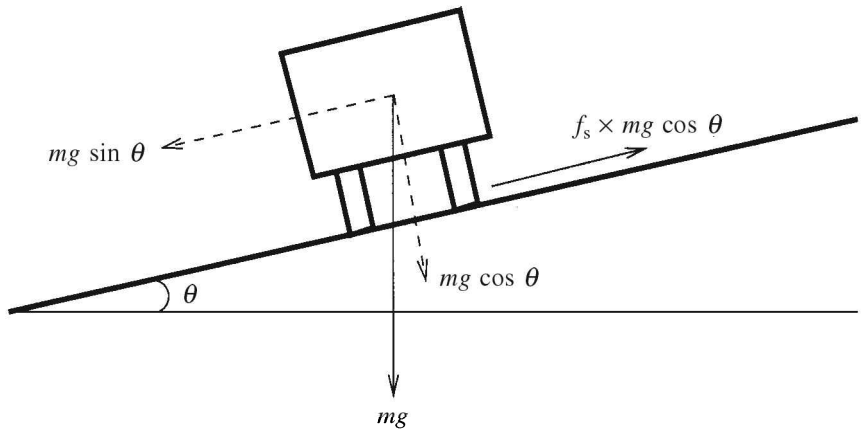


Figure 3.4 Free body diagram of a static vehicle on a circular horizontal curve.

For very low speeds, AASHTO [3] suggests maximum friction value close to 0.3, and IRC suggests (recall, IRC suggestions are independent of speed) a value of 0.15. This indicates that the maximum value of e, (i.e. e_{max}), that can be used is at least as high as 0.15. However, the practical maximum limits of e, as suggested by IRC and AASHTO, are much lower than this value. AASHTO suggests using e values less than 0.1 (for expressways through open area) with several other lower limiting values which depend on the terrain and environmental factors. Similarly IRC (see IRC:73–1980 [79] and IRC: 86–1983 [80]) suggests the following maximum limits on e value: 0.07 for plain and rolling terrain and for snow-bound areas, 0.1 for hilly terrains (without snow), and 0.04 for urban roads with frequent intersections.

Methods of attaining superelevation

The superelevation (or banking) of the road can be obtained in many ways. The most common way of banking the road is to revolve the road surface about the centre-line of the road. In this section, this method is explained by drawing the longitudinal profile of a road with a superelevated curve. The longitudinal profile of a road is a view of the road similar to an elevation drawing, the only difference being that the distances in the horizontal direction are always made equal to the true distances.

Figure 3.5(b) shows the longitudinal profile of a road with a superelevated curve. The road (in its plan view) is shown in Figure 3.5(a). The centre-line profile indicates the level of the centre line of the road on the straight stretch as well as on the curve. Since the road surface is rotated around the centre line, there is no change in the vertical coordinate of the centre line. Before the point marked A in the profile, the outer edge and the inner edge of the road are both slightly below the centre line (as the straight road has a normal camber). From point A, the outer-half of the road is rotated till the outer edge reaches the same level as the centre line. This is achieved at point B. The distance

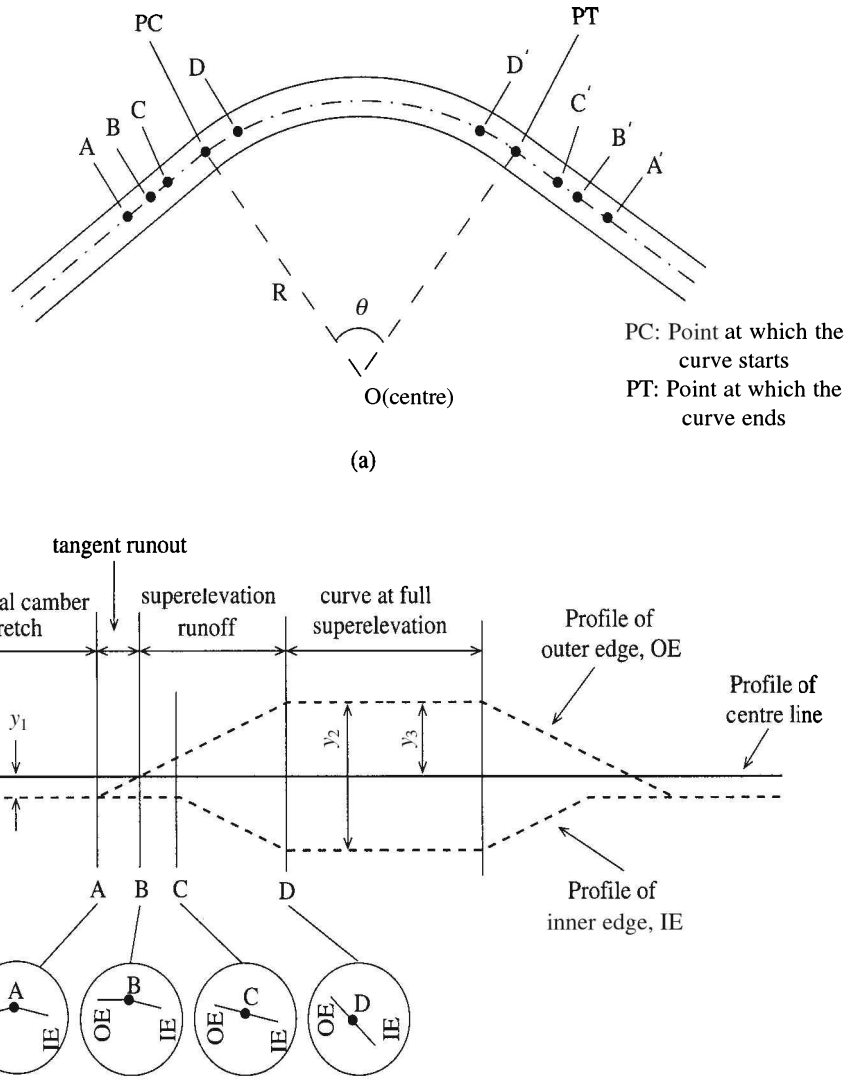


Figure 3.5 (a) Plan view of a road with a simple circular curve; (b) longitudinal profile of the road in (a).

required to complete this rotation is called the *tangent runout*. From point B, the outer edge is further rotated till this edge, the centre line and the inner edge all lie on the same plane. This point is named point C in the figure. From this point onwards, the entire plane is rotated about the centre line till the slope of the plane becomes equal to the superelevation rate. This is achieved at point D. The distance between points B and D

is called the *superelevation runoff* or simply *runoff*. From point D onwards, the plane is kept at the same slope till it is time to reverse the process in order to achieve normal camber on the straight stretch at point A′. The profile is symmetric between points A and A′. Further, the figure shows the transverse sections of the road at points A through D in bubbles marked under A, B, C, and D in part (b) of the figure.

The rate at which these rotations can take place are specified through specification of the *runoff* for different values of design speed, superelevation rate, and road width. The distance of the *tangent runout* is determined by assuming the same rate of rotation as implied by the *runoff* in a given case. That is, the rate of change of slope (of the road) from A to B is the same as that from B to D. That is why the line denoting the profile of the outer edge is a straight line in Figure 3.5(b). In the following, all the distances required to draw a longitudinal profile of a curved road (as shown in the figure) are given.

The distance y_1 is the height between either of the edges and the centre line in the normal camber section of the road. If the normal camber slope is s_c and the width of the entire road (that is, the distance between the inner and the outer edges) is w, then since s_c is small

$$y_1 = \frac{w}{2} s_c \tag{3.5}$$

The distance y_2 is the difference in heights between the outer and inner edges of the road at the fully superelevated section of the roadway. It is equal to twice the value of y_3. The distance y_3 is the difference in heights of the centre line and the outer edge at the fully superelevated section of the roadway. If the superelevation rate is e, then since the values of e are small

$$y_3 = \frac{w}{2} e \tag{3.6}$$

The superelevation runoff r is generally specified in the codes (see for example, IRC: 73–1980 [79] or IRC:38–1988 [85] or AASHTO [3]) and is dependent on e, the design speed v, and the width of the roadway. However, the bottom line in the determination of the superelevation runoff is that the slope of the outer edge with respect to the centre line should be less than a permissible limit. That is, the slope of the line representing the profile of the outer edge between points B and D should be within a permissible limit. The limit is generally 1 in 150 for roads in plain and rolling terrain and about 1 in 60 in mountainous terrain. Denoting this slope as s_r, we can obtain r as

$$r = \frac{y_3}{s_r} \tag{3.7}$$

The tangent runout t_r is calculated in a similar fashion and is given by

$$t_r = \frac{y_1}{s_r} \tag{3.8}$$

Before leaving this section it should be pointed out that when a straight road is joined by a simple circular curve, then about two-thirds of the runoff is provided on the straight section of the road and one-third on the circular curve. However, if a transition curve (see latter parts of this chapter for a discussion on transition curves) joins the circular curve to the straight section, then the entire superelevation runoff is provided on the transition curve. In either case, the tangent runout is provided on the straight section of the road.

3.3.2 Available Sight Distance

Figure 3.6 shows a typical horizontal curve joining two straight edges (the straight edges are generally referred to as tangents). The first point (in terms of the distance from some benchmark point) at which the curve starts is referred to as the point of curvature (PC), and the second point at which the curve ends and the tangent starts is referred to as the point of tangency (PT). The hypothetical point where the roads would have intersected

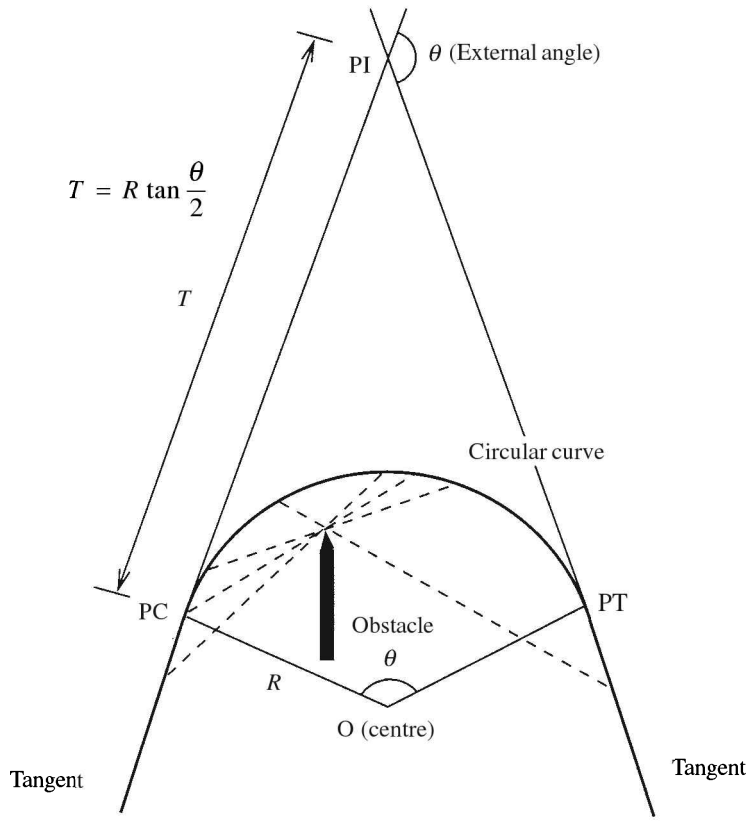


Figure 3.6 A typical circular horizontal curve with sight obstruction.

is referred to as the point of intersection (PI). The angle between the roads θ is called the external angle. The figure also shows an obstacle, which limits the extent of the road that is visible. The dotted lines are different lines of sight of the driver from different locations on the road. For any given line of sight, the corresponding distance along the road (or along the centre line of the inner-most lane of the road) is the sight distance available to the driver.

Obviously, in the vicinity of the obstacle, the shortest available sight distance will correspond to the shortest line of sight. It can be shown that the shortest line of sight (or the chord going through the corner of the obstacle) will correspond to that chord whose middle point is the corner point of the obstacle. Given this fact, it can be easily shown that (see Figure 3.7) the least available sight distance (ASD_{hc}) for a horizontal curve is given by

$$ASD_{hc} = 2R \cos^{-1}\left(1 - \frac{M}{R}\right) \tag{3.9}$$

where M is the middle ordinate distance of the obstacle from the curve and R is the radius of the curve.

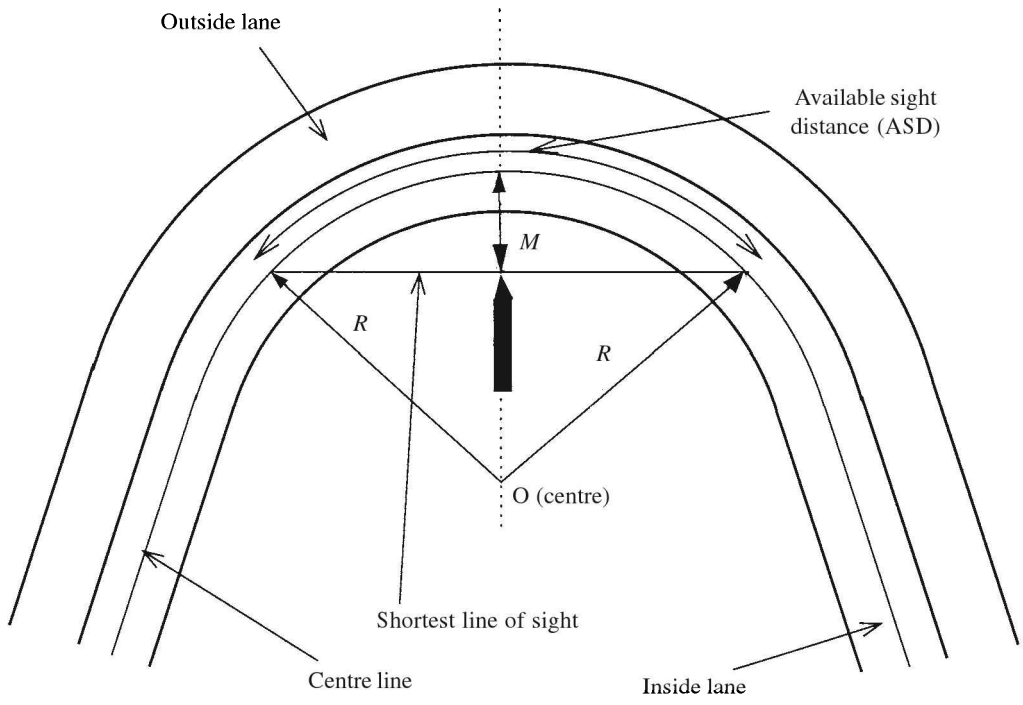


Figure 3.7 Available sight distance on horizontal curves (ASD_{hc}).

While designing a horizontal curve, care should be taken to see that the available sight distance is greater than the distance required for stopping (discussed in Chapter 2). If not, then either the speed limit is reduced or the obstacle is moved. Note that the curve could also be moved more towards the PI (and hence away from the obstacle) but this would mean that the radius of the curve will reduce. Sometimes, when the least radius [determined by assuming the maximum values of e and f in Eq. (3.2)] is being provided, moving the curve towards the PI is not a viable option. The curve could also be moved away from the PI so as to have the obstacle outside the curve, but this may not always be practical. Similarly, if passing (or overtaking) is to be allowed on the curve then it must be ensured that the available sight distance is greater than the required overtaking distance (as discussed in Chapter 2).

EXAMPLE 3.1

Determine the design radius of a horizontal curve (in plain terrain with rural settings) joining two straight stretches which meet at an external angle of 90°. The point of intersection is at the coordinates (1500, 1500). All distances are measured in metres. The critical corner of an important building (which cannot be moved) is at (1495, 1469). Assume IRC recommendations for all design parameters and a design speed of 40 kmph (or 11.11 m/s). The benchmark point from where the coordinates have been determined is an upstream point on the road which lies to the left of the building when pointing towards the PI from the building.

Solution

First, the minimum radius R_{min} of the curve which can be provided in this case is calculated taking the maximum allowable values of e and f_s as per IRC specifications [see Eq. (3.2)]. Therefore,

$$R_{min} = \frac{(11.11)^2}{9.81(0.07 + 0.15)} = 57.19 \text{ m}$$

Now consider the data shown in Figure 3.8. The benchmark point (0, 0) lies on the road marked A. Hence the equation of the line representing Road A is $y = x$. The distance of PC from PI is given by $R \tan(\theta/2)$. Since $\theta = 90°$, the distance of PC from PI is 57.19 m. Thus, PC is at coordinates of (1459.56, 1459.56). Similarly, the equation of the road marked B (which is at an angle 90° with road A and passes through PI) is $y = 3000 - x$ and the coordinates of PT (which again is at a distance of $R \tan(\theta/2)$ from PI) are (1540.44, 1459.56).

From the coordinates of PC and PT and the fact that lines from the centre of the circular curve to PC and PT will be at right angles to the roads A and B, respectively, the coordinates of the centre can be easily obtained. The coordinates of centre are (1500, 1419.12).

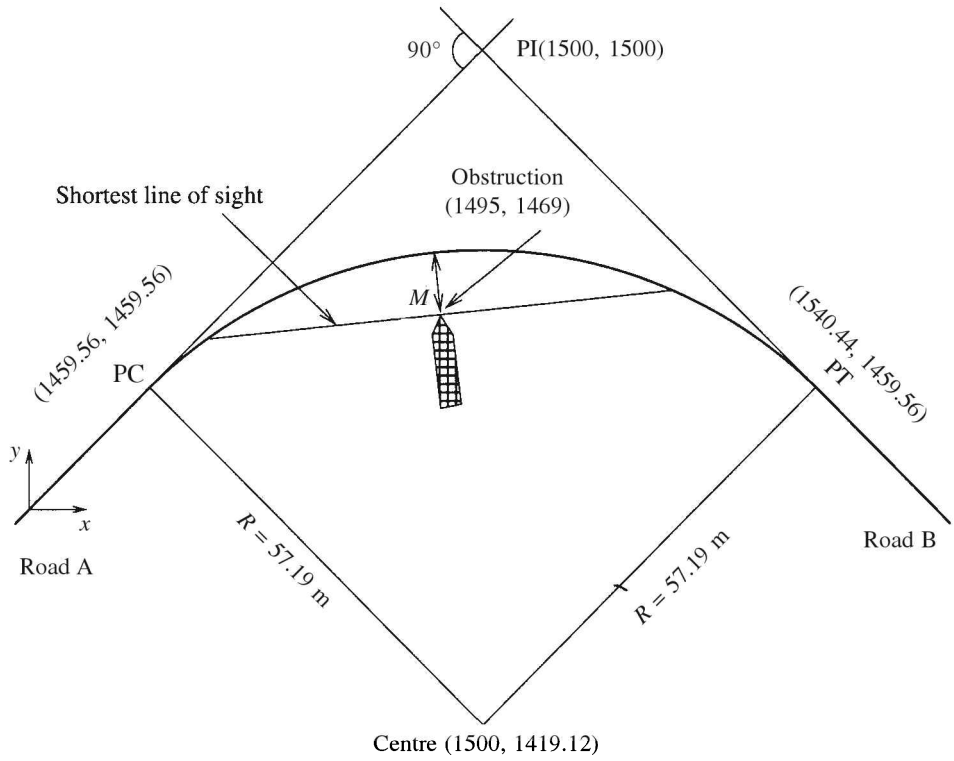


Figure 3.8 The horizontal curve designed in Example 3.1.

Hence, M the middle ordinate distance of the obstacle (for the shortest line of sight) is R – (distance of the obstacle from the centre). This distance can be obtained from the coordinates of the obstacle, which are (1495, 1469), and the coordinates of the centre. The distance so obtained is 50.13 m. Hence, M = 7.06 m. Thus, from Eq. (3.9), ASD_{hc} = 57.43 m.

From the discussions on the required stopping distances in Chapter 2, it can be seen that the stopping sight distance requirement for a design speed of 40 kmph, a perception–reaction time of 2.5 s (as per IRC specifications), and a coefficient of rolling friction of 0.4, is 43.51 m or 44 m. Hence adequate sight distance for stopping is available. However, overtaking cannot be allowed on the curve, since the required overtaking sight distance is greater than 57.43 m (see IRC:73–1980 [79]).

3.3.3 Transition Curves

Transition curves are provided in order to gradually introduce the centrifugal acceleration that drivers experience when negotiating a curve. The gradual introduction of the centrifugal acceleration is achieved by providing a clothoid spiral as the transition curve joining the straight stretch with the circular curve. The clothoid spiral has the

property that the curvature (or the reciprocal of the radius) varies linearly with the length of the curve. At the beginning, the transition curve has a zero curvature (or infinite radius) and at the point where it meets the circular curve, it has a curvature equal to the reciprocal of the radius of the circular curve.

The length of the transition curve, the only design variable for the transition curve, is determined by setting a permissible limit on the jerk (or the rate of change of acceleration) experienced by drivers while negotiating the curve. The AASHTO [3] suggests the use of a value between 0.3 and 0.9 ft/s^3. The IRC code (for example, refer to [79]) suggests the use of a value between 0.5 and 0.8 m/s^3.

The length of the curve can be easily determined using the following analysis. The centrifugal acceleration changes from zero (at the beginning of the transition curve) to v^2/R at the end of the transition curve. Since the curvature increases at a constant rate along the length of the transition curve, the rate of increase in the centrifugal acceleration is also constant along the length of the curve. Assuming that drivers move at a constant speed along the transition curve of length L_T, the jerk J faced by drivers while on the curve is given by

$$J = \frac{v^2}{R} \div \frac{L_T}{v} \tag{3.10}$$

From this, we can say that in order to keep the jerk less than a permissible limit J^*, the minimum length of the transition curve, L_T^{min}, that needs to be provided is

$$L_T^{min} = \frac{v^3}{RJ^*} \tag{3.11}$$

However, when transition curves are provided the entire superelevation runoff (see Section 3.3.1) is provided on the transition curve. Hence, the length of transition curve to be provided should be the greater of the superelevation runoff and L_T^{min}.

Circular horizontal curve combined with transition curves

If a circular horizontal curve is connected to the straight edges through transition curves, then the geometry of the combined curve differs from the case when there are no transition curves. In this section, the geometric features of the combined circular and transition curves are illustrated. Figure 3.9 shows such a combined curve.

In Figure 3.9, the Points 1, 2, 3, 4, 5, and 6 are, respectively: (i) ST, the point at which the transition curve leaves the straight stretch, (ii) PI, the point at which the circular curve of radius R would have joined an imaginary road parallel to the actual straight road, (iii) TC, the point at which the transition curve meets the circular curve, (iv) CT, the point at which the circular curve ends and another transition curve starts, (v) a point similar to Point 2, and (vi) TS, the point at which the transition curve meets the straight road. The quantities of interest are: (i) the coordinates (X, Y) of any point on the transition curve as measured from ST with the actual straight road as the abscissa,

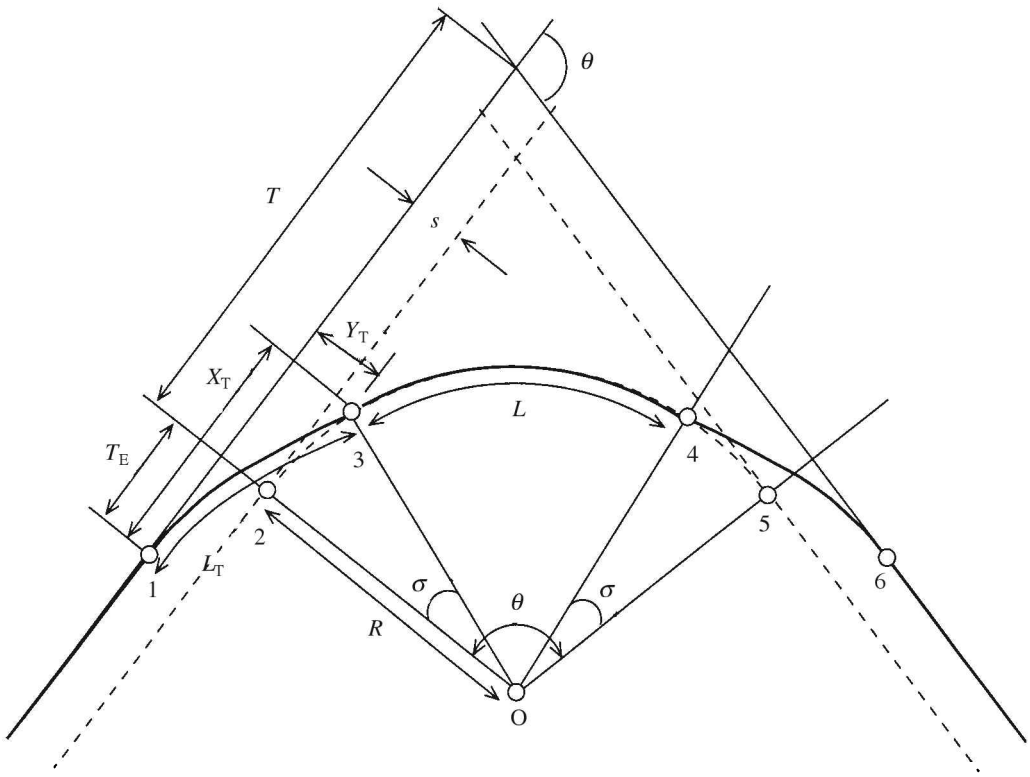


Figure 3.9 Combined circular and transition curve.

(ii) the shift s between the actual straight roads and a parallel line which is tangential to the extended circular curve, (iii) the distance T from the point of intersection PI, (iv) the distance T_E, (v) the angle σ, (vi) the length of the circular curve L, and (vii) the coordinates, X_T and Y_T, of the end point of the transition curve, TC. Mathematically, these quantities are as stated below.

The coordinates (X, Y) of any point at distance ℓ (measured along the curve) from ST are given by

$$X = \ell - \frac{\ell^5}{40R^2L_T^2} + \frac{\ell^9}{3456R^4L_T^4} + \cdots \tag{3.12}$$

$$Y = \frac{\ell^3}{6RL_T} - \frac{\ell^7}{336R^3L_T^3} + \frac{\ell^{11}}{42240R^5L_T^5} + \cdots \tag{3.13}$$

The shift s is twice the value of Y at length of transition curve equal to $L_T/2$. A good approximation of s is given by

$$s \approx \frac{L_T^2}{24R} \tag{3.14}$$

The distance T can be easily seen to be the following:

$$T = (R + s) \tan\left(\frac{\theta}{2}\right) \tag{3.15}$$

The distance T_E can be reasonably approximated to $L_T/2$. We could also use the fact that the length of the transition curve becomes half its total length when $X = T_E$ and then determine the value of X using the relation of the coordinates (X, Y) to the length of the transition curve.

The spiral angle σ in radians can be approximated quite well by assuming that the arc length from TC (Point 3) to Point 2 is equal to the transition curve length between the same points. Since the length of the transition curve between these points is $L_T/2$, the angle σ can be obtained as

$$\sigma \approx \frac{L_T}{2R} \tag{3.16}$$

The length of the circular curve L is given by

$$L = R(\theta - 2\sigma) \tag{3.17}$$

The quantities X_T and Y_T can be computed precisely by substituting L_T for ℓ in Eqs. (3.12) and (3.13). However, they can also be closely approximated using the following:

$$X_T \approx L_T - \frac{L_T^3}{40R^2} \tag{3.18}$$

$$Y_T \approx \frac{L_T^2}{6R} \tag{3.19}$$

Before leaving this section it may be pointed out that since the reciprocal of the radius of the transition curve varies linearly over its length from zero to $1/R$, the radius, R_T^{ℓ} at any length ℓ can be obtained as

$$R_T^{\ell} = \frac{\ell}{RL_T} \tag{3.20}$$

3.3.4 Curve Widening

As discussed in Chapter 2, when vehicles take a turn, the front and rear axles do not follow the same path. Owing to this, the effective width of vehicles on turns is greater than that on straight edges. Further, vehicles while negotiating a curve tend to move towards the outer edge. Both these factors necessitate that the width of the road section on the curve be larger than that on the straight stretches. Both IRC codes (see IRC:73–1980 [79] and IRC:86–1983 [80]) and AASHTO [3] suggest how the curve widening should be achieved. The AASHTO guidelines are somewhat more

comprehensive than the IRC guidelines. The interested reader may refer to both IRC codes and AASHTO guidelines for a more complete understanding of this topic.

3.4 VERTICAL CURVES

When roads change grades, generally parabolic curves are provided to join the two straight roads (at different grades). Parabolic curves are provided because the rate of change of slope on the curve is constant. The primary issue in the design of such a curve is to determine a length of the curve which provides ample sight distance. It should be noted that, unlike in the case of horizontal curves, the primary concern here is the availability of sight distance. This is because, unlike horizontal curves, vertical curves by their very geometry impose restrictions on the sight distance availability. This can be seen in Figure 3.10. Here, two types of vertical curves are shown together with some of the terminology used in the context of vertical curves. Whenever the elevation of the

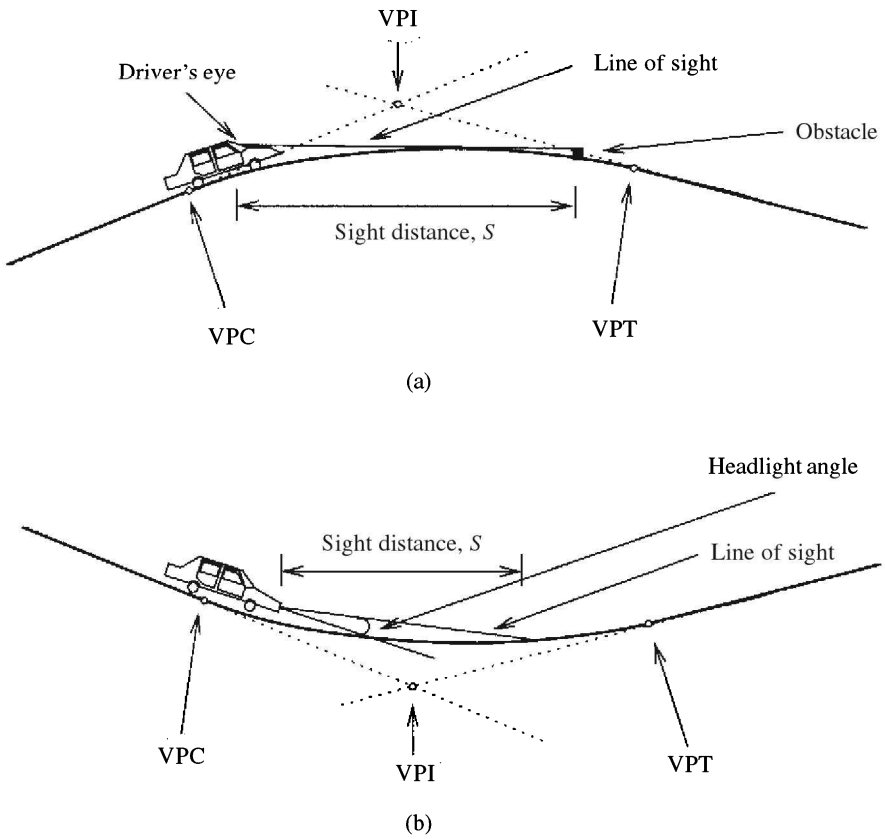


Figure 3.10 (a) A *crest* vertical curve, (b) a *sag* vertical curve.

vertical point of intersection, VPI, is greater than or equal to the elevation of any point on the curve, the curve is referred to as a *crest* or *summit* curve [see Fig. 3.10(a)]. Whenever the elevation of the *vertical point of intersection*, VPI, is less than or equal to the elevation of any point on the curve, the curve is referred to as a *sag* or *valley* curve [see Fig. 3.10(b)]. Further, where the first straight road meets the curve, the point is referred to as the *vertical point of curvature*, VPC. The point at which the curve meets the second straight stretch is referred to as the *vertical point of tangency*, VPT.

3.4.1 Length of Vertical Curves

The length L of a vertical curve, measured along the horizontal (i.e. L is equal to the horizontal distance between VPC and VPT), as required to provide a minimum sight distance S is given by the following equations [see Eqs. (3.21) and (3.22) below]. These equations, which are different for *crest* curves and *sag* curves are derived from basic geometry and trigonometry (the reader may refer to IRC Special Publication 23 [255] for the derivations). It may be noted here that if the grade changes between the straight edges are small, then no vertical curve is required from the sight distance standpoint. In most of these cases, however, some minimum length of curve is provided from the viewpoints of comfort and aesthetics. For Indian conditions, these minimum lengths are specified in IRC Special Publication 23 [255].

Length of crest curves

It can be shown that the minimum length of curve L, as required to provide a sight distance of S [also see Figure 3.10(a)] is given by

$$L = \begin{cases} \dfrac{AS^2}{2\left(\sqrt{h_1} + \sqrt{h_2}\right)^2} & \text{when } S \leq L \\ 2S - \dfrac{2\left(\sqrt{h_1} + \sqrt{h_2}\right)^2}{A} & \text{when } S \geq L \end{cases} \tag{3.21}$$

where

$A = |g_1 - g_2|$ with g_1 and g_2 being the grades of the two straight stretches being joined by the vertical curve expressed in fractions and not percentages.

S is the sight distance provided by a parabolic vertical crest curve of length L; this is made either equal to (i) the passing (or overtaking) distance required if overtaking is to be allowed, or equal to (ii) the stopping distance required if overtaking is not to be allowed.

h_1 is the height of the driver's eye; this is generally assumed to be 1.2 m.

h_2 is the height of the obstacle; this is generally assumed to be (i) 0.15 m if the stopping distance requirement is used for S, or (ii) 1.2 m if the passing (or overtaking) distance requirement is used for S.

The procedure for determining the lengths of curves is further illustrated through the solved example following Section 3.4.2.

Length of sag curves

Unlike the crest vertical curves which impose restriction on sight distance due to the raised elevation of the middle part of the curve, sag vertical curves impose restriction on sight distance only at night when the road is illuminated by the headlights. This can be clearly seen in Figure 3.10(b).

It can be shown that the minimum length of curve L, as required to provide a sight distance of S is given by

$$L = \begin{cases} \dfrac{AS^2}{2(h_3 + S\tan\beta)} & \text{when } S \leq L \\ 2S - \dfrac{2(h_3 + S\tan\beta)}{A} & \text{when } S \geq L \end{cases} \tag{3.22}$$

where

$A = |g_1 - g_2|$ with g_1 and g_2 being the grades of the two straight stretches being joined by the vertical curve expressed in fractions and not percentages.

S is the sight distance provided by a parabolic vertical sag curve of length L; this is made either equal to (i) the passing (or overtaking) distance required if overtaking is to be allowed, or equal to (ii) the stopping distance required if overtaking is not to be allowed.

h_3 is the height of the headlight; this is generally assumed to be 0.75 m.

β is the upward headlight angle as shown in Figure 3.10(b); the value of β is generally taken as 1°.

3.4.2 Geometry of Curves

The various aspects related to the geometry of the curve are explained through Figure 3.11. In the figure, two straight edges with gradients g_1 and g_2 expressed as fractions are joined by a parabolic curve of the form given in equation

$$y = a + bx + cx^2 \tag{3.23}$$

where x and y are measured from the VPC as the origin.

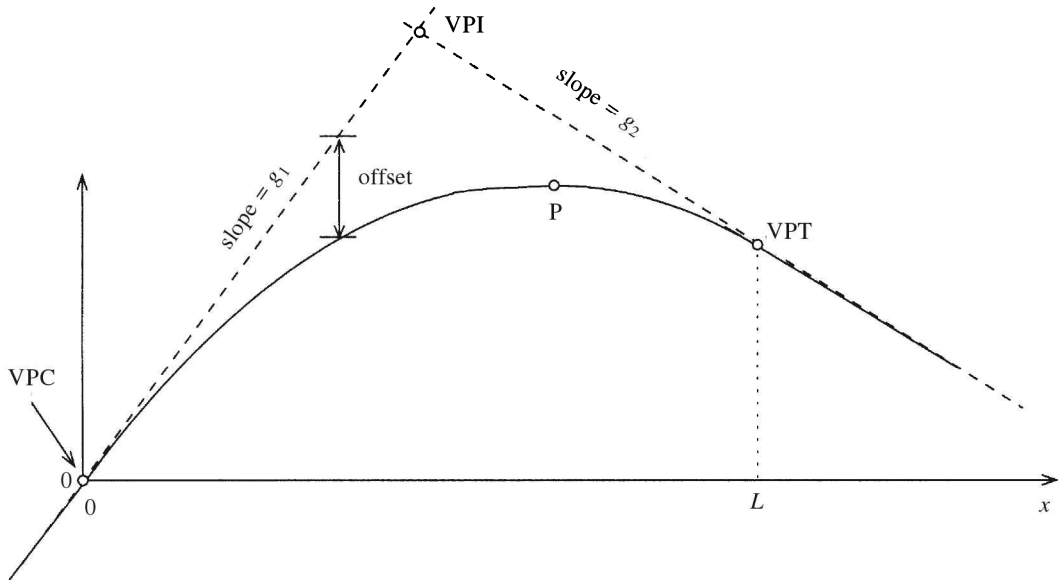


Figure 3.11 Geometry of vertical curves.

Equating the slope of the curve (which is $b + 2cx$) to g_1 at $x = 0$ and to g_2 at $x = L$, and using the fact that $y = 0$ when $x = 0$, the following equation for the curve can be obtained

$$y = g_1 x + \frac{g_2 - g_1}{2L} x^2 \tag{3.24}$$

If the coordinates of VPC are (x_{VPC}, y_{VPC}) instead of (0, 0) as assumed in Eq. (3.24), then the coordinates of any point on the curve can be obtained easily as $(x + x_{VPC}, y + y_{VPC})$, where (x, y) are the coordinates of the same point obtained from Eq. (3.24).

Another point of interest is the location of VPI in relation to either VPC or VPT. Note that if this location is known, then we can easily determine either VPC or VPT. VPI is a point where the two straight stretches meet (or would have met had they been extended). The equation of the first straight stretch is $y = g_1 x$; the equation of the second straight stretch (which has a slope of g_2 and passes through the point $(L, g_1 L + (g_2 - g_1)L/2)$—the coordinates of VPT) is $y = g_2 x + (g_1 - g_2)(L/2)$. Equating these two equations of straight lines, we can easily find the coordinates of VPI as $(L/2, g_1 L/2)$. These coordinates show that irrespective of g_1 and g_2 the VPI always lies vertically above or below the middle of the curve (recall that the length of the curve is L and is measured along the horizontal).

Another feature of interest is the offset of the curve as shown in Figure 3.11. The offset from the first tangent at any value of x can be determined by subtracting the y

coordinate of the first tangent from the y coordinate of the curve at the given value of x. The offset obtained is, thus

$$\text{offset at } x = \frac{g_2 - g_1}{2L} x^2 \tag{3.25}$$

The value of the offset at VPI, i.e. O_{VPI} can be obtained from Eq. (3.25), by substituting $x = L/2$. Thus,

$$O_{\text{VPI}} = \frac{g_2 - g_1}{8} L \tag{3.26}$$

Since in crest curves, the VPI lies above the curve, O_{VPI} must be negative and vice versa for sag curves (recall the definition of crest and sag curves given earlier). This together with the above equation for O_{VPI} indicates that if $g_2 - g_1$ is negative then the curve will be a crest curve and if $g_2 - g_1$ is positive then the curve will be a sag curve.

The only other quantity of interest that remains to be discussed is the determination of the highest point (in the case of crest curves; for example see point P in the figure) and the lowest point (in the case of the sag curves). The highest and the lowest points have the same characteristic in that the slope of the curve at both the points is zero. Using this fact, the highest/lowest point will be at a point where

$$x = \frac{g_1 L}{g_1 - g_2} \tag{3.27}$$

However, x could be greater than L or less than 0, indicating that the highest/lowest point on the curve is beyond the range in which the curve is being actually provided on the field. Thus, using the fact that the curve is monotonic on either side of the point where the slope of the curve is zero, the following equation for the location of the highest/lowest point, $x_{\text{h/l}}$, can be written as:

$$x_{\text{h/l}} = \begin{cases} 0 & \text{if } (g_1 L)/(g_1 - g_2) < 0 \\ L & \text{if } (g_1 L)/(g_1 - g_2) > L \\ (g_1 L)/(g_1 - g_2) & \text{otherwise} \end{cases} \tag{3.28}$$

The elevation of the highest/lowest point can be calculated from the offset of that point from the first tangent using the relation given earlier.

The procedure to layout a vertical curve can be found in any introductory book on surveying and, therefore, is not discussed here. The following two examples, one each for crest and sag curves, illustrate the various concepts spelled out in this subsection.

EXAMPLE 3.2

A vertical curve is required to join a road with + 3% grade to a road with – 2.5% grade. The design speed of the road is 100 kmph. The VPI is located at coordinates (1000, 100). Further, the midpoint of an overhead electric transmission line of width 5 m crosses the road at a distance of 1100 m and elevation of 118 m. Determine the length of the vertical curve so that a stopping sight distance (i.e. the sight distance required for stopping safely) of 180 m (required for a design speed of 100 kmph) is available. Also, determine (i) the location of VPC and (ii) the minimum clearance from the transmission line. All distances are measured in metres from the same benchmark point.

Solution

$$g_1 = +3/100 = +0.03;\ g_2 = -2.5/100 = -0.025$$

$$g_2 - g_1 = -0.025 - 0.03 = -0.055 \text{ and } A = |g_2 - g_1| = 0.055$$

Since $g_2 - g_1$ is negative, the curve is a crest curve. The length of the curve required to provide the adequate sight distance for stopping can be obtained as follows:

Assume $h_1 = 1.2$ m, $h_2 = 0.15$ m. Further, assume $S \geq L$, then

$$L = 2 \times 180 - \frac{2\left(\sqrt{1.2} + \sqrt{0.15}\right)^2}{0.055}$$

or

$$L = 280 \text{ m}$$

In this case, therefore, S (= 180 m) is not greater than or equal to L. Hence, the earlier calculation is not valid. Therefore, assuming $S \leq L$, we have

$$L = \frac{0.055 \times 180^2}{2\left(\sqrt{1.2} + \sqrt{0.15}\right)^2}$$

or

$$L = 405 \text{ m}$$

Since in this case S is less than L, the design value of L is 405 m. Further, the minimum value of the length L, suggested by IRC Special Publication 23 [255] is 60 m. Hence, the final design value remains as 405 m.

The location of the VPC is calculated as follows:

(i) The horizontal distance of VPC from the benchmark is given as

$$\text{horz. dist. of VPC} = \text{horz. dist. of VPI} - \frac{L}{2}$$

$$= 1000 - \frac{405}{2} = 797.5 \text{ m}$$

(ii) The elevation of VPC is given as

$$\text{elev. of VPC} = \text{elev. of VPI} - \frac{L}{2}g_1$$

$$= 100 - \frac{405}{2}\,0.03 = 93.9 \text{ m}$$

Hence the coordinates of VPC are (797.5, 93.9).

Assuming VPC to be the origin, the transmission line can be thought of as a line (of width 5 m) parallel to the x-axis between the points (300, 24.1) and (305, 24.1). The question at hand, therefore, is to determine what is the highest elevation of the curve between the x-coordinates of 300 and 305.

With respect to the VPC as the origin, the highest point of the curve will be at the x-coordinate value of

$$\frac{g_1 L}{g_1 - g_2} = \frac{0.03 \times 405}{0.03 - (-0.025)} = 221 \text{ m}$$

From this, it can be said that the curve is steadily falling from the x-coordinate value of 300 to 305. Hence, the highest point in this region will be at the x-coordinate value of 300. The elevation at this point (assuming VPC to be the origin) is

$$\text{elev. of curve at } x \text{ equal to } 300 = \text{elev. of first tangent at } x + \text{offset at } x$$

$$= g_1 300 + \frac{g_2 - g_1}{2L}300^2$$

$$= 0.03 \times 300 + \frac{-0.025 - 0.03}{2 \times 405}300^2$$

$$= 9 - 6.1 = 2.9 \text{ m}$$

Therefore, the minimum clearance from the transmission line is 24.1 – 2.9 = 21.2 m.

EXAMPLE 3.3

A vertical curve joins a + 0.5% grade with a + 3.5% grade; the VPI is at coordinates (500, 50) from a benchmark point. Due to certain other traffic flow related considerations, a passing (or overtaking) sight distance (i.e. the sight distance required for overtaking safely) of 470 m is required on the curve. Determine (i) the length of the curve, (ii) the coordinates of VPC and VPT, and (iii) the coordinates of the lowest and the highest points on the curve. Assume a design speed of 80 kmph.

Solution

$$g_1 = +0.5/100 = +0.005;\ g_2 = +3.5/100 = 0.035$$

$$g_2 - g_1 = 0.035 - 0.005 = 0.03 \text{ and } A = 0.03$$

Since $g_2 - g_1$ is positive, the curve is a sag curve. The length of the curve required to provide adequate sight distance for overtaking can be obtained as follows:

Assume $h_3 = 0.75$ m and $\beta = 1°$. Further, assume $S \geq L$, then using Eq. (3.22)

$$L = 2 \times 470 - \frac{2(0.75 + 470 \tan 1°)}{0.03}$$

$$= 343 \text{ m}$$

In this case because S (= 470 m) is greater than L, the above value of 343 m is acceptable and unlike Example 3.2, no further calculation is necessary. Further, the minimum value of the length for a design speed of 80 kmph as suggested by IRC Special Publication 23 [255] is 50 m. Hence, the final design value remains as 343 m.

The location of the VPC is calculated as follows:

(i) The horizontal distance of VPC from the benchmark is given as

$$\text{horz. dist. of VPC} = \text{horz. dist. of VPI} - \frac{L}{2}$$

$$= 500 - \frac{343}{2} = 328.5 \text{ m}$$

(ii) The elevation of VPC is given as

$$\text{elev. of VPC} = \text{elev. of VPI} - \frac{L}{2} g_1$$

$$= 50 - \frac{343}{2} 0.005 = 49.14 \text{ m}$$

Hence, the coordinates of VPC are (328.5, 49.14).

The location of the VPT is given as follows:

(i) The horizontal distance of VPT from the benchmark is given as

$$\text{horz. dist. of VPT} = \text{horz. dist. of VPI} + \frac{L}{2}$$

$$= 500 + \frac{343}{2} = 671.5 \text{ m}$$

(ii) The elevation of VPT is given as

$$\text{elev. of VPT} = \text{elev. of VPI} + \frac{L}{2} g_2$$

$$= 50 + \frac{343}{2} 0.035 = 56 \text{ m}$$

Hence, the coordinates of VPT are (671.5, 56).

With respect to the VPC as the origin, the highest point of the curve will be at an x-coordinate value of

$$\frac{g_1 L}{g_1 - g_2} = \frac{0.005 \times 343}{0.005 - 0.035} = -57.16 \text{ m}$$

This indicates that the lowest point of the parabola representing the curve occurs before VPC. That is, the lowest point of the parabola is not inside the range of x-values, 0 to 343 m, with respect to VPC as the origin, over which the actual curve is provided. Further, since the entire curve actually provided is in the range of x-coordinate values greater than the x-coordinate value at the minimum, we can say that the actual curve is monotonically rising. Hence, the lowest point on the actual curve is the VPC and the highest point on the actual curve is the VPT.

3.5 CHANNELIZATION DESIGN

This refers to the use of geometric features in the control of traffic flow. For instance, a roundabout or rotary which bars vehicles from crossing an intersection directly is an example of channelization design. 'Turning bays' at intersections is another example of channelization design. In short, the purpose of channelization design is to promote safety and efficiency of traffic flow through the use of geometric features such as islands and lane markings.

Channelization design is predominantly an art and no set rules exist which suggest what kind of channelization is required at a particular location. However, a detailed description of some of the common types of channelization designs is provided in Chapter 5 on "Design of Traffic Facilities". Hence its description here is omitted.

EXERCISES

1. Draw the longitudinal profile (assuming a two-lane road of width 7 m) for the horizontal curve designed in Example 3.1. Use runoff values from the relevant code or use the limiting value as per the discussion in the text.
2. Determine the minimum radius of a circular horizontal curve for a two-lane road (3.5 m lane width) for design speeds of 30 kmph and 65 kmph.
3. Determine the minimum length of transition curves for the horizontal curves designed in Exercise 2.
4. Recalculate the quantities of Exercises 2 and 3 assuming maximum superelevation to be f_s (coefficient of side friction). Compare and comment on the results.
5. Refer to IRC:38–1988 [85] and find out other methods of attaining superelevation on curves. For these methods develop procedures to obtain the longitudinal profiles.

6. Refer to IRC:38–1988 [85] and write a critical review on the curve widening requirements and methods suggested therein.
7. By how much does the length of a circular curve reduce because of the use of transition curves?
8. For a given set of roads in one case, only a circular curve of radius R is provided; in another case a circular curve of radius R and transition curves of length L_T are provided. What is the distance between the centres of the two circular curves? Assume the external angle to be θ. You may also make any other necessary assumptions.
9. Identify the sag and the crest curves from the following curves:
   (a) Curve joining a + 2% slope to a – 2% slope
   (b) Curve joining a – 2% slope to a + 2% slope
   (c) Curve joining a + 2% slope to a 0% slope
   (d) Curve joining a + 2% slope to a + 3% slope
   (e) Curve joining a – 1% slope to a – 2% slope
   (f) Curve joining a – 3% slope to a – 2% slope
10. A vertical curve of length 183 m connects a + 4% grade to a – 2% grade. The VPI is at an elevation of 198.1 m and at a distance of 609.6 m with respect to Point A. An overbridge, 3.66 m wide, crosses the road. The location of the centre of the bridge (as it crosses over the road) is 567 m from Point A and at an elevation of 199.3 m with respect to Point A. Determine the smallest clearance between the overbridge and the road.
11. State whether the following statements are true or false. Give reasons for your answer.
   (a) When transition curves are provided, then two-thirds of the tangent runout is on the transition curve and one-third on the circular curve.
   (b) The radius of a transition curve is equal to one-half the radius of the circular curve to which it connects.
   (c) The vertical point of curvature is always above the vertical point of tangency.
   (d) Sight distance available on crest vertical curves is determined by night-time driving conditions.
   (e) Normal camber is provided on superelevated curves.
   (f) The longitudinal profile of a superelevated section of a curve is obtained through orthographic projection.
   (g) When transition curves are provided, the entire runoff is provided on them.

Traffic Flow

4.1 INTRODUCTION

This chapter is concerned with the study of traffic flow on and through various traffic facilities such as freeways (or expressways), signalized intersections, and unsignalized intersections. Such a study is imperative for better designing of traffic facilities. The chapter is divided into the following three sections. Section 4.2 presents the general fundamentals of traffic flow. Section 4.3 studies traffic flow behaviour when the flow of traffic is uninterrupted while the Section 4.4 studies traffic flow behaviour in situations where the flow is interrupted either due to slow moving vehicles or due to vehicle stoppages at intersections.

4.2 FUNDAMENTALS OF TRAFFIC FLOW

In this section, the basics of traffic flow characterization and analysis are described. The section is divided into two subsections. The first of these describes how traffic flow is characterized and the second describes the fundamental relation between the parameters characterizing traffic flow.

4.2.1 Flow Characterization

The condition of any traffic stream can be defined by two stream variables, namely *speed* and *density* . The *speed* of the traffic stream is defined as the average speed of the vehicles moving in that stream, and the *density* of the traffic stream is defined as the average number of vehicles per unit length of the stream. In the following, these parameters are described more precisely. In order to aid in the description of these parameters, a typical time–distance diagram of a traffic stream is shown in Figure 4.1. In the figure, each thin line represents the trajectory of a vehicle over time.

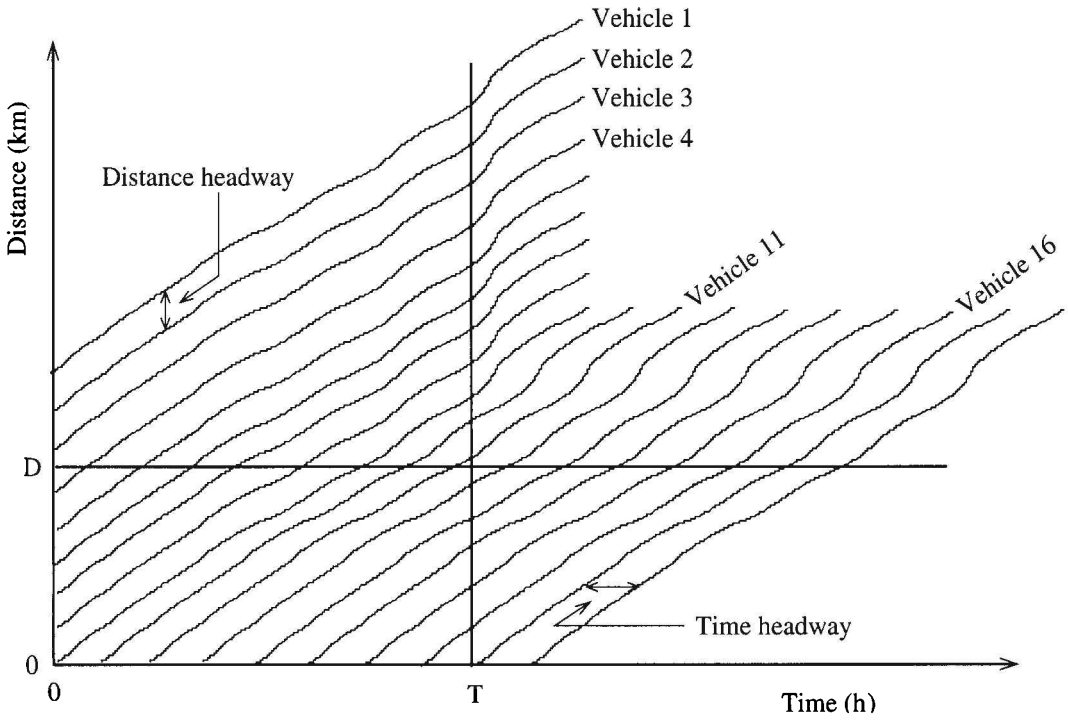


Figure 4.1 A typical time–distance plot of a traffic stream.

Speed, u

The speed of a traffic stream is, as described earlier, defined as the average speed of all the vehicles in the stream. The speed u_i of a vehicle i in a stream of N vehicles can be averaged in one of two ways as illustrated below:

$$\bar{u}_{\text{TMS}} = \frac{\Sigma_{\forall i} u_i}{N} \tag{4.1}$$

or

$$\bar{u}_{\text{SMS}} = \frac{1}{[\Sigma_{\forall i}(1/u_i)]/N} \tag{4.2}$$

Equation (4.1) gives the arithmetic mean of the speeds and is referred to as the *time mean speed* (TMS) of the traffic stream. This, in essence, means taking the speed of all the vehicles, say, crossing a point at distance D (see Figure 4.1), over a certain period of time, say T, and then taking the arithmetic mean of these. In this case, it will be the average of speeds of vehicles 4 through 11 at Point D.

Equation (4.2) is the harmonic mean of the speeds and is referred to as the *space mean speed* (SMS) of the stream. Note that in this case, the individual speeds are first translated into individual travel times (by taking the reciprocal) whose average is

determined and inverted to give the space mean speed. Physically, this means that first the average speed of each vehicle over a certain distance is determined (note that travel time is a direct measure of such an average speed) and the mean of that is obtained as the *space mean speed*. This measure of the average speed is more appropriate for the description of stream conditions as it gives a measure of the speed of the traffic stream over space.

Density, k

Density is defined as the average number of vehicles per unit distance. This is simply determined by dividing the total number of vehicles in a stream by the length of the stream. The unit of density is thus vehicles per km. With reference to Figure 4.1, for example, the density of the traffic stream at time T over a distance D will be the number of vehicles that cross the vertical line at time T, below the horizontal line at distance D, divided by the distance. In this particular case, density = $5/D$ vehicles per km.

Another related definition which should be mentioned here is that of *distance headway*. Distance headway is defined as the distance between corresponding points of two successive vehicles at any given time. Hence, the vertical gap between any two consecutive lines in Figure 4.1 gives the distance headway between the vehicles represented by the lines. Obviously, the reciprocal of density (at any time) gives the average distance headway between vehicles at that time.

Flow or volume, q

Although speed and density completely describe the stream conditions, another variable which is often used to describe a traffic stream is the *flow* or *volume*. Flow or volume of a traffic stream is defined as the number of vehicles of the stream that cross a fixed point on the road over a unit period of time. Generally, the period of time is taken as one hour and the unit of volume is stated as vehicles per hour. With reference to Figure 4.1, for example, the flow across point D over a time period of T is equal to the number of lines that cross the horizontal line through D to the left of the vertical line through point T divided by the time T. In this particular case, flow = $8/T$ vehicles per hour. As will be shown later, volume of a traffic stream can be easily obtained from the speed and density of the traffic stream.

Another related definition which should be mentioned here is that of *time headway* or simply *headway*. Time headway is defined as the time difference between any two successive vehicles when they cross a given point. Hence, the horizontal gap between any two consecutive lines in Figure 4.1 gives the time headway between the vehicles represented by the lines. Obviously, the reciprocal of flow (across a point) gives the average time headway between vehicles at that point.

Before leaving this section, a description of temporal variation of flow is provided. Although, speed, density, and flow all vary with time and space, only the temporal variation of flow is described here since it plays an important role in designing and

analyzing roadway sections. Flow varies with time—it varies from month to month, from day to day, from hour to hour and within the hour. In the design of expressways and other traffic facilities (discussed in Chapter 5) and level-of-service analysis (discussed later in this chapter), the variations which are of importance are (i) hour-to-hour variations and (ii) variations within the hour. Level-of-service analysis (as is design) is often done for the peak flow rate within the peak hour. The concepts of peak hour volume and within the hour peak flow rate are described in the following.

Hourly flow on an expressway (or any other traffic facility for that matter) varies over the year. Typically, such plots look like the one shown in Figure 4.2(a). In this figure, the abscissa is the rank of an hour; for example, the point marked 30th implies that in a year there are only 29 hours which have a higher flow than the flow in that hour. The ordinate gives the flow. Generally, the 30th highest hour is used as the peak hour volume.

Within each hour, the flow varies from minute to minute. Typically, such plots look like the one shown in Figure 4.2(b). If one analyzes (or designs) on the basis of peak hour volume (which when expressed as per minute flow is equal to the sum of all the minute volumes divided by 60), then there would be certain durations within the hour when the flow is actually larger than the hourly flow. Then, during these times the performance will be worse than what is envisaged (and may sometimes cause interruptions which take a long time to normalize). On the other hand, if one analyzes (or designs) on the basis of the highest minute flow rate, then the analysis (or design) will be highly pessimistic as for the rest of the hour the situation will be much better than envisaged. Hence, analysis is done for a flow value which indicates the highest flow for some period of time within the hour. Generally, the highest 15 minute period is taken.

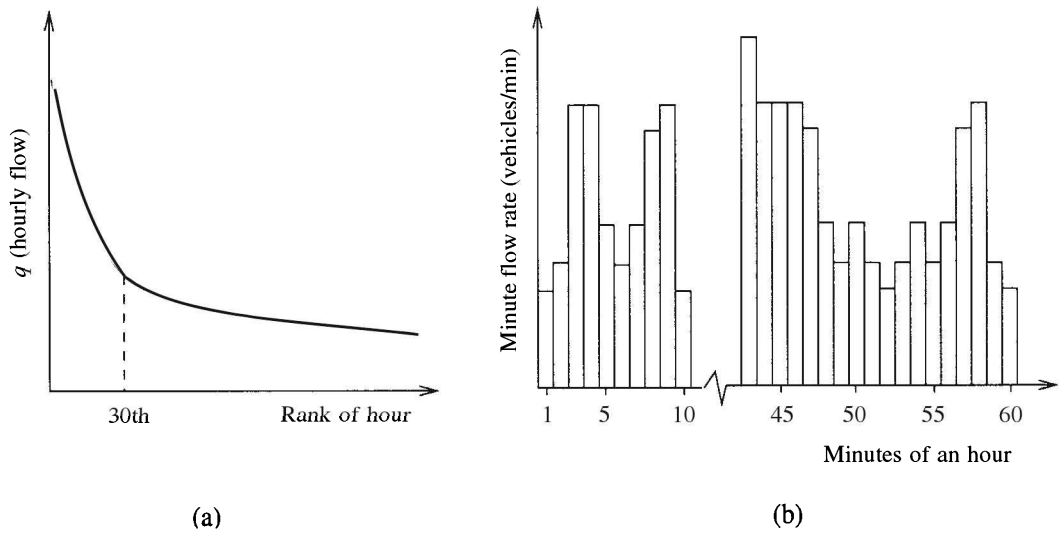


Figure 4.2 (a) Typical hourly variation of flow at a given section; (b) typical within the hour variation of flow at a given section.

The process of incorporating the highest 15 (or t) minute period within the hour is based on computing a factor called the *peak hour factor*, PHF_t. The PHF_t is computed as shown in Eq. (4.3). The hourly volume q is converted to within the hour peak flow by dividing it by PHF_t. Finally, the concept of peak hour factor (or PHF) is illustrated by Example 4.1.

$$PHF_t = \frac{\Sigma_{i=1}^{i=60} N_i}{(60/t)\max_{\forall k \in [1,60-t+1]}(\Sigma_k^{k+t-1} N_k)}$$

$$= \frac{q}{(60/t)\max_{\forall k \in [1,60-t+1]}(\Sigma_k^{k+t-1} N_k)} \tag{4.3}$$

where

N_i is the flow in minute i

q is the hourly flow

t is the within the hour duration for which the highest flow is determined (generally 15 minutes).

EXAMPLE 4.1

The following minute flow values were observed for the peak hour on an expressway section. Determine the 5 minute and 15 minute PHF. Also, convert the volume into within the hour peak flow using the 15 minute PHF.

Minute i	1–7	8–10	11–20	21	22–32	33–45	46–47	48–55	56–60
N_i	20	30	25	26	4	10	15	30	10

Solution

The total hourly volume q on this section is $\Sigma_{i=1}^{i=60} N_i$, and is given by

$$q = 7 \times 20 + 3 \times 30 + 10 \times 25 + 1 \times 26 + 11 \times 4 + 13 \times 10 +$$
$$2 \times 15 + 8 \times 30 + 5 \times 10 = 1000 \text{ vph}$$

Next the total volume in every 5 minute duration is calculated as

From minute 1 to minute 5, the total volume is: $5 \times 20 = 100$ vehicles
From minute 2 to minute 6, the total volume is: $5 \times 20 = 100$ vehicles
From minute 3 to minute 7, the total volume is: $5 \times 20 = 100$ vehicles
From minute 4 to minute 8, the total volume is: $4 \times 20 + 30 = 110$ vehicles

From minute 5 to minute 9, the total volume is: $3 \times 20 + 2 \times 30 = 120$ vehicles
From minute 6 to minute 10, the total volume is: $2 \times 20 + 3 \times 30 = 130$ vehicles
From minute 7 to minute 11, the total volume is: $20 + 3 \times 30 + 25 = 135$ vehicles
From minute 8 to minute 12, the total volume is: $3 \times 30 + 2 \times 25 = 140$ vehicles
From minute 9 to minute 13, the total volume is: $2 \times 30 + 3 \times 25 = 135$ vehicles
and so on for all the other 5 minute durations ending with:
From minute 56 to minute 60, the total volume is $5 \times 10 = 50$ vehicles.

Comparing all the above, the maximum number of vehicles in any 5 minute period is obtained as 150 vehicles (it occurs in all the five minute periods between minute 48 to minute 55).

Hence, the 5 minute peak hour factor, PHF_5, is given as

$$PHF_5 = \frac{1000}{(60/5)\times 150} = 0.555$$

In order to calculate the 15 minute PHF, as earlier, the total number of vehicles in each of the 15 minute periods need to be determined. This can be done as follows:

From minute 1 to minute 15, the total volume is: $7 \times 20 + 3 \times 30 + 5 \times 25 = 355$ vehicles;
From minute 2 to minute 16, the total volume is: $6 \times 20 + 3 \times 30 + 6 \times 25 = 360$ vehicles
and so on for all the 15 minute periods ending with:
From minute 46 to minute 60, the total volume is: $2 \times 15 + 8 \times 30 + 5 \times 10 = 320$ vehicles.

Comparing all the above, the maximum number of vehicles in any 15 minute period is obtained as 386 vehicles (it occurs in the interval from minute 7 to minute 21). Hence, the 15 minute peak hour factor, PHF_{15}, is given as

$$PHF_{15} = \frac{1000}{(60/15)\times 386} = 0.648$$

The hourly flow can be changed to within the hour peak flow rate as 1000/0.648 = 1544 vph. Note that 1544 vph is nothing but the hourly flow rate that would be obtained if for the whole hour the vehicles arrived at the rate at which the vehicles arrived during the peak 15 minute period (i.e. at the rate of 386/15 = 25.73 vehicles per minute).

4.2.2 Fundamental Relation of Traffic Flow

As stated earlier, three parameters, namely speed u, density k, and flow q, describe a traffic stream. In this section, a relation among these parameters (valid under all flow conditions) is developed.

Consider the flow situation described in Figure 4.3. The figure presents a 'snapshot' of a road section u km long taken at time, $t = 0$ h. Each dot with an arrow represents a vehicle moving from left to right at a speed of u kmph. The density on this section is k vpkm. An observer standing at location A starts counting the vehicles that cross him/her from time = 0 h and stops at time, $t = 1$ h.

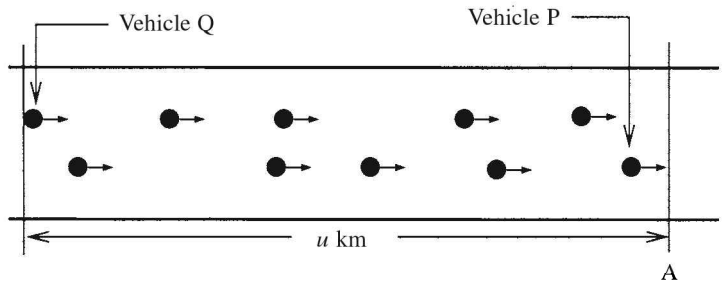


Figure 4.3 A snapshot of a road cross-section with density k vpkm and speed u kmph.

The first vehicle that the observer counts will be Vehicle P. Now, the section is u km long and Vehicle Q is at a distance of u km from the observer at time, $t = 0$ h. Further, since Vehicle Q is travelling at a speed of u kmph it will reach the observer exactly an hour later at time, $t = 1$ h. Hence, Vehicle Q will be the last vehicle that will be counted by the observer. Since, all the vehicles are travelling at the same speed, and since the observer counts Vehicle P as the first vehicle and Vehicle Q as the last vehicle, the observer would have counted, in one hour, all the vehicles in the section shown in the figure. Let this number be N. Now, the density on the section is k vpkm and the section is u km long; hence, $N = u \times k$. Thus, the observer would count $u \times k$ vehicles in an hour. Yet, by definition the number of vehicles the observer counts in an hour is the flow q vph. This implies that

$$q = u \times k \tag{4.4}$$

Equation (4.4) is the fundamental relation of traffic flow. It may be noted that the development given here is not a rigorous proof of the relation but gives the reader an intuitive basis for the relation.

4.3 UNINTERRUPTED TRAFFIC FLOW

In this section, the basis of describing and analyzing flow in an uninterrupted stream of traffic is discussed. First, the characteristics of uninterrupted traffic flow are discussed. Next, some methods of collecting data on the parameters describing uninterrupted traffic flow are provided. This is followed by two sections which describe the macroscopic and microscopic models of uninterrupted traffic flow.

4.3.1 Stream Characteristics

In uninterrupted traffic flow conditions, the traffic stream can move under the same stream conditions, that is, with the same speed and density, over an extended period of time. Of course, any uninterrupted stream sooner or later gets interrupted either by another stream moving under different conditions, or by some traffic control mechanism such as signals, stop signs, yield signs, and the like.

4.3.2 Data Collection

In this section, some methods of collecting data on speed, density, and flow are described. The section is accordingly divided into three parts, each devoted to the methods of data collection on one of the parameters.

Collecting speed data

The speed of vehicles can be measured either by using the principle of Doppler effect or by measuring the time it takes a vehicle to cross two closely-spaced sensors. The instruments which use the Doppler effect are of two types—the radar-based instrument and the laser-based instrument. These instruments are popularly referred to as *speed guns*. The speed measured using such equipment is the instantaneous speed of a vehicle and is often referred to as *spot speed* of the vehicle. A photograph of a typical speed gun is shown in Figure 4.4.

Figure 4.4 A typical speed gun which measures the spot speed of vehicles.

In the other method, two detectors which can detect the presence of a vehicle are kept at a fixed distance apart. The detectors record the times at which the axles of a vehicle cross the detectors. From the difference in times and the distance between the detectors the speed of the vehicle can be easily determined. These detectors could even be simple video cameras. Figure 4.5 shows the photograph of a typical tubular presence type detector laid on the road.

Collecting density data

Collecting data on density is a much more tricky issue. We can use aerial photography, or input-output study to determine the number of vehicles in a certain section, and divide that by the length of the section to compute the average density. Aerial photography is too costly and cumbersome, particularly over an extended period of time. Input-output studies, on the other hand, make an initial count of the vehicles in a section, and then

Figure 4.5 A typical tubular presence type detector.

count only the number of vehicles that enter and leave the section. In this manner, we can get the total number of vehicles that are present in a section at any given time. This method, although cost effective, gives only a gross measure of the density (especially, if the section is long) and is prone to errors as a miscount of the number of vehicles entering or exiting the section at any given time renders all subsequent values of density erroneous. Both the aerial photography and the input-output studies are no longer in common use.

The method which is most common today, uses the presence-type detectors to determine the duration for which the detectors are occupied by vehicles. Based on this occupancy time, the density can be easily calculated. In the following, the process of obtaining density data from occupancy data is described.

Consider a presence-type detector of width D_l (this could also be the distance between two closely-spaced tubular detectors); assume that a vehicle of average length V_l on an average spends a time t_o on the detector. (Note that the time a vehicle spends on the detector can be obtained by subtracting the time at which the front axle crosses into the detector from the time the rear axle crosses over the detector; see Figure 4.6 for an explanation of t_o.)

As depicted in Figure 4.6, the average speed can be obtained as

$$u = \frac{V_l + D_l}{t_o}$$

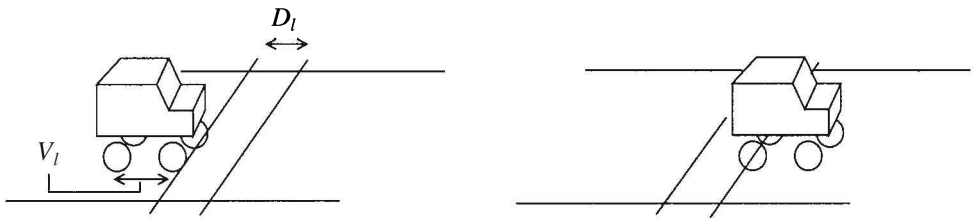


Figure 4.6 Figure explaining occupancy time t_o.

Note that the average speed here is the space mean speed as the average is based on the average travel time (the occupancy time is the time a vehicle takes to travel a distance of $V_l + D_l$). We can therefore write

$$t_o = \frac{V_l + D_l}{u}$$

If a total of N vehicles pass over the detector in a time period T and if the sum of the occupancy times of the vehicles is T_o, then

$$T_o = Nt_o$$

or

$$T_o = N\frac{V_l + D_l}{u}$$

Hence, the fraction of time the detector is occupied, $Oc = T_o/T$, is given by

$$Oc = N\frac{V_l + D_l}{uT}$$

Now by definition, $q = N/T$. Therefore, we get

$$Oc = q\frac{V_l + D_l}{u}$$

With density, $k = q/u$, we obtain

$$Oc = k(V_l + D_l)$$

or

$$k = \frac{Oc}{V_l + D_l} \tag{4.5}$$

Equation (4.5) can now be used to determine density from occupancy data, which is not very difficult to obtain. Further, we can easily determine the length of each vehicle (and use their average for V_l) by using another similar detector at a short distance away. It is left to the reader to determine how the length of a vehicle can be obtained by using two detectors and what basic assumption needs to be made in that case.

Collecting flow or volume data

Flow data is the easiest to collect as it entails counting the number of vehicles which cross a point on the road in a given time. We can use people, video cameras or presence-type detectors to count the number of vehicles. Given the axle distributions of different types of vehicles, most presence-type detectors can accurately classify vehicles and therefore, the volume of different types of vehicles can be easily obtained.

In this section, however, a low-cost method often used for determining volume and density of traffic streams is described. This method is referred to as the *moving observer method*. In this method, an observer in a test vehicle starts from a point A and moves against the stream being measured at a speed of v_a. After a distance L and time T_w, at point B the test vehicle swings around and starts moving with the stream at a speed v_w till it reaches point A (say, after a time T_a). During both the passes, the test vehicle counts the number of vehicles that it overtakes and the number of vehicles that overtake it. Based on this information and certain calculations, as illustrated here, stream variables q and k can be estimated.

Let M_o^w, M_p^w, M_o^a, and M_p^a, respectively, be the number of vehicles that overtake the test vehicle when the test vehicle is moving with the stream, the number of vehicles that the test vehicle overtakes when moving with the stream, the number of vehicles that overtake the test vehicle while the test vehicle is moving against the stream, and the number of vehicles that the test vehicle overtakes while moving against the stream.

Now, consider the following thought experiment. Let there be a stream of vehicles wherein some are moving at a speed u_1 and density k_1 and others are moving at a speed u_2 and density k_2. Let the test vehicle's speed v_w be greater than u_1 and less than u_2. In this case,

$$M_o^w = k_2(u_2 - v_w)T_w$$

$$M_p^w = k_1(v_w - u_1)T_w$$

or

$$M_w = M_o^w - M_p^w = k_2(u_2 - v_w)T_w - k_1(v_w - u_1)T_w$$

$$= (q_2 + q_1)T_w - (k_2 + k_1)v_wT_w$$

Since both the sub-streams are within the same stream (and the above construct is purely abstract), $q_2 + q_1 = q$ and $k_2 + k_1 = k$. Therefore,

$$M_w = (q - kv_w)T_w \tag{4.6}$$

Similarly, by noting that when the test vehicle moves against the stream (i) all the vehicles in the stream in effect overtake the test vehicle (overtaking being defined as reaching a point downstream before the test vehicle) and the test vehicle overtakes none, and (ii) the relative speed between the test vehicle and the stream is the sum of their speeds,

$$M_a = M_o^a - M_p^a = (qT_a + kv_aT_a) - 0$$

$$= (q + kv_a)T_a \tag{4.7}$$

Adding the expressions for M_w and M_a and using the relation $L = v_w T_w = v_a T_a$, the following is obtained

$$q = \frac{M_w + M_a}{T_w + T_a} \tag{4.8}$$

The value of q obtained in Eq. (4.8) can be substituted either in Eq. (4.6) or in Eq. (4.7) to obtain the value of k. Further, it can be shown that the value of q obtained from Eq. (4.8) will be the same as the flow observed at one of the end points of the section (over which the test vehicle runs) between the time when the test vehicle starts moving and the time $T_w + T_a$. The proof of this is left as an exercise for the reader.

4.3.3 Macroscopic Traffic Flow Models

Studies on traffic flow behaviour have shown that the three parameters (speed, density, and flow) describing an uninterrupted traffic stream are pair-wise dependent. That is, there is a relation between speed and density, flow and density, and speed and flow. Figure 4.7 shows typical plots of (a) speed–density, (b) speed–flow, and (c) flow–density data.

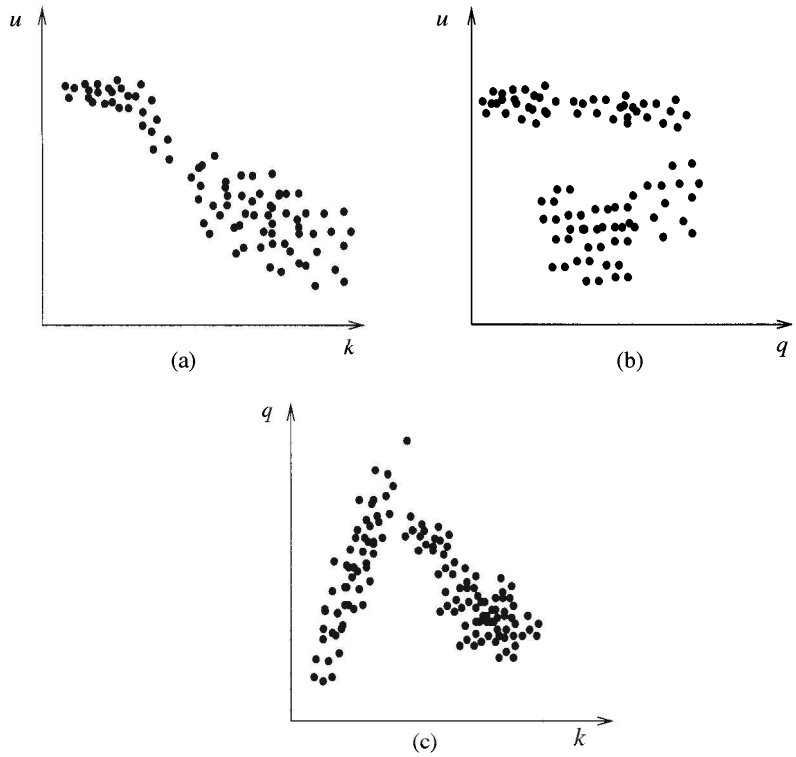


Figure 4.7 Typical plots of (a) speed–density, (b) speed–flow, and (c) flow–density for an uninterrupted traffic stream.

It is felt that, and the reasons for this will be stated shortly, of these three pair-wise relations, the basic dependence is between speed and density; the other relations are implied relations in the sense that once speed and density are related, all other relations become automatic because of the fundamental relation of traffic flow described in Eq. (4.4).

The reason for saying that speed and density dependence is the most basic of the three relations mentioned above is that drivers drive vehicles based on their immediate surroundings. If the immediate surroundings of a driver are cramped and congested then as density increases, drivers for safety reasons reduce their speed and vice versa. Thus, it can be stated that since a relation between speed and density is an outcome of the behavioural aspects of a driver it must be the basic relation. Even otherwise, it is difficult to imagine that relations among flow and speed, and flow and density are the basic relations because the driver (whose behaviour in a traffic stream ultimately gives rise to these relations) at no point has any indication as to what the flow might be.

Not surprisingly then, over the years, various models of speed–density relation have been proposed. Here, some of these models are mentioned. Before going into the models, however, it should be pointed out that (i) at very low densities the traffic moves at some finite speed referred to as the *free speed* u_f, and (ii) the speed is zero or very close to zero when the density of the stream reaches an upper limit known as the *jam density* k_j. Further, it must be realized that the relation between speed and density should be monotonically decreasing (given the fact that drivers reduce speed when they feel constrained).

Among the many models of speed and density (henceforth referred to as u–k relation or u–k model), the following are discussed here as they have historical importance and pedagogic value.

1. Linear model
2. Logarithmic model
3. Exponential models
4. Generalized polynomial model
5. Multi-regime models

Linear model

In the 1930s, Greenshields [82], based on limited data, proposed the following linear form of the u–k relation

$$u = a + bk \tag{4.9}$$

Using the boundary conditions that $u = u_f$ when $k = 0$ and $u = 0$ when $k = k_j$, the model in Eq. (4.9) reduces to the following form, often referred to as the Greenshields' model.

$$u = u_f\left(1 - \frac{k}{k_j}\right) \tag{4.10}$$

Logarithmic model

As more data sets became available, it was clear that a linear model was not a good model of the u–k relation. Based on observations and the theory of one-dimensional compressible fluid flow, Greenberg [81] in 1959 proposed the following relation:

$$u = a \ln\left(\frac{b}{k}\right) \tag{4.11}$$

On using the boundary condition that $u = 0$ when $k = k_j$ and the relation $q = u_k$, we can easily show that $a = u_o$ and $b = k_j$, where u_o is the speed at which q (as implied by the equation) becomes maximum. However, a drawback of this model is that as $k \to 0$, $u \to \infty$. This implies that the model's capability of predicting speeds at low densities is not very good.

Exponential models

Owing to the problems of the logarithmic model in the free flow region (where density is low), initially Underwood [248] in the early 1960s and later researchers at Northwestern University proposed an exponential form for the u–k relation

$$u = ae^{-b^{-1}(k/c)^b} \tag{4.12}$$

In the version of the model proposed by Underwood, $b = 1$, and that proposed by Northwestern University, $b = 2$. In either case, on using the boundary condition that $u = u_f$ when $k = 0$ and the relation $q = uk$, we can easily show that $a = u_f$ and $c = k_o$, where k_o is the density at which q (as implied by the equation) becomes maximum. However, a drawback of this model is that u becomes zero only when $k \to \infty$; this implies that according to this model the jam density is infinite.

Generalized polynomial model

This model was developed from the behavioural models of traffic flow (to be discussed under microscopic models later in the chapter) as a result of the work done by Gazis et al. [66, 67] and by May and Keller [153].

$$u^{1-a} = b^{1-a}\left\{1 - \left(\frac{k}{c}\right)\right\}^{d-1} \tag{4.13}$$

It can be shown that this model provides finite values of free flow speed and jam density when $0 \leq a < 1$ and $d > 1$. Further, it can be shown (by using the usual boundary conditions) that $b = u_f$ and $c = k_j$. In general, while calibrating the model, the values of a and d are chosen from a general idea of parameters like free-flow speed, jam density, etc. (See May [153] for a detailed description of the process.)

Multi-regime models

The above models of speed–density relation are referred to as single-regime models as they assume that the same relation between speed and density is valid for the entire range of densities seen in traffic streams. Some researchers question this assumption stating that human beings as drivers behave differently when the density is low than when the density is high. That is, according to this school of thought the speed–density relation is different in different regimes (zones) of density.

Many researchers proposed that the speed–density relation is different in the *free-flow regime* (where density is low) and in the *forced-flow* (or *congested*) *regime* (where density is high). Such two-regime models were proposed by Edie [59] and many others. Edie, for example, used an exponential model in the free-flow regime and a logarithmic model in the forced-flow regime.

Others have proposed three-regime models where it was assumed that in between the free-flow regime and the congested regime there is a transition zone where the flow can neither be characterized as free-flow nor be characterized as forced-flow.

Macroscopic speed–flow and flow–density relations

As stated earlier, once a speed–density relation is specified or assumed, the speed–flow and flow–density relations are automatically implied. For example, if we assume that the speed–density relation is given by the Greenshields' linear model then the implied flow–density and speed–flow models can be obtained as follows:

$$u = u_f\left(1 - \frac{k}{k_j}\right)$$

and

$$q = uk = u_f\left(1 - \frac{k}{k_j}\right)k$$

That is, the implied flow–density relation is a parabolic relation given as

$$q = u_f k - \frac{u_f}{k_j}k^2 \tag{4.14}$$

Similarly,

$$u = u_f\left(1 - \frac{k}{k_j}\right)$$

implies

$$k = k_j - \frac{k_j}{u_f}u$$

Hence

$$q = uk = u\left(k_j - \frac{k_j}{u_f}u\right)$$

That is, the implied speed–flow relation is a parabolic relation given as

$$q = uk_j - \frac{k_j}{u_f}u^2 \tag{4.15}$$

Using similar algebra, we could easily determine the implied relations once one of the three (i.e. either u–k, u–q, or q–k) relations is either assumed or known.

Calibrating macroscopic models

Determining the parameters of a macroscopic model (say, the u–k model) from data on the relevant variables (in this case, speed and density) is the subject of discussion here. In general, linear regression analysis is used to determine the parameters of the macroscopic model. It must be understood here that as long as the parameters of the model are linear, the linear regression analysis can be used.

In the simplest form of regression analysis (known as OLS or ordinary least squares estimation), the parameters are chosen such that the sum of the squares of the differences between the estimated values and the observed values is minimum. The theory of multivariable linear regression analysis is not discussed here; the interested reader may refer to Gujarati [95] for a good understanding of this topic. Here, some expressions are given which will help the reader to determine the two unknown parameters of the macroscopic models. The reason for choosing to provide expressions for determining only two parameters is that most macroscopic models have only two unknown parameters; even in the ones that have more than two parameters, most of the parameters are chosen from other considerations and only two are determined from the data.

Linear regression analysis (using OLS) suggests that for an equation of the type $g(y) = A + Bf(x)$, the parameters A and B can be determined as follows:

$$B = \frac{\Sigma_{\forall i}[f(x_i) - \overline{f(x_i)}][g(y_i) - \overline{g(y_i)}]}{\Sigma_{\forall i}[f(x_i) - \overline{f(x_i)}]^2} \tag{4.16}$$

$$A = \overline{g(y_i)} - B\overline{f(x_i)} \tag{4.17}$$

where $\overline{g(y_i)}$ and $\overline{f(x_i)}$ indicate the mean values of $g(y_i)$ and $f(x_i)$, respectively. The following two examples illustrate how the u–k models can be calibrated.

EXAMPLE 4.2

For the following data on speed and density, determine the parameters of the Greenshields' model.

u (km/h)	75	75	65	55	40	50	30	35	20	30	10	10
k (v/km)	10	20	30	40	50	60	70	80	90	100	100	110

Solution

Greenshields' equation is $u = a + bk$, where a is actually equal to u_f and $b = -(u_f/k_j)$. [see Eqs. (4.9) and (4.10).]

Using Eqs. (4.16) and (4.17), $g(y) = u$, and $f(x) = k$, we can obtain $b = -0.654$ and $a = 82.65$. Hence, according to this estimation, $u_f = 82.65$ km/h and $k_j = 126.4$ v/km. However, note that none of the properties of the linear regression (or OLS) estimates (see Gujarati [95]) are valid for the estimate of k_j as the regression equation used is not linear in the parameter k_j. Nonetheless, these values of u_f and k_j do produce the minimum sum of squared errors for the data given and the model used.

Figure 4.8 shows the plot of the data points as well as the calibrated Greenshields' model for the above example.

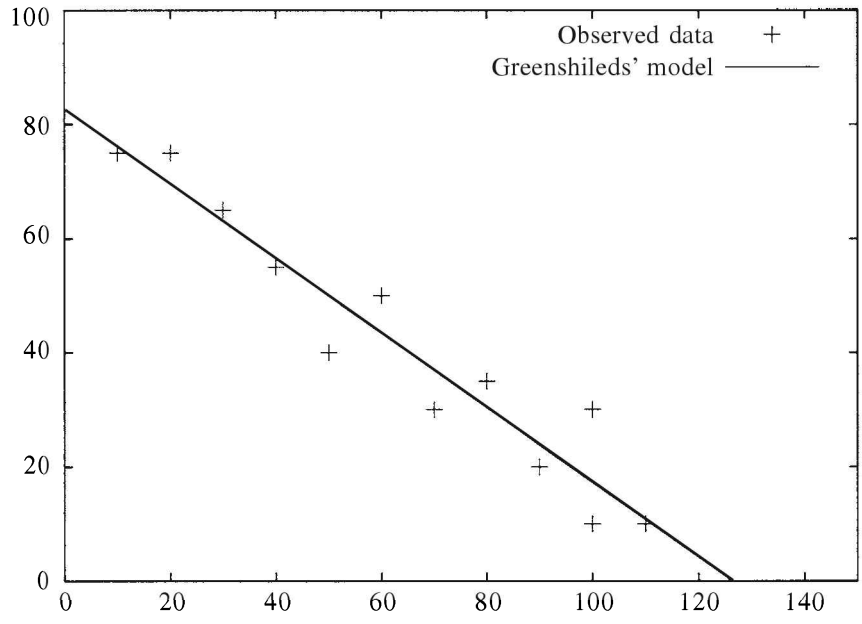


Figure 4.8 Calibrated Greenshields' model and the observed speed density data for Example 4.2.

EXAMPLE 4.3

For the data on speed and density given in Example 4.2, determine the parameters of the Northwestern University model.

Solution

As per Eq. (4.12), the Northwestern University model is

$$u = ae^{-0.5(k/c)^2}$$

This can be rewritten as

$$\ln(u) = \ln(a) - \frac{0.5}{c^2}k^2$$

or

$$\ln(u) = A + Bk^2$$

where $A = \ln(a)$ and $B = -0.5/c^2$.

Using Eqs. (4.16) and (4.17), $g(y) = \ln(u)$, and $f(x) = k^2$, we can obtain $B = -0.0001541$ and $A = 4.31$. Hence, according to this estimation, $a = 74.31$ and $c = 56.96$ which implies that $u_f = 74.31$ km/h and $k_o = 56.96$ v/km. However, note that none of the properties of the linear regression (or OLS) estimates (see Gujarati [95]) are valid for either of the estimates since the regression equation used is not linear in the parameters u_f and k_o. Nonetheless, these values of u_f and k_o do produce the minimum sum of squared errors for the data given and the model used.

Figure 4.9 shows the plot of the data points as well as the calibrated Northwestern University model for this example

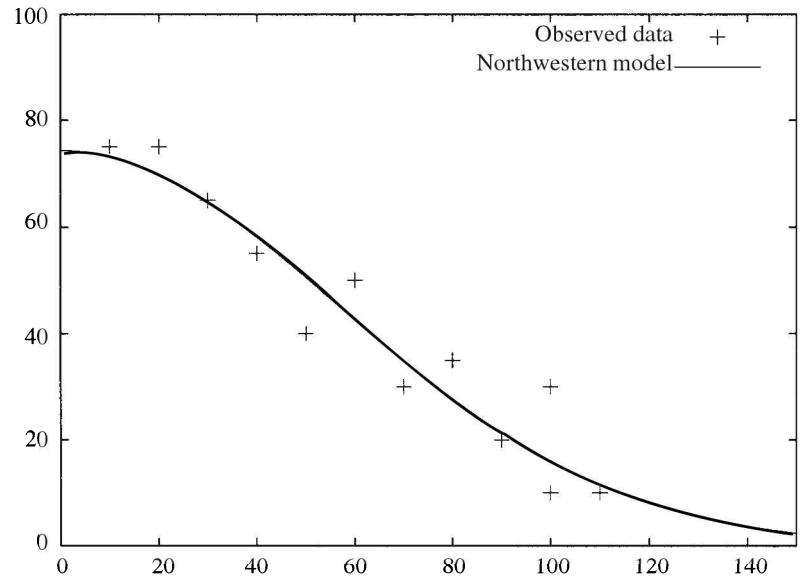


Figure 4.9 Calibrated Northwestern University model and the observed speed density data for Example 4.3.

4.3.4 Microscopic Traffic Flow Models

Microscopic models of traffic flow attempt to analyze the flow of traffic by modelling driver–driver, and driver–road interactions within a traffic stream. Modelling driver–driver interaction entails modelling how a driver reacts to the actions taken by another driver. On the other hand, modelling driver–road interaction entails modelling how a driver reacts to various features of the road such as narrow road width, curves, and so forth. Thus, it can be said that a microscopic model of traffic flow is basically a model of driver behaviour.

Over the years various models of driver behaviour in different driving situations have been developed. These can be broadly classified into two parts: (i) driver behaviour in the presence of static obstacles and (ii) driver behaviour in car-following situations (where one vehicle follows another). Very little and only exploratory work has been carried out on the first topic. The interested reader may refer to Michaels and Gozan [159] and Taragin [226] for work on this topic. However, a lot of work has been done on modelling driver behaviour in car-following situations and the entire body of work is sometimes collectively referred to as *theories of car-following*. The reason for this is that such models are directly associated with unidirectional traffic streams and find applications in many areas. In this section, two different car-following models are discussed. One is due to the research by the team at the General Motors Research Laboratory [36, 66, 67, 100, 101] and here is referred to as the GM model and the other is due to Chakroborty and Kikuchi [33, 135, 34] and here is referred to as Fuzzy Inference Model. A good and up-to-date review of car-following models can be found in Brackstone and McDonald [19]. However, before going into the models some of the important properties of the car-following behaviour are described.

Car-following behaviour

Car-following is a control process in which the driver of the following vehicle attempts to balance between maintaining a safe distance between his/her car and the vehicle ahead and maintaining a speed as close to his/her desired speed by accelerating or decelerating in response to the actions of the vehicle ahead. The general features of car-following behaviour are:

- Car-following behaviour is approximate in nature as the elements participating in the process are human.
- The driver of the following vehicle seems to react to certain stream variables such as distance headway between itself and the vehicle ahead and the rate of change of this distance headway. These variables (and possibly others) are referred to as *stimuli* to which the driver reacts.
- Response to stimuli in car-following behaviour is asymmetric in the sense that drivers respond differently (in terms of absolute value of acceleration or deceleration) to distance headway decrements than to distance headway increments.

This is possibly due to the fact that distance headway decrements pose safety hazards whereas distance headway increments do not.

- Car-following behaviour is stable. The behaviour is such that if the leading vehicle changes its speed for some time and then maintains a steady speed, the following vehicle eventually starts driving at the same speed and at a safe distance away. That is, although the distance headway and the relative speed between the vehicles initially vary with time, these values eventually stabilize at the safe distance headway and zero, respectively. This particular property of car-following is referred to as *local stability.*
- Car-following behaviour is such that any perturbations to distance headway and relative speed (or speed) introduced by the leading vehicle progressively reduce as they get transmitted upstream in a platoon of vehicles. This implies that, in a sufficiently long platoon, the perturbation introduced by the lead vehicle may not be felt at all by the vehicles near the end of the platoon. This property is referred to as *asymptotic stability.*
- Since car-following is a human control process, the stable condition reached is not absolutely stable in the mathematical sense of the word. There is some small and cyclic variation in the distance headway and speed around their respective stable positions. This is referred to as *drift.*
- Another feature which is seen is that the stable speed and the stable distance headway are not independent of one another—the higher the stable speed, the greater is the stable distance headway and vice versa. (Note that the same feature is seen, although at a macroscopic level, in the relation between speed and density).
- Finally, *closing-in* and *shying-away* patterns are observed in car-following behaviour. If the driver of the following vehicle finds that the distance headway is 'large', then he/she closes in on the leading vehicle irrespective of the actions of the leading vehicle. On the other hand, if the driver of the following vehicle finds that the distance headway is 'small', then he/she shies away from the leading vehicle irrespective of the actions of the leading vehicle.

GM model of car-following behaviour

The research team in General Motors, during the 1950s and 1960s [36, 66, 67, 100, 101], developed a difference differential equation based model of car-following. This, despite some of its obvious shortcomings, remains one of the important models of car-following. The model, referred to here as the GM model, proposes that drivers of the following vehicles react by accelerating (or decelerating) to only one stimulus, namely the rate of change of distance headway (which is relative speed). The extent of reaction, however, is inversely related to distance headway (implying that as distance headway increases the effect of the actions of the leading vehicle diminish) and directly related

to the speed of the following vehicle (implying that if the following vehicle is moving at a higher speed the driver is more sensitive to the actions of the leading vehicle). Schematically, the GM model views the car-following behaviour as shown in Figure 4.10.

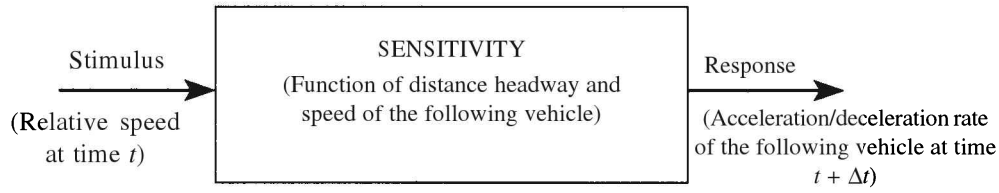


Figure 4.10 Schematic of the GM model view of the car-following behaviour.

Mathematically, the GM model can be expressed as

$$\ddot{x}_{n+1}(t+\Delta t) = \frac{\alpha_{\ell,m}(\dot{x}_{n+1}(t+\Delta t))^m}{[x_n(t)-x_{n+1}(t)]^\ell}[\dot{x}_n(t)-\dot{x}_{n+1}(t)] \tag{4.18}$$

where

$\ddot{x}_i(t)$ is the acceleration/deceleration of the ith vehicle at time t

$\dot{x}_i(t)$ is the speed of the ith vehicle at time t

$x_i(t)$ is the distance of the ith vehicle at time t from an arbitrary upstream benchmark point

Δt is the perception reaction time

$\alpha_{\ell,m}$ is a constant (whose units and meaning vary with the choice of ℓ and m).

It can be shown that with the proper choice of value of ℓ, m, and $\alpha_{\ell,m}$, the GM model does predict a behaviour which is locally and asymptotically stable. However, the GM model is a precise deterministic model of the human control process and does not incorporate the inherent vagueness of the human perception and reaction process. Owing to such deterministic representation, the GM model cannot represent *drift* in the stable condition. A much larger problem with the GM model is that it only considers relative speed (or rate of change of distance headway) as the stimulus, and hence it cannot model *closing-in* and *shying-away* behaviours. Another ramification of this single stimulus modelling is that the stable distance headway predicted by the GM model becomes sensitive to the initial conditions (although in reality the stable distance headway is only dependent on the final speed of the vehicles). This implies that, although the predictions from Eq. (4.18) may give consistent stable conditions (which it does) when the perturbations occur while the vehicles are moving under stable conditions, the predictions of the model are unable to properly describe how the stable condition is reached from an unstable initial condition. This, it is felt, is a major lacuna in the GM model. A good overview and critique of the GM model can be found in Chakroborty and Kikuchi [33].

However, notwithstanding these problems, the GM model formed a cornerstone of the theory of traffic sciences. One of the prime reasons for this is that the GM model under steady-state (or stable) conditions gives rise to various known models of speed–density relations. Figure 4.11 shows the m and ℓ values for which the GM model gives rise to the various models of u–k relation discussed in Section 4.3.3. Example 4.4 shows how a macroscopic u–k model can be derived from the GM model of car-following. Further, Example 4.5 is presented which shows how the GM model can be used to calculate the reactions of the following vehicle in response to the actions of the leading vehicle.

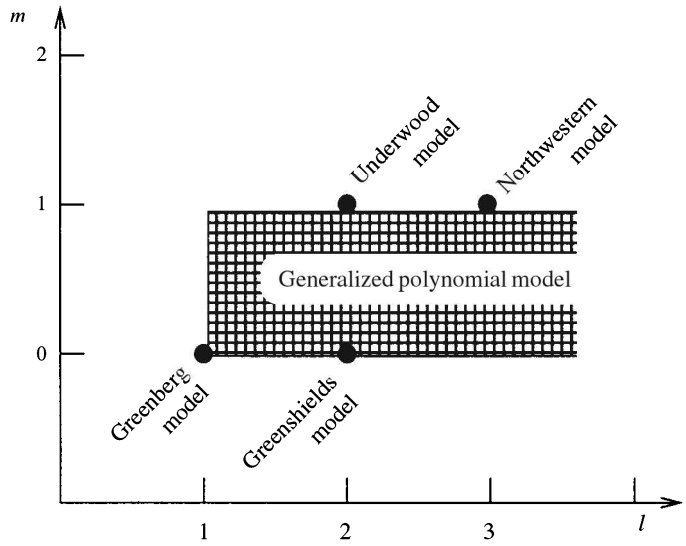


Figure 4.11 Schematic showing the mapping between the exponents m and ℓ of the GM model and the macroscopic u–k relations.

EXAMPLE 4.4

Show that with $m = 1$ and $\ell = 2$ the GM model of car-following in the steady state implies Underwood's macroscopic u–k relation.

Solution

For $m = 1$ and $\ell = 2$, the GM model takes the form:

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{\alpha_{2,1}\dot{x}_{n+1}(t + \Delta t)}{[x_n(t) - x_{n+1}(t)]^2}[\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$

Since the stream of traffic is flowing in steady state, there is no change in the system with time, and hence there is no distinction between time t and time $t + \Delta t$. Taking

$\dot{x}_{n+1}(t + \Delta t)$ as the denominator on the LHS and integrating both sides with respect to t yields

$$\ln(\dot{x}_{n+1}(t)) = -\alpha_{2,1}\frac{1}{x_n(t) - x_{n+1}(t)} + C$$

where C is the constant of integration.

Noting that (i) at steady state all vehicles are moving at the same speed (say, u) and at the same spacing (which implies that the reciprocal of $(x_n(t) - x_{n+1}(t))$ is equal to density, k) and (ii) the constant C can be written as ln (c), where c is another constant, we can write the above relation as

$$\ln(u/c) = -\alpha_{2,1}k$$

or

$$u = ce^{-\alpha_{2,1}k}$$

Using the boundary condition $u = u_f$ when $k = 0$, we get

$$u = u_f e^{-\alpha_{2,1}k}$$

From this, and (i) multiplying both sides by k, (ii) using the relation $q = uk$, (iii) differentiating both sides with respect to k, and (iv) determining the value of k_o at which q becomes maximum, we get $\alpha_{2,1} = 1/k_o$; thus the final form of the equation becomes

$$u = u_f e^{-(k/k_o)}$$

which is the Underwood's u–k relation.

EXAMPLE 4.5

Simulate the car-following behaviour for the following situation using a system update (or scan) time of 0.5 s: Two vehicles are moving at an initial speed of 16 m/s and distance headway of 28 m. At time = 1 s, the lead vehicle (LV) accelerates at 1 m/s^2 for 2 s; from time = 3 s, the LV decelerates at – 1 m/s^2 for 2 s.

Assume (i) perception–reaction time of 1 s, and (ii) GM model of car-following with $m = 0$, $\ell = 1$, and $\alpha_{1,0} = 13$ m/s. Based on the simulated data, plot (i) distance headway versus time, (ii) relative speed versus time, (iii) speed of LV and speed of FV versus time, and (iv) actions of LV and actions of FV versus time.

Solution

The car-following model to be used here is:

$$\ddot{x}_{FV}(t+1) = \frac{13}{[x_{LV}(t) - x_{FV}(t)]}[\dot{x}_{LV}(t) - \dot{x}_{FV}(t)]$$

or

$$\ddot{x}_{FV}(t) = \frac{13}{[x_{LV}(t-1) - x_{FV}(t-1)]}[\dot{x}_{LV}(t-1) - \dot{x}_{FV}(t-1)]$$

Further, since the system update time is 0.5 s, all the variables such as distances of FV and LV from an upstream point, speeds of FV and LV, actions of FV and LV, and so on should be calculated every 0.5 s. Table 4.1 shows all the system variable values calculated at 0.5 s interval. It is assumed that the upstream benchmark point is where the FV is at time t = 0 s. Further, all entries in the table indicate the value of the variable at the beginning of the interval, and acceleration/deceleration is assumed to remain constant over the 0.5 s update intervals. Also, the speed $\dot{x}(t)$ at time t, and the distance $x(t)$ at time t, are calculated using the following equations of motion:

$$\dot{x}(t) = \dot{x}(t-0.5) + \ddot{x}(t-0.5) \times 0.5$$

$$x(t) = x(t-0.5) + \dot{x}(t-0.5) \times 0.5 + 0.5 \times \ddot{x}(t-0.5) \times 0.5^2$$

Table 4.1 Table of values for Example 4.5 on car-following

Time	*LV*			*FV*			*Relative*	*Distance*
t	$\ddot{x}(t)$	$\dot{x}(t)$	$x(t)$	$\ddot{x}(t)$	$\dot{x}(t)$	$x(t)$	*speed*	*headway*
(s)	(m/s^2)	(m/s)	(m)	(m/s^2)	(m/s)	(m)	(m/s)	(m)
0	0	16	28	0	16	0	0	28
0.5	0	16	36	0	16	8	0	28
1	1	16	44	0	16	16	0	28
1.5	1	16.5	52.125	0	16	24	0.5	28.125
2	1	17	60.5	0	16	32	1	28.5
2.5	1	17.5	69.125	0.231	16	40	1.5	29.125
3	– 1	18	78	0.456	16.116	48.029	1.884	29.971
3.5	– 1	17.5	86.675	0.670	16.344	56.144	1.156	30.731
4	– 1	17	95.5	0.817	16.678	64.399	0.322	31.101
4.5	– 1	16.5	103.875	0.489	17.087	72.841	– 0.587	31.034
5	0	16	112	0.134	17.332	81.445	– 1.332	30.555
5.5	0	16	120	– 0.246	17.399	90.128	– 1.399	29.872
6	0	16	128	– 0.567	17.276	98.797	– 1.276	29.203
6.5	0	16	136	– 0.609	16.993	107.364	– 0.993	28.636
7	0	16	144	– 0.568	16.688	115.784	– 0.688	28.216
7.5	0	16	152	– 0.451	16.404	124.057	– 0.404	27.943
8	0	16	160	– 0.317	16.179	132.203	– 0.179	27.797
8.5	0	16	168	– 0.188	16.020	140.253	– 0.020	27.747
9	0	16	176	– 0.084	15.926	148.239	0.074	27.761
9.5	0	16	184	– 0.010	15.885	156.192	0.115	27.808
10	0	16	192	0.034	15.88	164.133	0.12	27.867

(Contd.)

Table 4.1 (Contd.)

Time	*LV*			*FV*			*Relative*	*Distance*
t	$\ddot{x}(t)$	$\dot{x}(t)$	$x(t)$	$\ddot{x}(t)$	$\dot{x}(t)$	$x(t)$	*speed*	*headway*
(s)	(m/s^2)	(m/s)	(m)	(m/s^2)	(m/s)	(m)	(m/s)	(m)
10.5	0	16	200	0.054	15.897	172.077	0.103	27.923
11	0	16	208	0.056	15.924	180.033	0.076	27.967
11.5	0	16	216	0.048	15.952	188.002	0.048	27.998
12	0	16	224	0.035	15.976	195.984	0.024	28.016
12.5	0	16	232	0.022	15.994	203.976	0.006	28.024
13	0	16	240	0.011	16.005	211.976	– 0.005	28.024
13.5	0	16	248	0.003	16.010	219.974	– 0.010	28.021
14	0	16	256	– 0.002	16.012	227.985	– 0.012	28.015
14.5	0	16	264	– 0.005	16.011	235.991	– 0.011	28.009
15	0	16	272	– 0.005	16.008	243.995	– 0.008	28.005
15.5	0	16	280	– 0.005	16.006	251.999	– 0.006	28.001
16	0	16	288	– 0.004	16.003	260.001	– 0.003	27.999
16.5	0	16	296	– 0.003	16.001	268.002	– 0.001	27.998
17	0	16	304	– 0.001	16	276.002	0.00	27.998
17.5	0	16	312	– 0.001	15.999	284.002	0.001	27.998
18	0	16	320	0.00	15.999	292.002	0.001	27.998
18.5	0	16	328	0.00	15.999	300.001	0.001	27.999
19	0	16	336	0.001	15.999	308.001	0.001	27.999
19.5	0	16	344	0.001	15.999	316.000	0.001	28
20	0	16	352	0.00	16	324	0	28
20.5	0	16	360	0.00	16	332	0	28
21	0	16	368	0.00	16	340	0	28
21.5	0	16	376	0.00	16	348	0	28
22	0	16	384	0.00	16	356	0	28

Based on the data in Table 4.1, the following graphs are plotted: (i) distance headway versus time (see Figure 4.12), (ii) relative speed versus time (see Figure 4.13), (iii) speed of LV and speed of FV versus time (see Figure 4.14), and (iv) actions of LV and actions of FV versus time (see Figure 4.15). It can be seen from Figure 4.12 how local stability is achieved (note that constant distance headway implies zero relative speed). The other figures also highlight the difference in response of FV and the actions of LV. In all the figures there is a lag between the time the LV acts and the FV acts; this is due to the perception–reaction time of the FV.

Fuzzy inference model of car-following behaviour

The fuzzy inference model of car-following is originally due to Chakroborty and Kikuchi [33, 135, 34]. The basic structure of the model is shown in Figure 4.16. As can be seen from the figure, the fuzzy inference model is multiple-stimuli based and

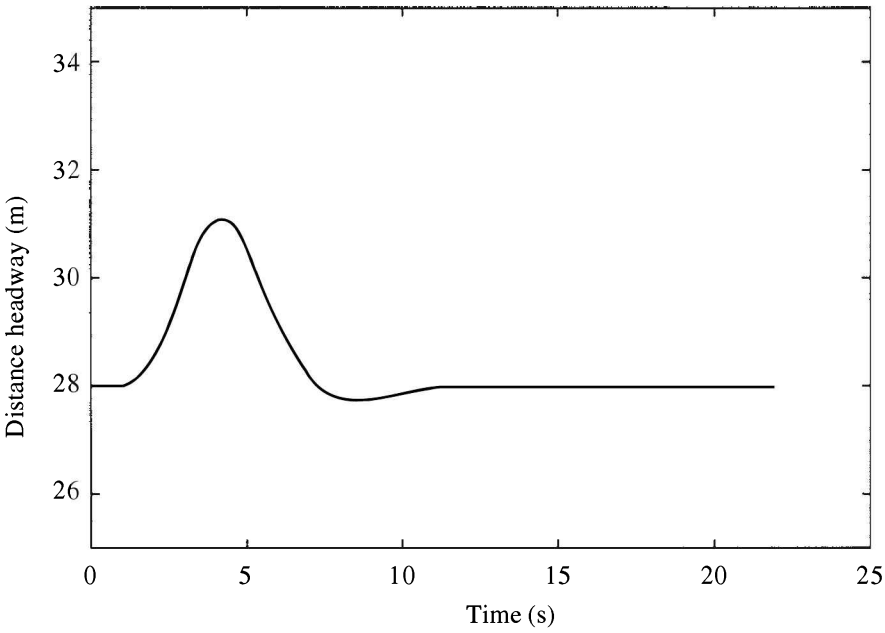


Figure 4.12 Distance headway versus time for Example 4.5.

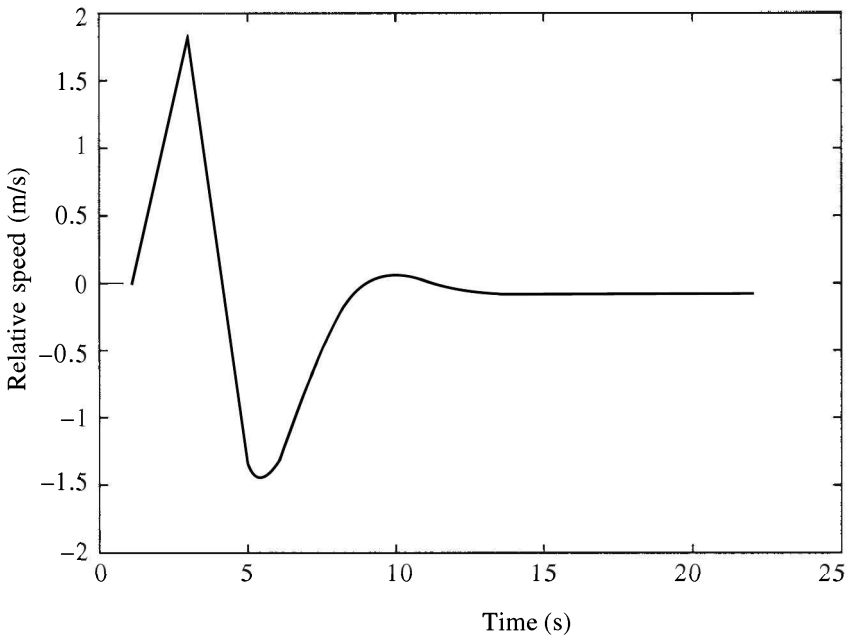


Figure 4.13 Relative speed versus time for Example 4.5.

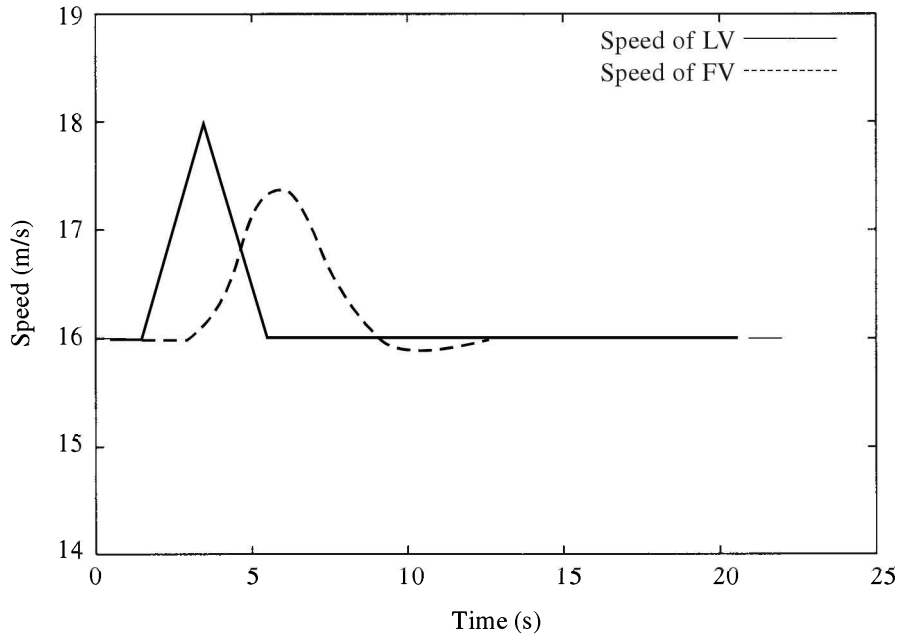


Figure 4.14 Speed versus time for Example 4.5.

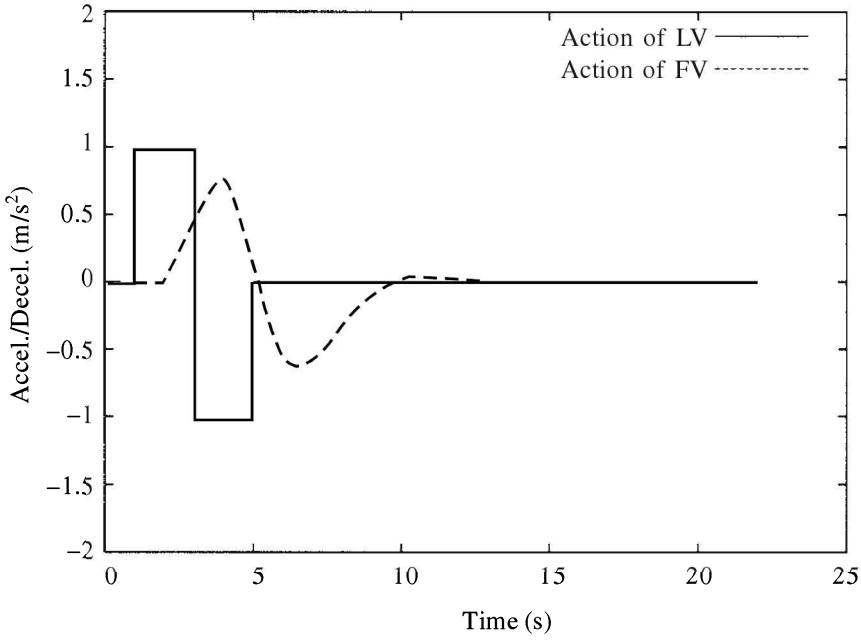


Figure 4.15 Acceleration/deceleration versus time for Example 4.5.

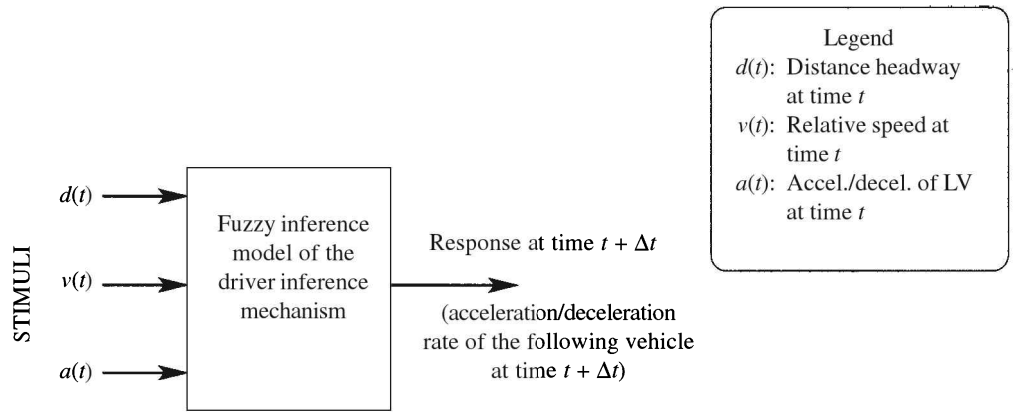


Figure 4.16 Schematic of the structure of the fuzzy inference based model of car-following.

acknowledges the approximate nature of the human control process by modelling the inference process as a fuzzy inference system. The fuzzy inference system is basically a set of rules of thumb where both the antecedents and consequents are fuzzy sets.

Examples of rules which constitute the inference system are given below:

Rule Example I

If (at time t) the *distance headway* is **LARGE** and the *relative speed* is **SMALL NEGATIVE** and *acceleration* of LV is **ZERO**
Then (at time $t + \Delta t$) the *acceleration* of FV should be **SMALL POSITIVE.**

Rule Example II

If (at time t) the *distance headway* is **SAFE** and the *relative speed* is **LARGE POSITIVE** and *acceleration* of LV is **ZERO**
Then (at time $t + \Delta t$) the *acceleration* of FV should be **SMALL POSITIVE.**

Rule Example III

If (at time t) the *distance headway* is **SMALL** and the *relative speed* is **POSITIVE** and *acceleration* of LV is **SMALL NEGATIVE**
Then (at time $t + \Delta t$) the *acceleration* of FV should be **MODERATELY NEGATIVE.**

The fuzzy inference based model has all the properties observed in car-following behaviour. However, more details of the model are not provided here as it requires substantial knowledge of the fuzzy inference mechanism. The interested reader may refer to Zimmermann [270] for an introduction to the topic of fuzzy inference and to the literature cited above for a detailed description of the fuzzy inference based car-following model.

4.3.5 Capacity and Level-of-Service Analysis of Basic Freeway (Expressway) Sections

Basic freeway section is a section of a roadway where interruptions to traffic flow are either absent or inconsequential. In this subsection, the capacity of such road-sections and the level-of-service of traffic streams on such road-sections are discussed. Such analysis is directly related to the design of road-sections where the interruptions are intended to be minimal. Before proceeding with the discussion, certain definitions related to capacity and level-of-service analysis are put forward.

Capacity. This is defined as the maximum number of vehicles that can be expected to cross a point (or line) on the road in a unit interval of time. Thus, the capacity of a road-section is the maximum flow $q_{\max}$ on the section.

Ideal capacity. This is defined as the maximum number of passenger cars (driven by drivers familiar to the area) that can be expected to cross a point (or line) on an ideal road in a unit interval of time. An ideal road-section is one which has ample width (at least 3.5 m wide lanes), wide-paved shoulders (at least 1.8 m wide) and zero gradient.

Level-of-service, LOS. This can be broadly defined as the prevailing conditions under which a driver has to drive. LOS is divided into six classes ranging from LOS(A) through LOS(F). In LOS(A) the driving conditions are the best; traffic is moving in free-flow conditions, a driver faces absolutely no hindrance from other vehicles on the road, the driver is able to choose his/her speed, and so on. In LOS(F), on the other hand, the driving conditions are the worst; traffic is moving in extreme forced-flow conditions, there are frequent stops, the driver is absolutely constrained by the other vehicles on the road, the driving is very taxing, and so forth. A complete definition of the different levels of service can be obtained from either the Indian codes (IRC:64–1990 [83] and IRC:106–1990 [84]) or the *Highway Capacity Manual* [103].

Capacity analysis

Figure 4.17(a) shows a traditional theoretical approximation of the u–q relation. The value of maximum flow $q_{\max}$ is the capacity. Figure 4.17(b) shows a more modern view of the theoretical approximation of the u–q relation. Here, the lower-half is marked as a vague fuzzy area, and no attempt is made to represent this region as a functional relation between u and q. In either case, however, it is quite clear that there exists a speed, u_o, where flow is maximum or reaches its full capacity. The topic of this section is to understand how to analyze or determine this maximum flow value for a given expressway section.

It has been seen that the behaviour of drivers changes (i) when the lane widths are narrow, (ii) when the shoulders are narrow, and (iii) when they are unfamiliar with the region. Further, the value of $q_{\max}$ changes when the traffic stream has heavy, slow moving vehicles.

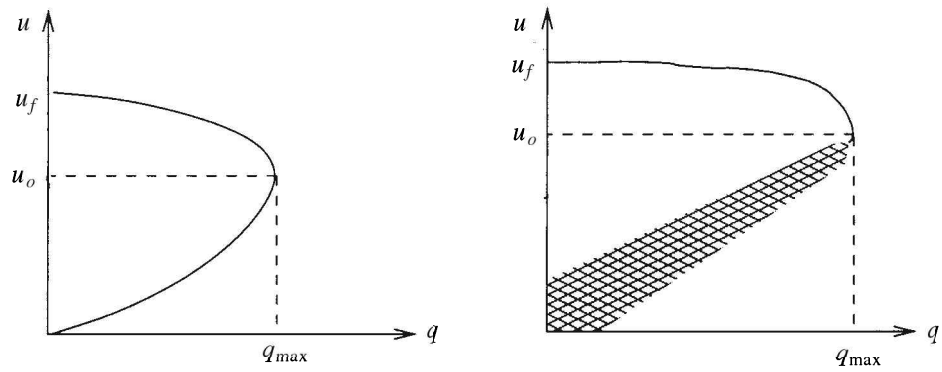


Figure 4.17 Approximations to the *u–q* relation (a) traditional and (b) modern views.

The reason for the behavioural changes in drivers is possibly due to the fact that under the three conditions mentioned above, drivers are more circumspect either because of safety concerns or because of the fear of getting lost. Therefore, under these conditions the speed at which people drive for a given density is lower than the speed at which people would drive at the same density had the conditions been better or ideal. This results in a fall in the value of capacity (as capacity is $u_o \times k_o$).

The reason for the fall in q_{max} in the presence of slow moving and heavy vehicles is twofold: (i) slow moving vehicles have operational problems maintaining the speed of the traffic stream and tend to cause interruptions and (ii) heavy vehicles also occupy more space than passenger cars do.

The problem at hand then is to determine (i) the capacity of roads under ideal conditions and (ii) how the capacity reduces when the actual conditions are different from the ideal conditions. In the following, the principles of two methods, which are philosophically similar, are described. The first method, referred to here as the *IRC/Old HCM method* describes the principles suggested in the IRC codes (which is close to the method proposed by the old *Highway Capacity Manual* [104]). The second method, referred to here as the *New HCM method*, is the one proposed in the 1998 version of the *Highway Capacity Manual* [103]. Both the methods are empirical in the sense that the methodology is purely based on observations.

IRC/OLD HCM METHOD

This method, in principle, views capacity c under a given condition as a reduced value of the capacity under ideal conditions. Hence, c can be expressed as

$$c = c_i f_{wl} f_{hv} f_p \tag{4.19}$$

where

c_i is the ideal capacity in passenger car units per hour

f_{wl} is a factor which modifies the ideal capacity for a non-ideal lane and shoulder widths (it is less than 1)

f_{hv} is a factor which modifies the ideal capacity due to the presence of non-passenger cars in the traffic stream (it is generally less than 1).

f_p is a factor which depends on the proportion of non-commuting drivers (i.e. drivers unfamiliar with the area) in the traffic stream. These factors are generally tabulated in the codes (for example, see HCM [104]) and can be used, given the prevailing conditions, to obtain the actual capacity c in vehicles per hour.

The IRC codes in this regard (for example, see IRC:64–1990 [83] and IRC:106–1990 [84]) are poorly written codes and provide little or no help in understanding the capacity calculations under Indian conditions. For example, (i) IRC codes do not provide any value for capacity of roads, they, however, give volume values representing the maximum allowable flow at one particular LOS for different types of road classes, (ii) only for two-lane rural roads they provide the f_{wl} factors, (iii) there is hardly any documentation on multi-lane facilities, and (iv) they ignore the effect of non-commuters on capacity. However, the IRC codes do suggest that when non-passenger cars are present, the volume should be modified (and not the capacity). To this effect, IRC codes provide passenger car equivalence factors for different vehicle types.

Next, the new HCM method is discussed in a slightly more detailed manner, as the HCM method is complete and provides an up-to-date and modern view on capacity of expressway sections.

NEW HCM METHOD

The new HCM method is based on the tacit implications that non-ideal conditions basically affect the driving environment which in turn affects the capacity of the facility. The effect of non-ideal conditions on driving environment is captured through the variable free-flow speed. That is, according to HCM, the changes in the driving environment can be seen through the variable free-flow speed. Next, the HCM provides different u–q relations for different free-flow speeds (these could be thought of as statements of different driving behaviours) from which capacity can be easily obtained. Since the exact relations and parameter values are not applicable in Indian conditions, the rest of the discussion is kept abstract. The method is presented in a step-by-step manner:

Step 1. Determine the prevailing conditions of the road in terms of lane width, number of lanes, shoulder width, and the number of interchanges (grade-separated intersections; discussed in Chapter 5) per unit length. *Note that a large number of interchanges within a short length reduce capacity due to excessive weaving movements.*

Step 2. Estimate the ideal free-flow speed of the road section based on the type of area the road goes through (for example, rural expressways seem to have greater free speeds than similar expressways through urban areas).

Step 3. Determine (i) the amount of free speed reduction based on lane width, (ii) the amount of free speed reduction based on shoulder width, (iii) the amount of free speed reduction based on the number of lanes, and (iv) the amount of free speed reduction based on the number of interchanges per unit length.

Step 4. Sum up all the reduction amounts and subtract the sum from the ideal free-flow speed.

Step 5. Based on a figure similar to Figure 4.18, determine the capacity in passenger cars per hour per lane (pcphpl) by reading the largest flow value from the correct speed versus flow relation. The correct relation is obtained by interpolating the lines given (for various free-flow speeds) for the free-flow speed obtained in Step 4.

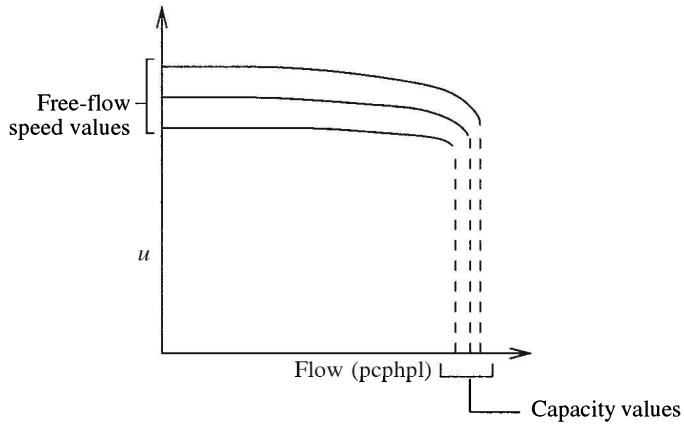


Figure 4.18 Determining expressway capacity according to the new HCM method.

Note that in the new method the effect of vehicle mix and driver population is incorporated in the conversion of the existing (or assumed) traffic volume from vehicles per hour to an equivalent number of passenger cars per hour. This conversion is now discussed in the subsection below.

Level-of-service, LOS, analysis

From the description of level-of-service given earlier in this section, it is clear that LOS being offered by a particular expressway-section at a given time is dependent on (i) the demand at that time and (ii) the capacity of the roadway (supply characteristic). The concept of capacity and its relation to road characteristics has already been discussed. However, no discussion on demand characteristics has been presented so far. Hence, before analyzing the level-of-service of expressway-sections, some features of the demand characteristics on such sections are described.

Demand on an expressway-section can be described as the number of vehicles that would like to cross a given point on the road during a specified interval of time. Most of the time the measured flow is equal to the demand. However, sometimes during forced flow conditions, the measured flow is less than the demand (which in some sense leads to the congested conditions). Measuring demand in such situations is an involved process and beyond the purview of this text. The interested reader may refer to May [154] for a good discussion on this topic.

The IRC codes do not provide any clear procedure to determine the LOS of an expressway-section under the prevailing conditions. The HCM [103] on the other hand provides a comprehensive procedure to determine LOS. Again, since the specific parameter values are not applicable to Indian conditions, the LOS determination procedure is described in an abstract manner. As per the 1998 HCM [103], the following steps can be used to obtain LOS under the prevailing conditions of demand.

Step 1. The first task is to convert the volume q, in vehicles per hour, to the within the hour peak flow rate f in (commuter) passenger cars per hour per lane (pcphpl). This conversion is achieved in two steps: (i) by dividing q by the peak hour factor PHF and the number of lanes N in the expressway, and (ii) by converting all non-commuting drivers to commuters (through an empirical factor f_p which tries to account for the differences in behaviour of a driver familiar with area from another driver unfamiliar with the area), and all non-passenger cars to passenger cars (through an empirical factor, f_{hv}, which converts all types of vehicles into passenger cars through empirically derived passenger car equivalence factors). Thus,

$$f = \frac{q}{\text{PHF} \times N \times f_{hv} \times f_p}$$

Step 2. Determine the free-flow speed for the expressway-section using Step 1 through Step 4 of the procedure described under the *New HCM method* for capacity determination.

Step 3. From a figure similar to Figure 4.19, determine the LOS. First, obtain the correct u–q relation for the free-flow speed obtained in Step 2. Next, for the given flow (in pcphpl) draw a vertical line till it intersects the correct relation at some point X. Now read the zone in which X lies. If X is in the zone termed i, then the LOS for the prevailing conditions is i.

It can be seen from the figure that the different level-of-service zones are determined based on the density values (recall that flow divided by speed is density). For example, for density values between k_2 and k_3 the LOS is C. Another point which should be mentioned here is that from the same figure we can determine the speed and density of the stream. For example, the point Y at which a horizontal line from point X intersects the abscissa gives the value of the speed of the stream under prevailing volume. Further,

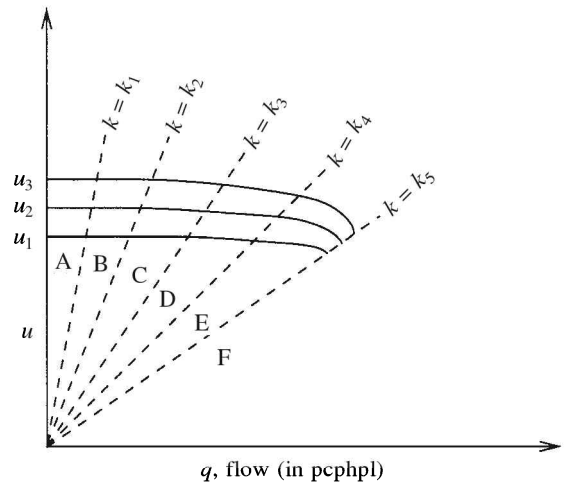


Figure 4.19 Determining the expressway level-of-service according to the new HCM method.

f divided by the speed at Y gives the density of the stream in passenger cars per km per lane.

4.4 FUNDAMENTALS OF INTERRUPTED TRAFFIC FLOW

In this section, the flow of traffic under traffic interruptions is studied. Traffic interruptions are collectively a general situation used to denote conditions where a stream of traffic flowing under certain conditions of speed and density meets another stream flowing under some other condition of speed and density. For example, if traffic flowing at high speed meets another stream flowing at a greater density and lower speed, the former traffic stream will face a traffic interruption. There are various ramifications of such interruptions on traffic flow and these are discussed in Section 4.4.1 on *shock waves*. Within this general description of traffic interruptions, two kinds of traffic interruptions are of special interest. These are the interruptions that take place at signalized and unsignalized intersections.

In this section, there are three subsections; the first studies *shock waves*. The second and third study the traffic flow at signalized and unsignalized intersections, respectively.

4.4.1 Shock Waves

Whenever a stream of traffic flowing under certain stream conditions (say, speed = u_A, density = k_A, and flow = q_A) meets another stream flowing under different conditions (say, speed = u_B, density = k_B, and flow = q_B) a shock wave is started. The *shock wave* is basically the movement of the point that demarcates the two stream conditions. This

demarcation point may move forward or backward or stay at the same place with respect to the road. The rate at which this demarcation point moves (the direction of motion of the vehicles is taken as the positive direction) is referred to as the speed of the shock wave.

In order to see the generation and movement of shock waves, consider the distance–time graph shown in Figure 4.20. This figure is drawn for the following situation. A slow moving vehicle with speed u_B (whose distance–time plot is shown as a dotted line) enters a traffic stream originally moving at u_A, k_A, q_A. This slow moving vehicle slows the traffic and creates another flow condition denoted by u_B, k_B, q_B. From the graph it can be seen that there exists a line (a bold line marked Shock wave 1) that is the locus of the point that demarcates the two flow conditions at any given time. After a while at point Q the slow moving vehicle leaves the traffic stream and the congested condition created by the slow moving vehicle is released, say, at some other stream condition (denoted by u_C, k_C, q_C). Obviously, the point that demarcates the flow conditions u_B, k_B, q_B from the flow conditions u_C, k_C, q_C also causes a shock wave. The locus of this point is marked as Shock wave 2 in the figure. At some time t these two shock waves meet signalling the end of the flow conditions u_B, k_B, q_B. But this time, the flow conditions u_A, k_A, q_A meet

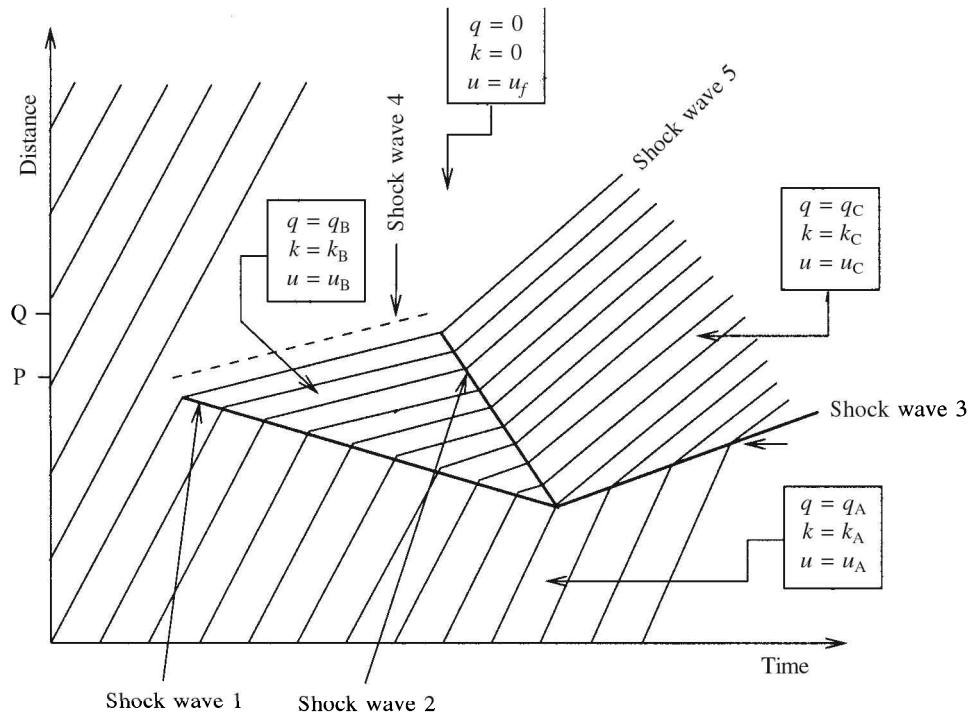


Figure 4.20 Distance–time diagram illustrating the creation and movement of shock waves.

the flow conditions u_C, k_C, q_C starting Shock wave 3 (see Figure 4.20). Interestingly, the vacant area in the figure bounded by the line denoting the motion of last vehicle to go without being caught behind the slow moving vehicle, the dotted line denoting the motion of the slow moving vehicle, and the line denoting the motion of the first vehicle to get released after the departure of the slow moving vehicle, also represent a flow condition, namely the *free-flow conditions*. In the figure, this zone is indicated as a stream with conditions $u = u_f$ (the free-flow speed), $k = 0$, and $q = 0$. Hence, when the stream condition u_B, k_B, q_B meets the stream condition $u = u_f$, $k = 0$, and $q = 0$, a shock wave should and does emanate. The only difference here is that the point which demarcates the two conditions is the same as the point representing the slow moving vehicle; hence the shock wave that emanates here moves with the slow moving vehicle and the time–distance diagram of this shock wave is the same as the time–distance diagram of the slow moving vehicle. Similarly, the time–distance diagram of the first vehicle that is released is the time–distance diagram of the shock wave that emanates when the flow condition u_C, k_C, q_C meets $u = u_f$, $k = 0$, and $q = 0$.

From the above discussion it is clear that the key parameters of a shock wave are (i) the point at which it starts, (ii) the point at which it ends, and (iii) the speed of the shock wave. As will be apparent later, of these, the only parameter which needs to be studied in detail is the speed of the shock wave which forms the topic of the next subsection.

Speed of shock waves

Let a stream flowing under condition A (with u_A, k_A, and q_A) meet another stream flowing under condition B (with u_B, k_B, and q_B). Assume that the speed of the resultant shock wave is u_{sw}. Then relative to this shock wave, vehicles in condition A are moving at a speed of $u_A - u_{sw}$ and those in condition B are moving at a speed of $u_B - u_{sw}$.

Now recall that the shock wave is a demarcation between the two traffic conditions. Hence, it can be said that in a time duration Δt the number of vehicles crossing over the shock wave from condition A is $(u_A - u_{sw})\Delta t \times k_A$. Similarly, it can be said that in a time duration of Δt the number of vehicles crossing over the shock wave from condition B is $(u_B - u_{sw})\Delta t \times k_B$. Note that, physically, some vehicles are crossing over the shock wave from one condition to the other. Since vehicles are neither created nor destroyed in the process of crossing over, the number of vehicles crossing over the shock wave from the perspectives of conditions A and B must be equal. Therefore,

$$(u_A - u_{sw})\Delta t \times k_A = (u_B - u_{sw})\Delta t \times k_B$$

Rearranging the terms and substituting $u_i k_i$ by q_i, the following relation can be obtained:

$$u_{sw} = \frac{q_A - q_B}{k_A - k_B} = \frac{q_B - q_A}{k_B - k_A} \tag{4.20}$$

It should be noted that the above equation for the speed of the shock wave can also be obtained using the principles of geometry from a distance–time plot of the type shown

in Figure 4.20. Further, the above equation also has a simple graphical interpretation; it states that the speed of the shock wave is given by the slope of the line joining the points representing the two conditions (on a q–k graph) whose confluence gives rise to the shock wave. Figure 4.21 shows this graphically.

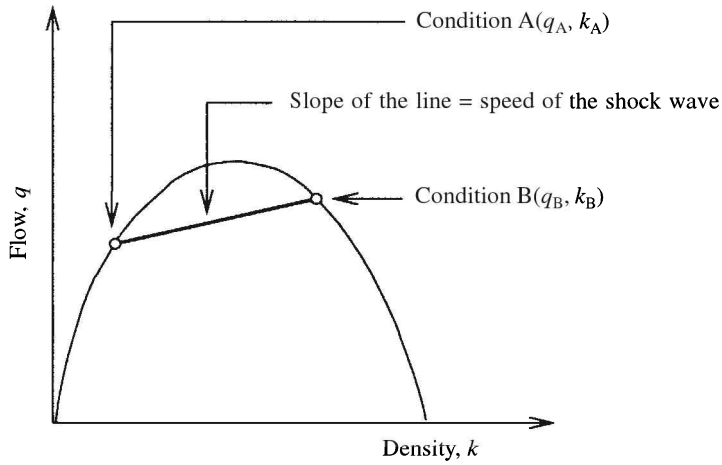


Figure 4.21 Illustration of the speed of the shock waves on a q–k plot.

There can be three types of shock waves: (i) the forward moving shock wave, i.e. the speed of the shock wave is positive [see Figure 4.22(a)], (ii) the stationary shock, i.e. the speed of the shock wave is zero [see Figure 4.22(b)], and (iii) the backward moving shock wave, i.e. the speed of the shock wave is negative [see Figure 4.22(c)]. As can be seen from the figure, the first type of shock wave will occur when a stream with lower flow and lower density meets a stream with higher flow and higher density or when a stream with higher flow and higher density meets a stream with lower flow and lower density. Stationary shock waves will occur when the streams meeting have the same flow value but different densities. The third kind of shock waves will occur when a stream with higher flow and lower density meets a stream with lower flow and higher density or when a stream with lower flow and higher density meets a stream with higher flow and lower density.

In the following, an example is worked out to show how the knowledge of shock waves can be used to obtain different traffic flow parameters of interest and also to illustrate how we can obtain information about where a shock wave starts and where it ends (which are the other two parameters related to the description of a shock wave).

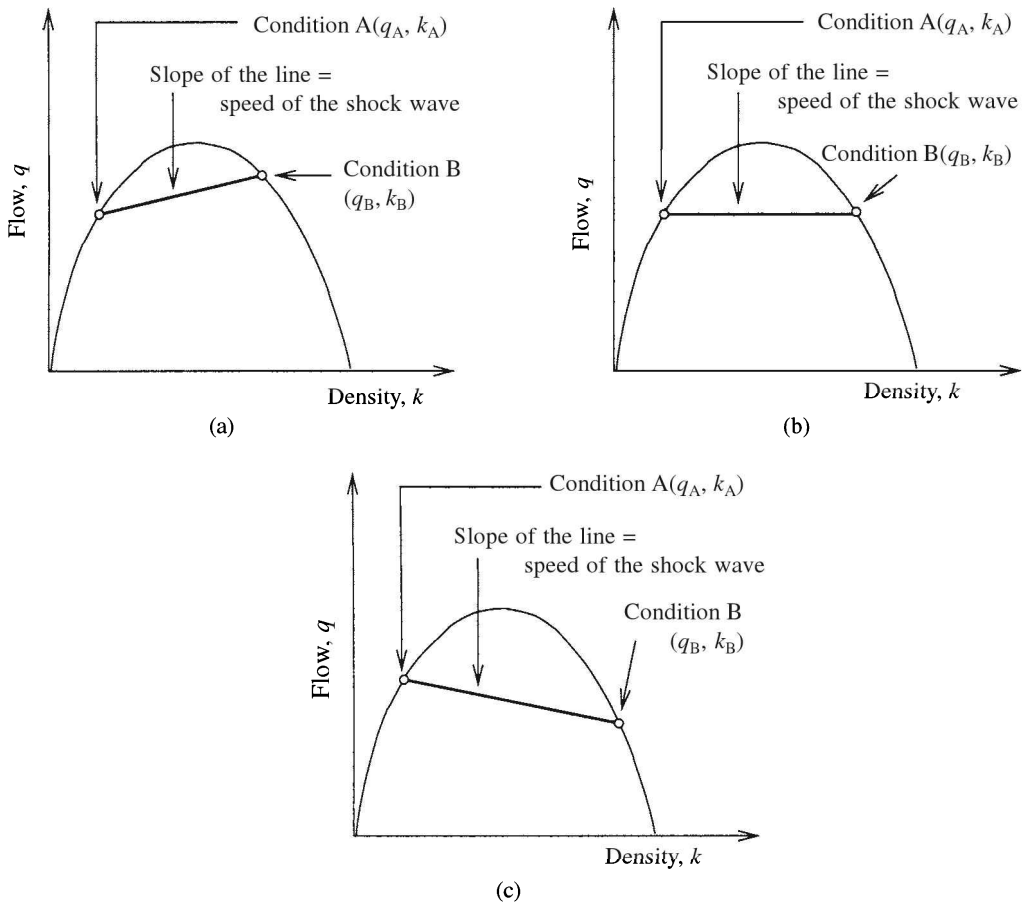


Figure 4.22 Illustration of different types of shock waves on a q–k plot.

EXAMPLE 4.6

Traffic is moving on a one-way road at $q_A = 1000$ vph, and $k_A = 16$ vpkm. A truck enters the stream at point P (which is at a distance of 1 km from an upstream benchmark point BM) at a speed of $u_B = 16$ kmph. Due to the decreased speed, the density behind the truck increases to 75 vpkm. After 10 minutes, the truck leaves the stream. The platoon behind the truck then releases itself at capacity conditions, $q_C = 1400$ vph and $k_C = 44$ vpkm. Determine (i) the speed of all shock waves generated, (ii) the starting point of the platoon (behind the truck) forming the shock wave, (iii) the starting point of the platoon dissipating the shock wave, (iv) the ending points of the platoon forming and platoon dissipating shock waves, (v) the maximum length of the platoon, and (vi) the time it takes for the platoon to dissipate, and also plot the (vii) location of the front of the platoon and the rear of the platoon versus time, and (viii) length of the platoon versus time.

Solution

Consider the distance–time diagram shown in Figure 4.23 plotted for the scenario described in the problem. This diagram is shown here to help the reader understand the problem better; strictly speaking the complete diagram is not necessary for solving the problem. However, an understanding of the physical scenario is of definite help.

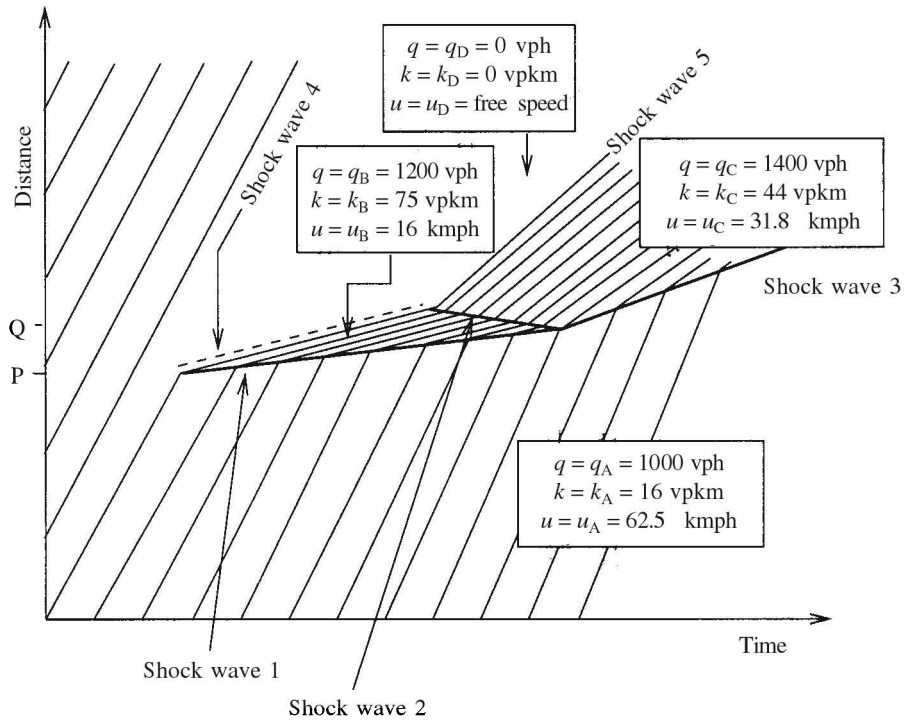


Figure 4.23 Distance–time diagram illustrating Example 4.6 on shock waves.

(i) Speeds of the various shock waves generated (shock wave i is denoted as SWi) can be obtained directly by using Eq. (4.20) as follows:

$$u_{SW1} = \frac{q_B - q_A}{k_B - k_A} = \frac{1200 - 1000}{75 - 16} = 3.39 \text{ kmph}$$

$$u_{SW2} = \frac{q_C - q_B}{k_C - k_B} = \frac{1400 - 1200}{44 - 75} = -6.45 \text{ kmph}$$

$$u_{SW3} = \frac{q_C - q_A}{k_C - k_A} = \frac{1400 - 1000}{44 - 16} = 14.29 \text{ kmph}$$

$$u_{SW4} = \frac{q_B - q_D}{k_B - k_D} = \frac{1200 - 0}{75 - 0} = 16 \text{ kmph}$$

$$u_{SW5} = \frac{q_C - q_D}{k_C - k_D} = \frac{1400 - 0}{44 - 0} = 31.8 \text{ kmph}$$

(ii) Shock wave 1 is the platoon forming shock wave. It starts at point P and at the time when the truck enters the stream.

(iii) Shock wave 2 is the platoon dissipating shock wave. It starts at point Q (i.e. where the truck leaves the stream) 10 minutes (note that the truck remains in the stream for 10 minutes) after the truck entered the traffic stream. Point Q is 16 × (10/60) = 2.67 km downstream of point P.

(iv) Both shock waves 1 and 2 will end if the platoon condition (i.e. condition B) ends. This condition will end whenever Shock waves 1 and 2 meet. Say, they meet at time t hours after the start of Shock wave 1, where their positions must be the same. Their positions at time t can be determined from their starting positions and distances by which they travel during time t. Thus, knowing that P is 1 km from BM and Q is 3.67 km from BM, we can write the following

$$1 + 3.39t = 3.67 - 6.45\{t - (10/60)\}$$

or

$$t = \frac{3.745}{9.84} = 0.381 \text{ h} = 22.84 \text{ min}$$

Hence, the two shock waves end 22.84 minutes after the start of Shock wave 1 and at a distance of 1 + 3.39 × 0.381 = 2.29 km downstream of BM.

(v) The maximum length of the platoon will be at the instant where Shock wave 2 is just about to start. The platoon at any given time is defined by the length between the front of the platoon (Shock wave 4) and the rear of the platoon (Shock wave 1). Hence, the length of the platoon grows at a speed of 16 − 3.39 = 12.61 kmph. The length is maximum at 10 minutes after the platoon starts forming. Hence the maximum length is equal to 12.61 × (10/60) = 2.1 km. In terms of the number of vehicles the maximum length of the platoon is $k_B \times 2.1 = 75 \times 2.1 = 157.5 \approx 158$ vehicles.

(vi) In part (iv), it was determined that the platoon ceases to exist 22.84 minutes after the start of platoon formation. Out of this for the first 10 minutes the platoon only grows (and there is no dissipation). Hence, it takes 12.84 minutes for the platoon to dissipate.

(vii) Figure 4.24(a) shows the required plot. In the plot, time is assumed to be zero when Shock wave 1 starts; the distances are as measured from BM. In the figure, Lines 1 and 2 represent the front of the platoon and Line 3 represents the rear of the platoon. Further,

the slope of Line 1 is equal to u_{SW4}, the slope of Line 2 is equal to u_{SW2}, and the slope of Line 3 is equal to u_{SW1}.

(viii) Figure 4.24(b) shows the required plot. In the plot, time is assumed to be zero when Shock wave 1 starts. Slope of Line 1 is equal to the rate of growth of the platoon. This value, as determined in part (v), is 12.61 kmph. The slope of Line 2 is basically the rate of dissipation of the platoon; however, this need not be calculated since we know the maximum length of the platoon and when the platoon completely dissipates.

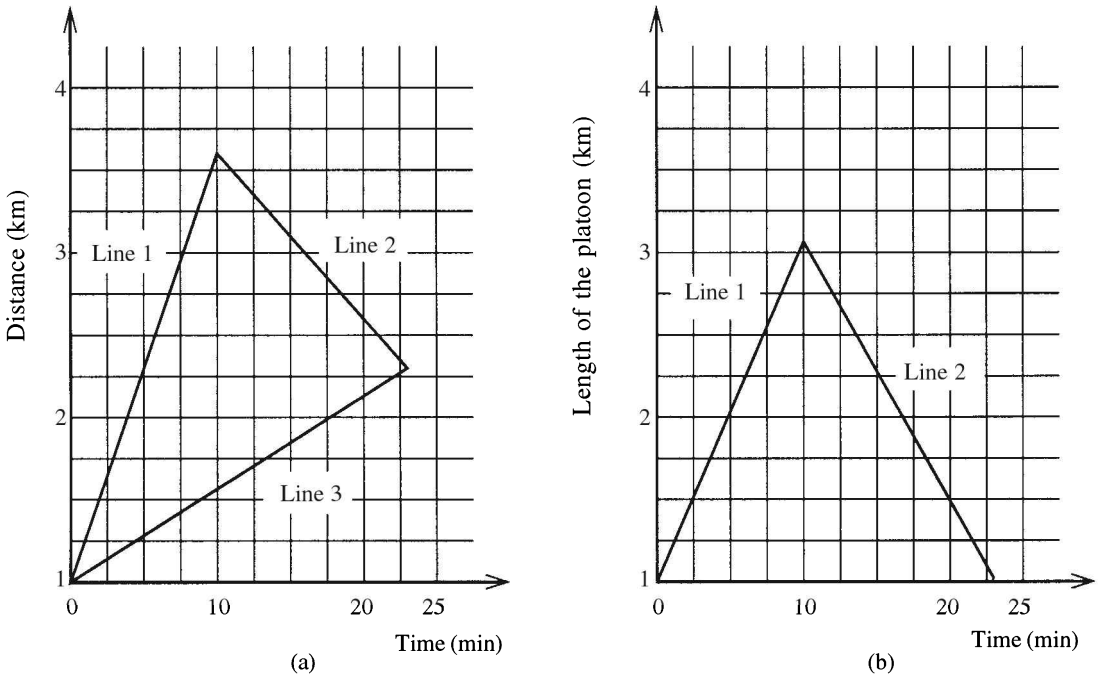


Figure 4.24 (a) Example 4.6: Plot of platoon front and rear locations versus time (b) plot of length of platoon versus time.

4.4.2 Traffic Flow at Signalized Intersections

An intersection is a location where two or more roads carrying traffic streams in different directions cross. The space which is common to all these roads is referred to as the intersection. At such a location, obviously, different traffic streams compete with one another for the use of the common space or the intersection. If left on its own, the flow at an intersection will always be chaotic; the safety and efficiency at such locations will be low. Hence, various strategies are used to control the flow of traffic at an intersection in order to improve the safety and efficiency of traffic flow. Among the strategies that are used, signalization is the most common.

At a signalized intersection, the common space is periodically given to certain flows while the other conflicting streams are barred from entry at that time. In a manner of speaking, the common space is *time-shared* among the various flows. Although, there are various kinds of time-sharing strategies like *pre-timed*, *partially actuated*, and *fully actuated* signalizations, in this section the *pre-timed* signalization and its effect on flow is studied. The other strategies, where the time-sharing mechanism changes more frequently than in the pre-timed strategy, are not studied here as the basic traffic flow analysis process is the same as that of the *pre-timed* strategy.

In the pre-timed signalization, the time sharing between the different conflicting flows occurs according to a pre-defined strategy which repeats at a fixed interval. This fixed interval is referred to as the *cycle length*. During the cycle length, the time for which a particular stream can utilize the intersection is referred to as the *green time* for that stream or movement, the time during which a particular movement cannot utilize the intersection is referred to as the *red time* for that movement. Invariably during the change-over from green to red an amber signal is shown to warn the driver that a red signal is impending. During the amber time for a movement, the vehicles of that movement can use the intersection. Of course, the sum of green, amber and red times for a particular movement is equal to the cycle time. A detailed discussion and design of *pre-timed* signalization is provided in Chapter 5. The above brief introduction is given here in order to initiate the reader to the various terminologies used in signalized intersection analysis and design.

Obviously, at a signalized intersection, there is interruption to flow of traffic. The type of interruption and its effect on the flow is described in the next several subsections.

Flow characteristics

The interruption to traffic flow at a signalized intersection is orderly and deterministic. Consider the following scenario. The signal has just turned red for a particular stream or movement. All vehicles on this stream come to a stop and remain stopped (thereby forming a queue) till the light turns green. Once the light turns green the first vehicle in the queue departs followed by the other vehicles. Movement of vehicles continues unabated till the light turns amber at which point vehicles close to the intersection generally go through while the ones farther away from the intersection initiate manoeuvres to come to a stop. The same pattern follows for every cycle. Given this interruption pattern, the following processes become important for analysis: (i) the arrival process of vehicles, (ii) the departure process of vehicles, (iii) the queue of vehicles, and (iv) the delay to vehicles.

In this subsection, the first two processes are described. The next subsection analyzes the other two processes. The last subsection in this topic analyzes the related matter of capacity and level of service at signalized intersections.

ARRIVAL PROCESS

The arrival processes at intersections could be of three kinds: (i) random arrivals,

(ii) grouped arrivals, or (iii) mixed arrivals. In random arrivals vehicles seem to arrive at the intersection randomly. Such an arrival pattern is seen at isolated intersections, i.e. intersections at locations where there are no other upstream intersections in the vicinity (say within 3 to 4 km). In these cases the inter-arrival times (or headways) are often distributed more or less according to the *negative exponential distribution*. That is, if h is the headway between vehicles, then the probability that a particular headway is between H_1 and H_2 is given by

$$\mathrm{P}(H_1 \leq h \leq H_2) = e^{-\lambda H_1} - e^{-\lambda H_2} \tag{4.21}$$

where λ is the reciprocal of the mean or average headway. The above relation is derived by integrating the negative exponential probability density function (for headways in this case) between H_1 and H_2. The negative exponential probability density function $f(h)$ (for all $h > 0$) is given as

$$f_h = \lambda e^{-\lambda h} \tag{4.22}$$

An assumption (or observation) of negative exponential distribution for headways also implies that the vehicle arrival process is a Poisson process. That is, the probability that the number of vehicles N_t, that arrive in a time interval t, is equal to k is given by

$$\mathrm{P}(N_t = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{4.23}$$

where λ is the reciprocal of the mean or average headway or, alternatively, it is the mean or average arrival rate.

Grouped arrivals are seen at intersections which are located close to (say, within 2 km) another upstream intersection. In such cases, the arrival process seems to be uniform and vehicles can be assumed to arrive at reasonably constant headways. This phenomenon occurs because vehicles arriving at the intersection are the ones which have been released by an earlier intersection and therefore are in a platoon.

Mixed arrivals are seen at intersections which are located at intermediate distances (say, between 2 to 4 km) from another upstream intersection. Here, the arrival cannot be characterized either as purely random or as purely grouped. This is because the distance between the upstream intersection and the intersection being studied is large enough for many of the released vehicles to disperse from the discharged platoon and arrive independently; yet the distance is not large enough for the entire platoon to disperse. Hence, some vehicles still arrive in a grouped manner.

DEPARTURE PROCESS

When a signal turns green from red, the first of the stopped vehicles initiates manoeuvres to move and cross the intersection. The next to cross the intersection is the second vehicle in the queue and so on. If one measures the headways, i.e. the time gaps between successive vehicles when they cross a pre-specified point on the intersection (generally

the stop line on the road), then an interesting and expected pattern emerges. Figure 4.25 shows a typical plot of headways versus position in the queue. The ordinate value for an abscissa value of i gives the headway between the ith and the $(i - 1)$th vehicles in the queue when they cross the pre-specified point in the intersection. Further, the ordinate value for an abscissa value of 1 indicates the time-gap between the light turning green and the first vehicle crossing the pre-specified point in the intersection.

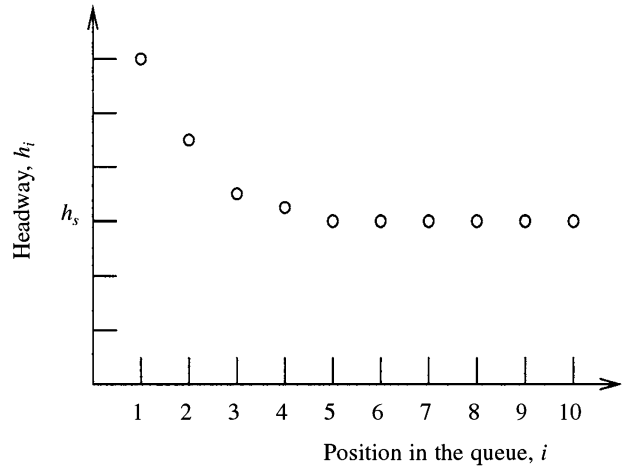


Figure 4.25 A typical plot of departure headways versus position in the queue at a signalized intersection.

From Figure 4.25, two features emerge: (i) the headway stabilizes to a value h_s, referred to as the *saturation headway*; this value basically states the maximum number of vehicles that can ever be released during a specified green time and (ii) the initial headways are larger than h_s. The second feature highlights that although vehicles can move at a headway of h_s, the initial vehicles take a longer time due to perception–reaction time (to the light turning green) and the extra time taken to accelerate to a reasonable speed (note that the latter vehicles more or less achieve this speed when they cross the specified point as they start moving from a distance further upstream from the specified point. In a sense then, some time is lost due to the fact that the initial vehicles take longer than h_s. The sum of these excess times is referred to as the *start-up lost time*, l_s. That is,

$$l_s = \sum_{\forall i} (h_i - h_s) \tag{4.24}$$

Near the end of the departure process some time is also lost. This happens because invariably some part of the amber time remains unutilized because vehicles come to a stop even when some part of the amber time is still remaining. This loss of time, referred

to as *movement lost time* (or sometimes as clearance lost time) l_m, is primarily due to the fact that drivers are never aware of the remaining amber time.

Delay and queue analysis

In this section the delay faced by vehicles at signalized intersections and the queues developed at signalized intersections are studied. Consider the plot shown in Figure 4.26. In this figure, the abscissa is time and the ordinate is the cumulative number of arrivals as well as the cumulative number of departures for a given stream (or approach) at an intersection. There are two lines in the figure. One shows a typical graph for cumulative number of arrivals on the given approach at a signalized intersection, the other shows the cumulative number of departures from the given approach at the signalized intersection. On the abscissa, the time is divided into slots named Cycle I, Cycle II, and so on. These slots represent the cycles at the intersection. Each cycle is further subdivided into R and G. The R represents the duration of effective red (i.e. the time during which no vehicle on this particular approach crosses the intersection); similarly the G represents the duration of effective green (i.e. the time during which vehicles on this particular approach cross the intersection).

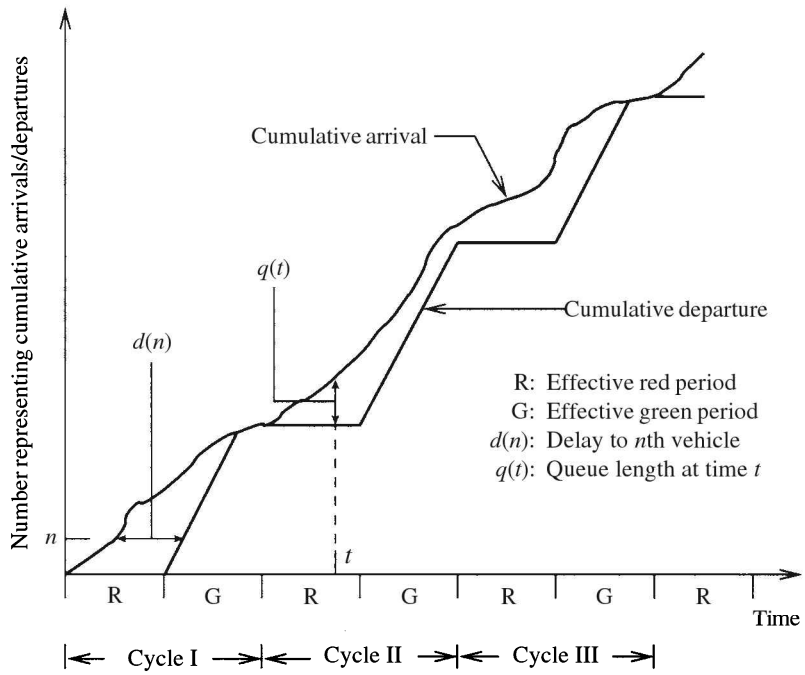


Figure 4.26 A typical plot of cumulative number of arrivals and departures on an approach to a signalized intersection.

Figure 4.26 gives a reasonably complete picture of the arrival and departure processes at the intersection. From such a plot, we can obtain information about both delay and queues. The horizontal distance at a value of n on the ordinate, for example, will give the delay faced by the nth vehicle to arrive at the intersection. Hence summing all such horizontal distances (or equivalently the area between the two lines) will give the total delay faced by all the vehicles arriving at the intersection. The total delay divided by the total number of arrivals will provide the average delay.

Similarly, the queues on the approach can be easily determined from this figure. For example, the vertical distance between the two lines at time t will give the queue length at the intersection approach at time t. This implies that the queue lengths at any given time can be obtained easily from the figure and hence the parameters such as average queue length, variance of queue lengths, and the like can also be obtained.

However, obtaining such graphs for each and every intersection at all times is not feasible. Hence it is imperative that we analyze the delay to vehicles and queues with an aim to derive equations which can give these quantities once the data on arrival rates, cycle lengths, green times, red times, etc. are known. In the following, such a description of the analysis procedures is provided.

DELAY ANALYSIS

To begin with assume that the arrival process is deterministic and vehicles arrive at a uniform rate. Further, assume that the system is unsaturated, that is, the total number of vehicles that arrive in a period is less than the total number of vehicles that can be served by the system. These two assumptions mean that the arrival rate is such that all the vehicles that come in a cycle are cleared within the same cycle (like the situation in Cycle I of Figure 4.26).

The average delay to vehicles for this case can then be easily determined from Figure 4.27. The figure shows a typical cumulative arrival/departure graph against time

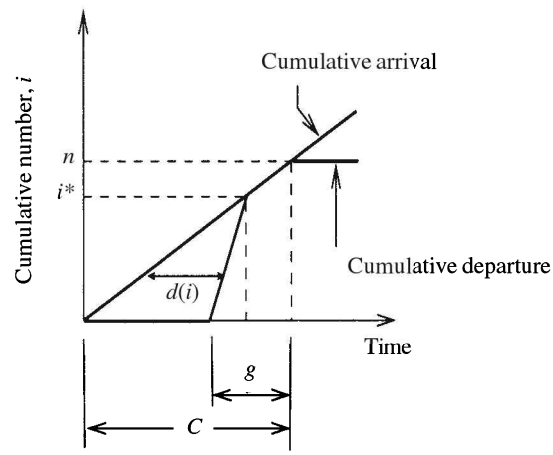


Figure 4.27 A typical plot of cumulative number of arrivals and departures of vehicles on an unsaturated, uniform arrival rate approach to a signalized intersection.

for an unsaturated, uniform arrival rate approach to an intersection. The slope of the cumulative arrival line is v, where v is the uniform arrival rate in vehicles per unit time. The slope of the cumulative departure line is sometimes zero (when the light is red) and sometimes s (when the light is green); where s is the saturation flow rate obtained as the reciprocal of the saturation headway explained earlier; s is expressed as vehicles per hour of green per lane or vphgpl.

From Figure 4.27 (where C is the duration of the cycle length and g the duration of the effective green period) it can be seen that the total delay (under the assumptions stated above) $TD_{u,us}$ is given by

$$TD_{u,us} = \sum_{i=1}^{i=n} d(i) \tag{4.25}$$

Assuming that n is large enough so that the discrete sum of $d(i)$ is equal to the area of the triangle in the figure, the following can be written:

$$TD_{u,us} = 0.5i^*(C - g) \tag{4.26}$$

From the figure, i^* can be easily determined by noting that

$$i^* = vt$$

where

$$vt = s[t - (C - g)]$$

Determining t from the above relation, i^* can be written as

$$i^* = \frac{vs(C-g)}{s-v}$$

Hence,

$$TD_{u,us} = \frac{vs(C-g)^2}{2(s-v)} \tag{4.27}$$

From the above and noting that n, the total number of vehicles that arrived, is vC the average delay under the above assumptions, $D_{u,us}$, can be obtained as

$$D_{u,us} = \frac{TD_{u,us}}{vC} = \frac{(C-g)^2 s}{2(s-v)C} = \frac{C(1-g/C)^2}{2(1-v/s)} \tag{4.28}$$

However, an approach to an intersection may not always stay unsaturated. There may be periods of oversaturation during which the arrivals from one cycle spill over to the next and so on. A similar situation can be seen in Cycle II of Figure 4.26. Under this assumption, and the assumption that the arrival rate is still deterministic and uniform, the plot of cumulative arrivals/departures versus time can be drawn (see Figure 4.28). In the figure it is assumed that the arrival rate from 0 to time T is v, and that v is large enough to cause oversaturation (i.e. vC is greater than sg, the maximum number of vehicles that can be served during a cycle).

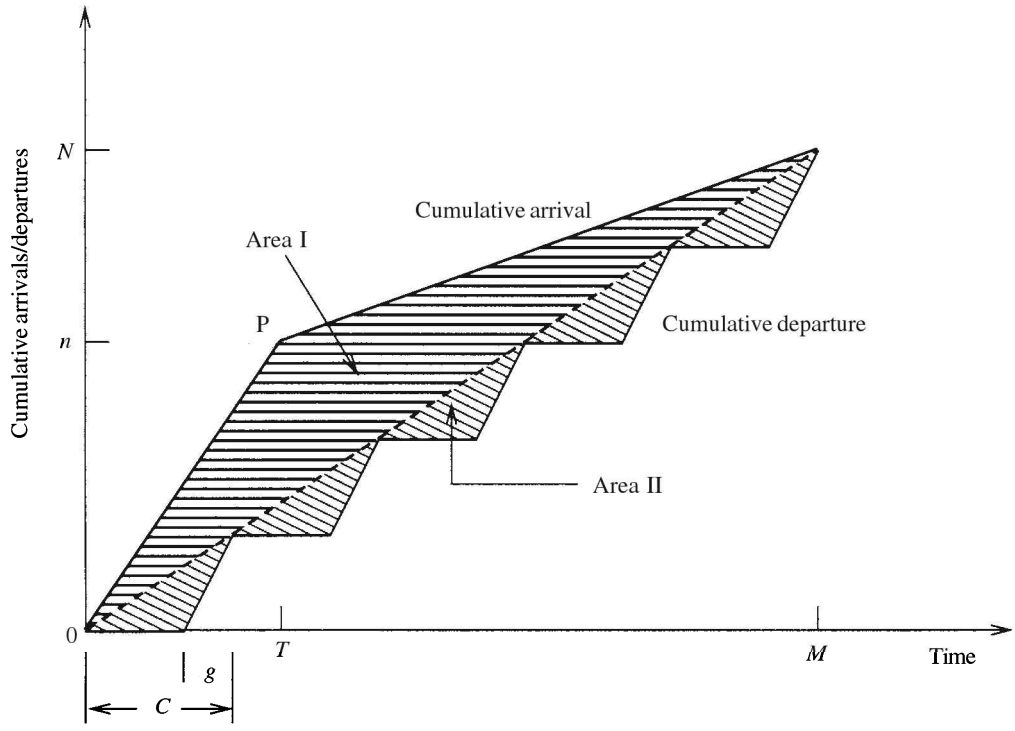


Figure 4.28 A typical plot of cumulative number of arrivals and departures on an oversaturated, uniform arrival rate approach to a signalized intersection.

From Figure 4.28 it can be seen that the total delay under these assumptions, i.e. $TD_{u,os}$ is given by the sum of the area marked with horizontal stripes (Area I) and the area marked with inclined stripes (Area II). This implies that the average delay, i.e. $D_{u,os}$ under these assumptions is the sum of the average delay due to Area I and the average delay due to Area II.

The average delay due to Area II can be easily determined by assuming the dashed line as the cumulative arrival line and using Eq. (4.28). However, first the slope of the dashed line needs to be determined and later substituted for v in Eq. (4.28). If the slope of the dashed line is taken as σ (and noting that the slope of the inclined part of the cumulative departure line is s, the saturation flow rate) then

$$\sigma C = sg$$

or

$$\sigma = \frac{g}{C}s$$

Substituting this expression of σ for v in Eq. (4.28), we can write the average delay due to Area II, i.e. AD_{II} as

$$AD_{II} = \frac{C(1-g/C)^2}{2(1-gs/Cs)}$$

$$= \frac{C-g}{2} \tag{4.29}$$

The average delay due to Area I can be determined by looking at the average time between the cumulative arrival line and the dashed line. If the horizontal distance (i.e. time in this case) between point P and the dashed line is taken as Z, then it can be said that the time between the cumulative arrival line and the dashed line increases linearly from 0 to Z over a time period of 0 to T. From this it can be said that the average delay for vehicles arriving between time 0 and T is $Z/2$ (Note that the area of the triangle formed by the cumulative arrival line, a horizontal line from P and the dashed line is given by $Zn/2$, where n is the number of vehicles to arrive till time T).

Similarly, the time between the cumulative arrival line and the dashed line decreases linearly from Z to 0 over a time period of T to M. Following the same logic as above, the average delay due to Area I to vehicles arriving between times T and M is $Z/2$. Hence, it can be said that the average delay due to Area I is $Z/2$ irrespective of when a vehicle arrives.

If one assumes the vertical distance of point P from the dashed line as y, then

$$\frac{y}{Z} = \text{slope of the dashed line, } \sigma$$

and

$$y = vT - \sigma T$$

hence,

$$Z = \frac{T(v-gs/C)}{gs/C} \tag{4.30}$$

or

$$Z = T\left(\frac{v}{c} - 1\right)$$

where, $c = (g/C)s$, represents the maximum number of vehicles that can cross the intersection from a given approach per unit time.

Thus the average delay i.e. $D_{u,os}$ is given as

$$D_{u,os} = \frac{T}{2}\left(\frac{v}{c} - 1\right) + \frac{C-g}{2} \tag{4.31}$$

In reality, however, more often than not the arrival is not deterministic, it is stochastic. Once it is assumed that the arrival is stochastic the above relations cannot be used and can at best function as approximate estimates. Assumption of stochastic arrivals

will lead us to analysis which is beyond the scope of this book. Hence, in this text, only the relations generally used to determine delay under the above assumptions are described.

One of the relations often used to determine delay is due to Webster [259]. Webster assumed that the arrivals are according to a Poisson distribution, departures occur uniformly and at a maximum rate of s, the average arrival rate v is such that $vC \leq gs$, and that the queueing process runs under similar arrival and departure conditions long enough for the system to stabilize to a steady state (where, for example, cycle to cycle variations in average delay, average queue length, etc. are minimal). Based on these assumptions and some simulation runs (where the arrival and departure processes at a signalized intersection are simulated for the various arrival pattern and signal settings) Webster proposed the following relation for the average delay, $D_{s,web}$:

$$D_{s,web} = \frac{C(1-g/C)^2}{2(1-v/s)} + \frac{(v/c)^2}{2v(1-v/c)} - 0.65(c/v^2)^{1/3}(v/c)^{2+(5g/C)} \tag{4.32}$$

The first term in Eq. (4.32) is the same as that given in Eq. (4.28), the second term is the additional term which results from analyzing the process by assuming stochastic arrivals, the third term is a correction factor obtained from simulation studies. It is seen that this term is often between 5 and 15 per cent of the sum of the first two terms. Hence the following simplified form of the Eq. (4.32) is sometimes used.

$$D_{s,web} = 0.9\left(\frac{C(1-g/C)^2}{2(1-v/s)} + \frac{(v/c)^2}{2v(1-v/c)}\right)$$

In practice, it is found that the $D_{s,web}$ estimates of delay are not good for the entire range of v/c values. In general, when v/c values are close to one (that is, vC is close to gs, this implies an increase in the chances of oversaturation when the stochasticity in the arrival rate causes the number of arrivals to be greater than gs), $D_{s,web}$ overestimates the average delay. One possible reason for this is that the derivation of the quantity assumes steady-state behaviour, which is never achieved at real intersections as oversaturation by design occurs only in short spells. Other researchers around the world have developed other equations which are supposed to predict delays more realistically. All of them predict values which are close to $D_{s,web}$ when v/c is not high (say less than 0.8) while their estimates are much lower when v/c is high (say greater than 0.95). The 1985 *Highway Capacity Manual* of USA [104] proposes the use of one such expression for delay, $D_{s,hcm85}$, reproduced here as Eq. (4.33). This relation was developed based on a large database on intersection delay.

$$D_{s,hcm85} = 0.76\frac{C(1-g/C)^2}{2(1-v/s)} + 173(v/c)^2\left[(v/c-1) + \sqrt{(v/c-1)^2 + 16(v/c^2)}\right] \tag{4.33}$$

The 1985 HCM [104] cautions users against using this relation for $(v/c) > 1.2$. In the 1998 HCM [103] the calculation of delay has been further modified. That expression is not provided here as it uses many site specific empirical constants which are not valid for Indian conditions.

QUEUE ANALYSIS

An involved discussion on analysis of queues at signalized intersections is not possible in this text as it requires substantial knowledge of stochastic queueing processes. However, certain characteristics of the arrival and departure processes from the point of view of queueing analysis are presented here so that the interested reader may pursue this topic further.

The primary purpose of doing a queueing analysis at a signalized intersection is to be able to determine the probability distribution of queues that form. That is, the aim should be to answer questions like, what is the probability that there will be n vehicles in a particular queue at any time. This is important since not only values such as average queue length are important, but what is more important is an idea of the queue length distribution so that we can determine lengths of auxiliary lanes[1] (see Chapter 5 for details) for pre-designated values of probability of overflow (where the queue of vehicles is larger than the space provided by the auxiliary lane) and probability of blockage (where the queue of vehicles on the lane adjacent to the auxiliary lane is so large that it blocks the entrance to the auxiliary lane).

However, the theory of queues available so far is not able to model the queueing process at a signalized intersection in a simple manner. The primary reason for this is that the departure process is not a Poisson process. If we assume (and correctly so) the time a vehicle waits at the top of the queue to be its service time, then it can be easily seen that the service time duration (to which the departure process is integrally connected) is not negative exponential. In fact, the service time is either zero (if one reaches the top of the queue when the light is green) or equal to the duration of the red period (if one is the next to the last vehicle to have crossed the intersection); the probability of either of these service times being the service time of a particular vehicle is not easy to calculate. Matters can be further complicated if there is a permitted phase in the signal (as is sometimes the case for turning movements) where a vehicle can cross the intersection (during the non-green time) if a sufficient gap in the opposing stream exists.

Some researchers (see for example Kikuchi et al. [136]), however, have attempted to determine the probability distribution of queues through a Markov chain analysis of the queueing process. The interested reader may refer to Kikuchi et al.'s work cited above for a good understanding of the analysis procedure.

[1]An auxiliary lane is a limited length lane provided at intersections for turning vehicles.

EXAMPLE 4.7

On an approach to a signalized intersection, the effective green time and the effective red time are 30 s each. The arrival rate of vehicles on this approach is 360 vph between 0–120 s, 1800 vph between 120–240 s, and 0 vph between 240–420 s. The saturation flow rate for this approach is 1440 vphgpl. The approach under consideration has one lane. Assume that at time = 0 s the light for the approach has just turned red.

1. Plot the arrival rate of vehicles versus time.
2. Assuming the arrival and departure processes to be continuous, plot the cumulative number of arrivals and departures versus time.
3. Determine the average delay to vehicles arriving between 0–120 s.
4. Determine the average delay to vehicles arriving between 120–240 s.
5. Determine the average delay to vehicles arriving between 0–240 s.
6. Determine the delays to the fourth and the sixtieth vehicles that arrive at the intersection.
7. Determine the maximum delay faced by a vehicle on this approach.
8. Determine the maximum queue length on this approach. At what time does the queue length first become equal to the maximum?
9. Determine the percentage of time for which there exists a queue on this approach.
10. Determine the average queue length between 120 and 420 s.

Solution

The following parameter values are provided: g = 30 s, C = 30 + 30 = 60 s, s = 1440 vphgpl, v from 0–120 s = 360 vph, v from 120–240 s = 1800 vph, and v from 240–420 s = 0 vph, and T, the time for which there exists a flow higher than c, is 120 s. Note, $c = (g/C)s$ = 720 vph.

(i) Figure 4.29 gives the plot of arrival rate of vehicles versus time.

(ii) Figure 4.30 gives the plot of cumulative arrivals and departures versus time.

(iii) Between 0–120 s the intersection is operating under unsaturated conditions. Further, the arrival is deterministic and uniform. Hence we can use Eq. (4.28) to determine the average delay. Therefore,

$$D_{u,us} = \frac{C(1-g/C)^2}{2(1-v/s)} = \frac{60(1-30/60)^2}{2(1-360/1440)} = 10 \text{ s}$$

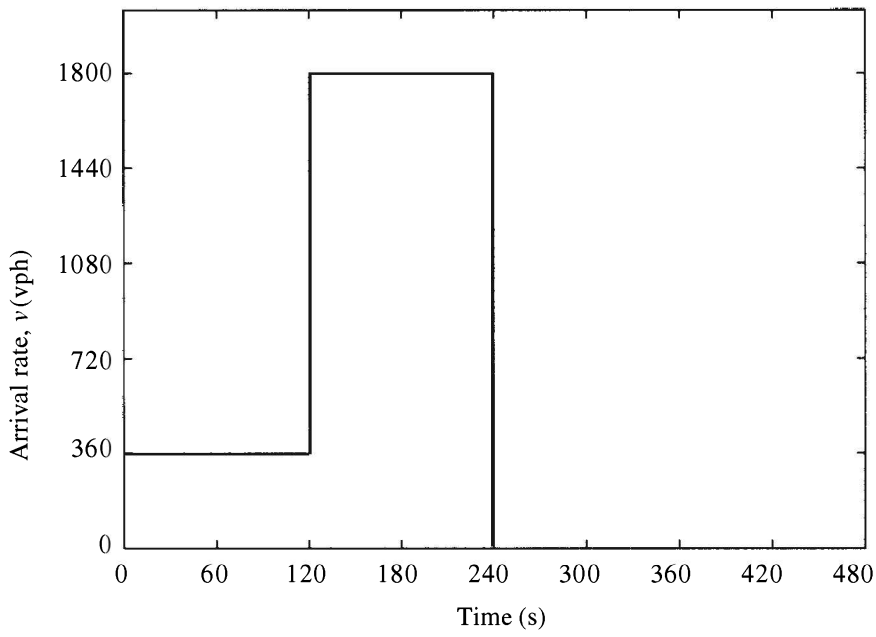


Figure 4.29 Example 4.7: Plot of arrival rate of vehicles versus time.

We can also determine the average delay directly from the graph, by noting that the area of either Triangle I or Triangle II in Figure 4.31 (which is the same as Figure 4.30 but with few extra annotations) divided by the total number of arrivals during a cycle will give the average delay. Therefore,

$$\text{Average delay} = \frac{\text{Area of Triangle I or II}}{\text{Number of arrivals in a cycle}}$$

$$= \frac{0.5 \times 30 \times 4}{6} = 10 \text{ s}$$

(iv) Between 120–240 s the intersection is operating under oversaturated conditions. The arrival is deterministic and uniform. Hence we can use Eq. (4.31) to determine the average delay. Thus,

$$D_{u,os} = \frac{T}{2}\left(\frac{v}{c} - 1\right) + \frac{C-g}{2} = \frac{120}{2}\left(\frac{1800}{720} - 1\right) + \frac{60-30}{2} = 105 \text{ s}$$

We can also determine the average delay directly from the graph (see Figure 4.31), by noting that,

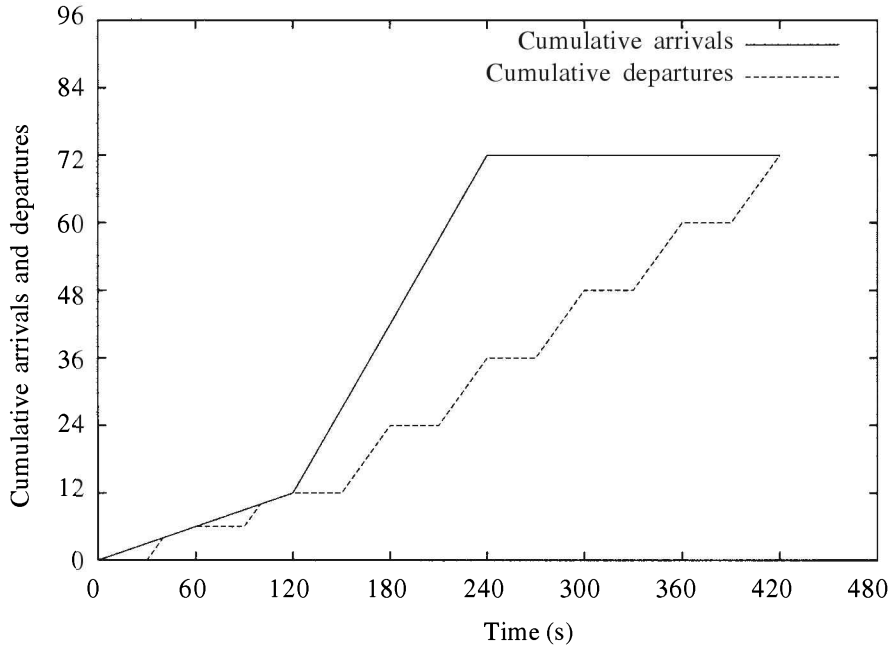


Figure 4.30 Example 4.7: Plot of cumulative number of arrivals and departures of vehicles versus time.

$$\text{Average delay} = \frac{\text{Area of Triangle III} + 5 \times \text{Area of Triangle IV}}{\text{Number of arrivals from } 120 - 240 \text{ s}}$$

$$= \frac{0.5 \times 180 \times 60 + 5 \times 0.5 \times 12 \times 30}{60} = 105 \text{ s}$$

(v) The average delay to all the vehicles between 0–240 s can be obtained by dividing the total delay (faced by all vehicles) by the total number of vehicles. Hence,

$$\text{Average delay} = \frac{n_1 d_1 + n_2 d_2}{n_1 + n_2}$$

where

n_1 is the number of vehicles that arrive during 0–120 s

d_1 is the average delay to a vehicle coming during 0–120 s

n_2 is the number of vehicles that arrive during 120–240 s

d_2 is the average delay to a vehicle arriving during 120–240 s.

Hence,

$$\text{Average delay} = \frac{12 \times 10 + 60 \times 150}{12 + 60} = 89.2 \text{ s}$$

Of course, we can also find the average delay here from the graph (the reader should do this).

(vi) The arrival rate of vehicles from 0–120 s is 360 vph or 0.1 vps. Assuming that the fourth vehicle arrives before the expiry of 120 s, the time of arrival of the fourth vehicle is 4/0.1 = 40 s. (Hence the assumption is not violated.)

The departure rate of vehicles is 1440/3600 = 0.4 vps. The time of departure of the fourth vehicle, assuming that the fourth vehicle gets discharged during the first green, is 30 + 4/0.4 = 40 s. (Since the departure time is less than the start of the next red, the assumption is valid.)

The delay to the fourth vehicle therefore is

$$\text{departure time} - \text{arrival time} = 40 - 40 = 0 \text{ s}$$

The same observation can be made from Figure 4.31.

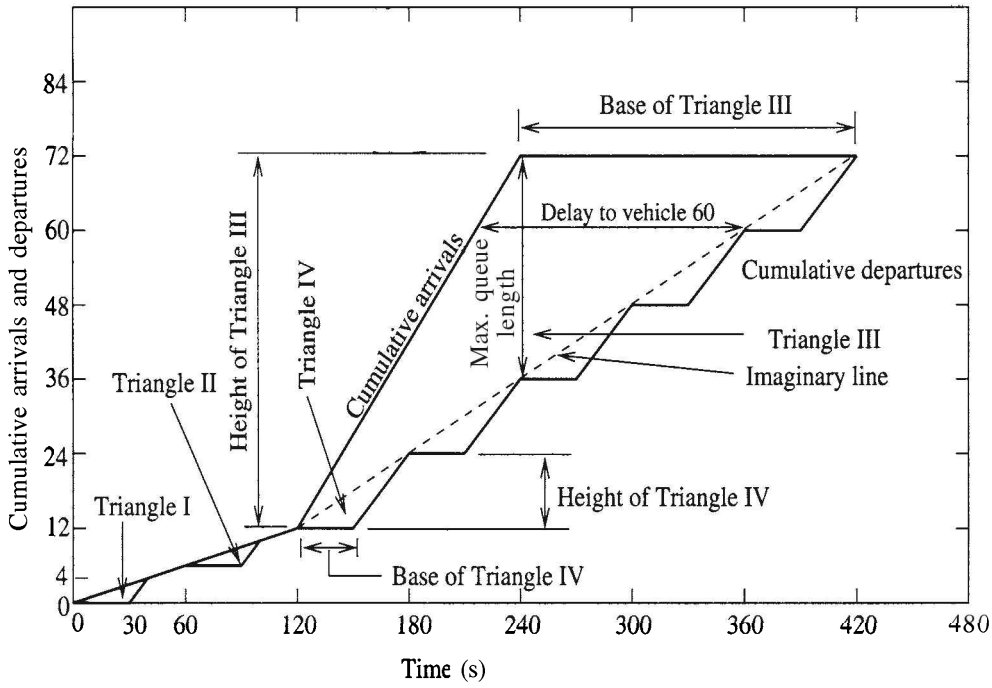


Figure 4.31 Example 4.7: Plot of cumulative number of arrivals and departures of vehicles versus time.

The delay to the sixtieth vehicle can also be read from Figure 4.31 as 144 seconds.

(vii) As can be seen from Figure 4.31 the maximum delay is 180 seconds.

(viii) As can be seen from Figure 4.31 the maximum queue length is 36 vehicles. At time = 240 s, the queue length first becomes equal to 36 vehicles.

(ix) As can be seen from Figure 4.31 there are no queues from 40–60 s and from 100–120 s. For the rest of the time, there is a queue at the intersection. Hence, the percentage of time for which there is no queue at the intersection is (40/420)100 = 9.52%. Hence the percentage of time when there exists a queue is 100 – 9.52 = 90.48%.

(x) We can determine the average queue length directly from Figure 4.31, by noting that

$$\text{Average queue length} = \frac{\text{Area of Triangle III} + 5 \times \text{Area of Triangle IV}}{\text{Total time from 120–240 s}}$$

$$= \frac{0.5 \times 180 \times 60 + 5 \times 0.5 \times 12 \times 30}{300} = 21 \text{ vehicles}$$

Data collection

At signalized intersections, other than collecting data on arrival rate and pattern (which can be done in the usual manner of counting volume and recording time headways), data on delay and saturation flow rates may need to be collected. In this section, procedures for collecting data on delay and saturation flow rates are described.

COLLECTING DATA ON AVERAGE DELAY

In this section a procedure which can be easily used to collect data on stopped delay at a signalized intersection is described. The procedure relies on the principle that the area between the cumulative arrival and departure plots (see Figure 4.26) gives the total time that all the vehicles spend stopped at the intersection; its unit is *vehicle seconds*. This area divided by the total number of arrivals obviously gives the average delay.

The area can be obtained by summing either all the delays or all the queue lengths, that is,

$$\text{Total area between cumulative arrivals and departure plots} = \sum_n d(n) = \int_t q(t)dt$$

where all the variables are as explained in Figure 4.26.

Hence the average delay can be obtained as

$$\text{Average delay} = \frac{\int_t q(t)dt}{\text{Total number of arrivals}}$$

The data collection procedure uses the above relation to evaluate the average delay. The method relies on determining the area by observing the queue lengths at short intervals of time over the entire experiment time period, and counting the total number of vehicles that arrive during the entire test period. The procedure is explained step-by-step as follows:

Step 1. Decide the time period P for which the data will be collected. Decide the interval of time I at which, the queue length at the intersection will be counted. The cycle length C should not be an integral multiple of I. Let the number of intervals in P be m.

Step 2. Count the queue length q_i at the end of each interval. Continue counting till P is over. Also over the time P, count the total number of vehicles that arrive at the intersection, i.e. V_{total}.

Step 3. Estimate the area as $\sum_{i=1}^{m}(I \times q_i)$. See Figure 4.32, which is a figure similar to Figure 4.26. The area obtained using $\sum_{i=1}^{m}(I \times q_i)$ is generally seen to be more than the actual area, and hence the estimate obtained here is reduced by 10% in order to get a closer estimate of the true area.

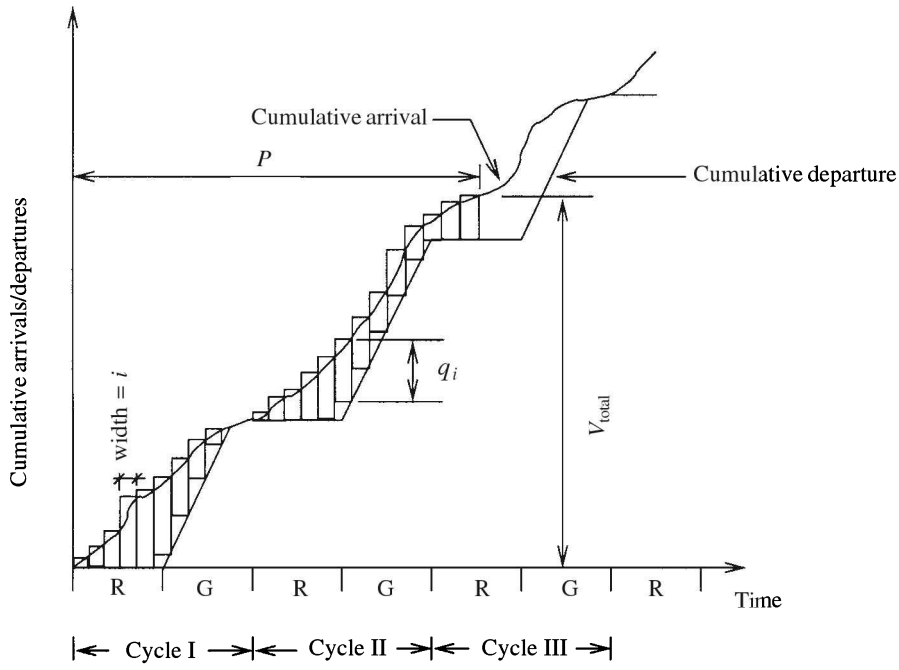


Figure 4.32 Plot of cumulative number of arrivals and departures versus time showing the area obtained by numerically integrating the queue lengths.

Step 4. Estimate the average delay as

$$\text{Average delay} = \frac{0.9 \times I \sum_{i=1}^{m} q_i}{V_{\text{total}}}$$

COLLECTING DATA ON SATURATION FLOW RATE

The saturation flow rate is the reciprocal of the saturation headway. The saturation headway, as suggested in the subsection on DEPARTURE PROCESS, is the headway at which latter vehicles discharging from a queue cross the stop line. In general, the HCM [103] suggests that the average value of the saturation headway s_i for sample i can be obtained using

$$s_i = \frac{T_{L,i} - T_{4,i}}{L - 4}$$

where $T_{j,i}$ is the time at which the jth vehicle of the queue crosses the stop line for the ith sample, and L stands for the last vehicle in the queue. The assumption here is that from the fourth vehicle onwards all vehicles more or less maintain the saturation headway. According to the HCM [103], the average value of the saturation headway should be estimated as the mean of all s_is after repeated sampling.

Capacity and level-of-service analysis of signalized intersections

In this subsection the capacity and level-of-service analyses of signalized intersections are described. The description presents the principles on which the analyses are based and does not go into the specifics. This is because the Indian codes' (see IRC Special Publication 41 [86] and IRC:93–1985 [93]) description of the process is somewhat incomplete and in any case the principles of the analysis process are more important than the exact details of the process. The section is divided into two parts, the first part discusses the capacity analysis and the second part the level-of-service analysis.

CAPACITY ANALYSIS

The capacity of signalized intersections as such is not meaningful. What is meaningful is the capacity of an approach or a lane or a lane group[2] of an intersection. The capacity c_i of lane i to a signalized intersection, is defined as

$$c_i = s_i \times \frac{g_i}{C}$$

where

s_i is the saturation flow on lane i in vehicles per hour of green per lane

g_i is the green time for the lane

C is the cycle length at the intersection.

[2]A lane group is a group of lanes on an approach which have vehicles of the same movements; for example there could be two lanes only for through traffic, in this case these two lanes could be grouped into a single lane group.

Thus, the primary factor which needs to be determined is the saturation flow for the lane. Typically, saturation flow on a lane or a lane group depends on a number of factors such as (i) the number of lanes in the lane group and width of lanes or alternatively the width of the lane group, (ii) the gradient of the lane, (iii) percentage of turning traffic, (iv) vehicle mix, (v) the number of parking manoeuvres, and (vi) the number of bus stoppings. The HCM [103] provides a detailed and complete overview of how the effect of these factors can be and should be incorporated in the calculation of saturation flows. The manual assumes a set of ideal conditions and also assumes an ideal saturation flow for these conditions. This saturation flow is then reduced by multiplying it with a series of correction factors. Generally, each correction factor quantifies the detrimental effects of the non-ideal conditions with respect to each of the factors mentioned above. These correction factors are provided in a series of easy-to-follow tables.

The IRC Special Publication 41 [86] suggestion on determining the saturation flow does not incorporate all the features that affect the quantity. For example, though the code acknowledges the effect of vehicle mix on saturation flow it does not give any procedure to quantify this effect. Similar is the case for turning traffic's effect on the saturation flow of a lane or lane group (although the code attempts at quantifying the effect of right-turning traffic, the procedure is poorly explained and never illustrated). However, the code makes clear suggestions on the effect of width and gradient on saturation flow. For example, the code states that the saturation flow from a lane group of width w (for $w \geq 5.5$ m) is $525 \times w$ passenger cars per hour of green (pcphg); a table is provided for smaller values of w. The code also states that this value of saturation flow is for flat roads and should be reduced by 3% for every 1% uphill slope and increased by 3% for every 1% downhill slope. The reader may refer to [86] for more details on these aspects.

LEVEL-OF-SERVICE ANALYSIS

The level-of-service of different lanes and lane groups at signalized intersections should be determined through a measure which directly gives the level of discomfort (or comfort) of drivers using these lane or lane groups at the intersection. One such measure is the average delay to vehicles of different lanes and lane groups. There are various equations, each with certain shortcomings, which can give the average delay (see the subsection on DELAY ANALYSIS provided earlier). Surprisingly, the IRC codes are silent on the matter of level-of-service at signalized intersections. However, other codes like the HCM [103] of USA does provide a relation between delay to vehicles and level-of-service.

4.4.3 Traffic Flow at Unsignalized Intersections

An unsignalized intersection functions quite differently from a signalized intersection. While in the signalized intersection the common intersecting space is 'time shared,' at the unsignalized intersection the sharing is a lot more complex. The characteristics of the

flow are described in detail in the following subsection. Later sections analyze the queues formed at unsignalized intersections and the capacity and level-of-service of unsignalized intersections.

Flow characteristics

Traffic flow of a movement at an unsignalized intersection is guided by the hierarchical position of the movement specified either tacitly (by rules of driving) or decreed (through static signs, such as 'STOP' or 'YIELD'). At any unsignalized intersection there are various types of movements, like (i) through movement on major street, (ii) right turn movement from major street, (iii) left turn movement from major street, (iv) through movement on minor street, and so forth. Each of these movements has a place in the hierarchy specifying their claim on the right-of-way at the common intersecting space. For example, in general, first in the hierarchy is the through movement on the major street and slightly lower down is the right turn from major street. Now if in a situation there is a vehicle on the right turn movement and another on the conflicting through movement, then the latter will use the intersection and the former has to wait till the latter clears the intersection. If some movement is still lower down the hierarchy (like the right turn from minor street) then a vehicle on that movement has to wait till the vehicle on movement higher up in the hierarchy has cleared the intersection. As can be seen, the departure process is purely stochastic and extremely complex to model. A detailed discussion on the arrival and departure processes is provided in the following subsections.

However, before going into the discussion on arrival and departure processes it must be understood that unsignalized intersections work very efficiently if the total conflicting volume is not very high. For example, if at the intersection of a major street with a minor street, the traffic to and from the minor street is low then the intersection works quite well irrespective of the volume on the major street. If, however, conflicting movements have reasonable volumes then unsignalized intersections become inefficient and tend to cause large delays to the low priority (i.e. lower in the hierarchy) movements. This is when signalization becomes imperative. In Chapter 5, some conditions which justify or warrant signalization are described.

ARRIVAL PROCESS

The arrival processes of vehicles obviously do not depend on the type of intersection at which they arrive. Hence they are like those at the signalized intersections and no separate discussion is therefore provided here.

DEPARTURE PROCESS

The departure processes from unsignalized intersections are quite different from those at the signalized intersections. The departure process of a movement is determined by the hierarchical position of the movement and the type of control ('STOP' or 'YIELD')

on the movement. If a movement is at the top of the hierarchy and is not controlled (or 'YIELD' controlled) then vehicles on the movement always have right of way at the intersection and their flow is not interrupted. However, for the majority of the movements, their position is not at the top of the hierarchy and are often 'STOP' controlled. For these movements, the departure process is quite complex and is therefore explained here through an example.

Consider the situation shown in Figure 4.33. In this situation, two through vehicles (marked T1 and T2) and two right-turning vehicles (marked R1 and R2) on the left-to-right stream are shown. Numerous vehicles on the right-to-left stream are also shown. Further consider the arrival and departure processes for the left-to-right stream at a proper location on the road (say the stop line). Vehicles arrive at this point as has been described earlier. The departure process for the two types of vehicles shown, however, is different. The through vehicles represent those vehicles which always have the right of way. Hence for these vehicles the arrival time at the stop line is always equal to the departure time from the stop line. The right turning vehicles represent those vehicles which are lower in the hierarchy and have to wait for gaps in the opposing stream to complete their manoeuvre. For example, vehicle R1 waits at the stop line and evaluates each of the gaps in the opposing stream. Only when a gap is greater than some value (at which the driver is comfortable) does the driver of the vehicle accept the gap and make the right turn. In the figure shown this could be Gap III. Hence, vehicles which have to look for gaps in the opposing stream (or streams), sometimes have to wait at the stop line before departing. Since the arrival of gaps is a stochastic process, the departure process of vehicles (or the waiting time at the stop line), is also a stochastic process.

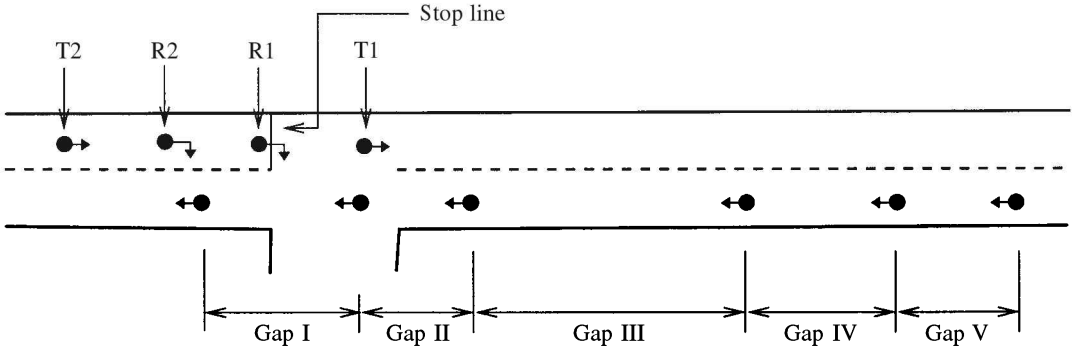


Figure 4.33 A snapshot of an unsignalized T intersection.

Once the vehicle at the stop line departs, all those vehicles waiting behind it move up. In case it is a single-lane (in each direction) road, the vehicles waiting behind could be of any kind (T or R in this case). If it is a through vehicle then as soon as it reaches the stop line, it departs. If, on the other hand, it is a right turning vehicle (or any other kind lower in the hierarchy) then the vehicle again may have to wait at the stop line for an adequate gap.

Before leaving this section, a brief discussion on this concept of adequate gap is necessary. The minimum value of the adequate gap is referred to as the *critical gap*. In the deterministic view of things, a driver accepts a gap whenever the gap is greater than the critical gap and rejects it whenever the gap is less than the critical value. In reality, however, this is not true and the critical gap is only an idealization of the observation that drivers tend not to choose gaps which are 'small' and choose gaps which are 'large.' The analysis of unsignalized intersections, nonetheless, assumes that a critical gap exists.

Delay and queue analysis

Unlike in the signalized intersection, the delay and queue analysis is taken up here together. From a queueing theory standpoint, the flow on any of the approaches at an unsignalized intersection must be viewed as having the following characteristics:

- Various types of vehicles arrive at the intersection; these vehicles differ from one another in the time they spend at the stop line (henceforth referred to as *service time*). For vehicles which are at the top of the hierarchy, service time is nil and deterministic. For all the other types of vehicles the service time is stochastic and follow different distributions.
- The queue formed at the stop line has a 'first-in-first-out' queue discipline. The queue may contain more than one type of vehicle depending on the number of lanes on the approach under consideration. For example, if there are separate turning and through lanes then the queue will only contain turning vehicles as through vehicles do not queue. If, however, the approach has a lane which is shared by many types of vehicles then the queue has all the types of vehicles.

In analyzing any stochastic queueing system we need to determine the arrival distribution and the service time distribution. In this case the arrival distribution may be assumed to be Poisson (if the intersection is away from any other source of interruption) or deterministic (if the intersection receives vehicles released by some other nearby signalized intersection) or a combination of the two. The service time distribution, however, needs to be determined given the modalities of the departure process.

The service time distribution will depend on the number of gaps that a vehicle rejects before accepting a gap and on the distribution of gaps themselves. Such an analysis is far beyond the scope of this book. The interested reader may refer to Drew [53] for a good development of this topic.

Even more complex is the analysis of the queue distribution and the delay to vehicles. The complexity arises primarily due to the complex and different service time distributions of the various types of vehicles in the queue. Some idea of the process of analysis may be obtained from Chakroborty et al. [35].

Indian codes do not provide any relation which can be used to determine delay at unsignalized intersections under Indian traffic conditions. The 1998 HCM [103] does provide relations to determine queue lengths and delay, but these are based on highly

empirical considerations and are not provided here as Indian conditions at unsignalized intersections are quite different from those under which these equations were developed. Although, the relevant relations are not available, the factors on which queue length and delay at unsignalized intersections depend are described below. In this description, the delay and queue of the movement being studied is referred to as TM.

Conflicting volume. The volume of traffic in which the vehicles of TM look for gaps affects the queue length and delay. This is so, because, as the conflicting volume increases the number of adequate gaps decrease; this then increases the service times and hence the delay and the queue lengths.

Movement type. Vehicles in movements which are lower in the hierarchy, generally have to wait longer than vehicles of movements which are higher in the hierarchy. The reason for this is that often vehicles of low priority movements cannot accept an adequate gap because there is a vehicle of the higher priority also waiting and will use that gap. For example, consider a situation where there are two vehicles waiting; one belongs to the through movement from the minor stream (travelling North) while the other belongs to the right turning movement from the major (westbound) stream (a movement which is higher on the hierarchy than the former type of movement). Now a gap arrives in the eastbound major stream which could have been used by either of the vehicles; in such a situation the vehicle on the through movement from the minor stream will have to wait while the right turning vehicle uses the gap.

Critical gap. As the critical gap increases the number of acceptable gaps in the conflicting movements reduce. This again increases the service times and hence the delay and the queue lengths.

Arrival rate. As the arrival rate of vehicles in the TM increase, the queue lengths increase and hence the delay also increases.

Speed. It is seen that as the speed of conflicting streams increases, the critical gap for drivers in the TM also increases. (This is possible because, with increased speeds drivers want to be very sure before accepting gaps.) Increases in critical gaps have the effects described earlier.

4.4.4 Data Collection

Two types of data are generally collected at unsignalized intersections, namely the data on (i) average delay and (ii) critical gap. The data on average delay may be collected in the same way as in signalized intersections, because the approximation made there (see Figure 4.32) is also valid here except for the fact that the cumulative departure line does not follow any fixed pattern (since the departure process is also stochastic).

Data on critical gap is slightly more difficult to obtain in the field. The reason for this is that a single driver may reject a lot of gaps but always accepts only one gap. From

this fact we can only say that the critical gap for the person is greater than the largest rejected gap and smaller than the accepted gap. Hence such a method of data collection will only yield ranges, being estimates of the critical gaps for different drivers. Further these ranges could be quite large making a reasonably precise estimate of the critical gap almost impossible.

Therefore, we have to often assume that the drivers coming at the intersection exhibit reasonably similar approach towards their gap acceptance behaviour in order to obtain a large number of rejected and accepted gaps. A simple determination of the critical gap from the largest rejected and the smallest accepted gaps is not feasible as often the former may be greater than the latter due to differences in gap acceptance behaviour of different drivers. Drew [53] suggests the following procedure to determine the critical gap.

- Observe an unsignalized intersection and obtain data on many drivers in respect of the largest gap rejected and the smallest gap accepted.
- Divide the time scale into small intervals (say of 0.5 s duration) and determine for every time class t (say, the gap size between 2 and 2.5 s) the number of such gaps accepted and the number of those rejected.
- From this, calculate the number of gaps less than t which are accepted and the number of gaps greater than t which are rejected.
- Plot the cumulative curves on the same graph of 'number of gaps' versus 't'.
- Report the value of t where these plots intersect as the critical gap. This value of t represents that gap size for which the number of gaps greater than t which have been rejected are equal to the number of gaps smaller than t which have been accepted.

Capacity and level-of-service analysis

In this section, the principles of capacity and level-of-service analyses for unsignalized intersections are described. The description does not go into the specifics and the reader can refer to IRC Special Publication 41[86] for the specific details. Another reason for not going into the specifics is that none of the methodologies suggested by the IRC have been developed taking into account the flow characteristics of Indian unsignalized intersections (a fact acknowledged by the IRC in [86]).

This subsection is divided into two parts. The first part describes the capacity analysis and the second the level-of-service analysis.

CAPACITY ANALYSIS

As in the case of signalized intersections, it is meaningful only to talk about capacities of different movements, lanes, or approaches to the unsignalized intersection, and not about the intersection as a whole. Given the earlier description of the flow characteristics at such intersections, it is evident that the capacity of a movement at an unsignalized intersection depends on the following factors:

Gap availability. Any traffic stream at an unsignalized intersection moves by accepting gaps in the conflicting streams. Obviously then, the more the number of conflicting streams the lower is the gap availability; similarly, the higher the volume in the conflicting streams the lower is the gap availability. Further, the greater the size of the critical gap, the smaller are the number of acceptable gaps. Therefore, the total number of gaps available for use by a traffic stream depends on the number and volume of conflicting streams and the size of the critical gap. Hence, by implication the capacity of a movement is also affected by the same factors.

Hierarchical position. Not all gaps available for use can, however, be used by the vehicles of a movement. As stated earlier, there exists a hierarchy of movements at any unsignalized intersection. The movements which are higher in this hierarchy get a preference over the movements lower in the hierarchy when it comes to actually using an available gap. In effect, the lower the priority of a movement the more it is impeded from using a gap because the chances of higher priority movements wanting to use the same gap are more. Hence, even if a particular movement has a lot of available gaps it may get to use only a few of them because of its poistion in the hierarchy.

Gap accessibility. Sometimes, at unsignalized intersections, a lane may be shared by more than one movement. In such cases, it may often happen that a particular gap of interest to a given movement never gets used because the vehicle of that movement is caught behind vehicles of other movements (to whom the gap may not be of interest) sharing the same lane. That is, even if a gap is available it may not be used because the vehicle which could have used it had other vehicles (from other movements) ahead of it in the queue. This situation can occur more frequently if the number of movements sharing the lane is more.

This understanding of the factors which affect the capacity of a movement, is formalized for capacity calculations in the following manner.

Step 1. The movement whose capacity is to be determined is identified (this is referred to as the test movement, TM). Given the TM and the intersection geometry, all conflicting streams are identified.

Step 2. The volumes of the conflicting streams are added. The sum, referred to as the conflicting volume, and the critical gap size, are then used to determine the potential capacity $c_{p,i}$ for the movement i (in this case, i is TM) from a set of empirically deterimined curves. This capacity is basically a statement of the total number of available gaps for TM.

Step 3. Given the hierarchical position of TM, all streams which are higher in the hierarchy (and compete with TM for gaps) are identified. For each such higher priority movement, an impedance factor is determined. The impedance factor from a particular movement, in a way, states the number of times that movement uses a gap which TM could also have used. The impedance factor from a given higher priority movement,

j, is determined from an empirically derived graph which plots impedance factor versus the $v_j/c_{p,j}$, where v_j is the volume on the movement j.

Step 4. Once all the impedance factors, P_js, have been determined, the movement capacity of movement i (here i is TM) is determined as

$$c_{m,i} = c_{p,i} \times \prod_{\forall j \in J} P_j$$

where J is the set of all movements higher in the hierarchy than movement i and competes for gaps with movement i. The movement capacity, therefore, is a statement of the number of gaps which TM can possibly use given its position in the hierarchy.

Step 5. As stated earlier, sometimes different movements share a lane. In such cases, the movement capacities are not very meaningful as all the gaps counted in the determination of these values may not be used due to some of them not being accessible. The capacity of shared lanes c_{sh} is given as

$$c_{sh} = \frac{\sum_{\forall i \in I} v_i}{\sum_{\forall i \in I} (v_i / c_{m,i})}$$

where I is the set of all movements which share the lane. If there is only one movement which uses the lane, then the shared lane capacity is equal to the movement capacity.

In this case, the volume is in vehicles per hour (per lane) when calculating the conflicting volume and in passenger cars per hour (per lane) in all other calculations. The translation of vehicles to equivalent number of passenger cars can be obtained using the passenger car equivalence factors given in the code [86]. All capacities are also in passenger cars per hour (per lane). The critical gap sizes may also be obtained from [86].

LEVEL-OF-SERVICE ANALYSIS

The level-of-service of different movements at unsignalized intersections, like in the case of signalized intersections, should be determined through a measure which directly gives the level of discomfort (or comfort) of drivers using the intersection. One such measure is the average delay to vehicles using the intersection. Although the recent HCM [103] suggests the use of delay, the IRC still follows the old HCM [104] method of using the concept of *reserve capacity* to determine the level-of-service. Reserve capacity for a movement (if a lane has only one movement) or a lane (if a lane has more than one movement) is obtained by subtracting the volume in the movement (or the lane) from the movement capacity (or the shared lane capacity).

It is envisaged that the higher the reserve capacity the lower will be the average delay and therefore better will be the level-of-service. The IRC Special Publication 41 [86] provides a table which relates reserve capacity to level-of-service.

EXERCISES

1. What are the parameters which characterize flow? How are they related?

2. Calculate the PHF_t and the corresponding peak flow rate for t = 1, 5, 10, and 15 minutes for the minute flow rate data given in the following table. Comment on the difference in the peak flow rates obtained for the different values of t.

Minute i	1–7	8–10	11–20	21	22–32	33–45	46–47	48–55	56–60
N_i	15	25	20	36	9	5	10	25	15

3. Show that the higher the value of the peak hour factor the more uniform is the flow within the hour.

4. Show that $0.25 \leq PHF_{15} \leq 1$.

5. Show that $a = u_o$ and $b = k_j$ if $u = a \ln (b/k)$.

6. If $u = u_o \ln (k_j/k)$, what are the implied u–q and q–k relations?

7. If the u–k relation is modelled through the Greenshields' equation, then show that the maximum flow q_{max} is one-quarter of $u_f \times k_j$.

8. For the following u–k data, use the ordinary least squares technique to determine the parameter values of (a) the Underwood's model and (b) the Greenberg's model.

u, kmph	10	20	30	30	40	50	60	60	70	80	90	100	100
k, vpkm	110	100	80	70	70	60	50	45	40	25	20	10	5

9. For the u–k data given in Exercises, determine u_f and k_j values using the generalized polynomial model with various assumed values of a and d [see Eq. (4.13)]. Compare the fits and suggest the best combination of a and d. Use the ordinary least squares technique for fitting the equations.

10. Given the following q–k data, determine the parameters of the Greenshields' model by fitting (i) the implied q–k model and (ii) the u–k model to the implied u–k data. Is there any mismatch in the estimates? If so, give reasons.

u, kmph	10	20	30	40	50	60	80	90	100
q, vph	1100	2000	2400	2800	3000	2700	2000	1800	500

11. Show that with $m = 0$ and $\ell = 2$, the GM model of car-following implies a linear u–k relation.

12. Derive the generalized polynomial model of u–k relation from the GM model of car-following.

13. In a car-following situation the LV (leading vehicle) decelerates at 2 m/s^2 for 2 s, then decelerates at 3 m/s^2 for another 2 s. After this the LV accelerates at 2 m/s^2 for 5 s. Simulate the behaviour of the following vehicle (FV). For all other relevant parameters use the data given in Example 4.5.

14. For the situation given in Example 4.5 on car-following, simulate the asymptotic behaviour of a platoon of six vehicles. All the vehicles in the platoon are initially moving at 16 m/s and at distance headways of 28 m.

15. For the situation given in Example 4.5 on car-following, simulate the behaviour of a pair of vehicles for increasing values of $\alpha_{1,0}$. Study how the local stability behaviour changes.

16. On a road, the speed–density relationship is $u = 70 - 0.7k$. The traffic on the road was moving at a speed of 40 kmph when it was stopped by a flag-person (say, at time t_0). Movement on the stream was stopped for 5 minutes. At time $(t_0 + 5)$ minutes the vehicles were released from their stopped condition and started moving at 30 kmph. For the above conditions:

(a) Draw a neat schematic diagram for the distance–time plots of the vehicles moving on this road.

(b) Identify all the shock waves on the diagram that are created on this road due to the interruption caused by the flag-person.

(c) Determine the speeds of all the shock waves.

(d) Locate the starting point of the platoon forming shock wave.

(e) Locate the starting point of the platoon dissipating shock wave.

(f) Locate the end points of the platoons in (d) and (e) above.

(g) Determine the maximum length of the platoon.

(h) Determine the time it takes for the platoon to dissipate.

(i) Plot the location of the front of the platoon and the rear of the platoon versus time.

(j) Plot the length of the platoon versus time.

17. Plot the average delay to vehicles at signalized intersections as predicted by Eq. (4.32) and Eq. (4.33) versus v/c ratio. Compare the predictions and comment on the results.

Design of Traffic Facilities

5.1 INTRODUCTION

This chapter presents in a concise manner the principles of design of certain traffic facilities. The term *traffic facilities* means any transportation feature or structure that facilitates the movement and storage of roadway vehicles such as automobiles, two-wheelers, buses, etc. Traffic facilities by this definition, therefore, include (i) various types of roadways such as freeways (expressways), arterials, collector roads, and local roads, (ii) special roadway features such as signalized intersections, unsignalized intersections, grade separated intersections or interchanges, (iii) vehicle storage facilities such as parking lots and road-side parking spaces, and (iv) miscellaneous aids such as direction signs, street lighting, etc. Further, the principles of design discussed in this chapter exclude the principles of geometric design which were covered in Chapter 3.

It is a difficult task for a textbook of this nature to cover all the different traffic facilities and all the aspects of designing such facilities. Hence, the chapter concentrates on a few traffic facilities (which are varied in their characteristics) and attempts to present only the principles of designing such facilities. Details of designing these facilities may be found in the various design manuals published by various transportation organizations such as the Indian Roads Congress, Transport Research Laboratory (of the UK), AASHTO (of the USA), various state departments of transportation in the USA, etc. Another document which is very useful in this regard is the *Highway Capacity Manual* [103]. It may be noted here that although the design specifications vary from one place to another, the basic principles of designing the facilities remain more or less the same throughout the world. It is these principle that are illustrated in the book and not the design specifications which rightly form the subject matter of design codes.

The traffic facilities which are discussed here have been chosen based on two factors, namely (i) these facilities form important constituents of any traffic and transportation system, and/or (ii) expose the reader to a variety of design principles used in traffic facilities design. The facilities discussed here are:

- Freeways (or Expressways)

- Intersections
  - Unsignalized intersections
  - Signalized intersections
- Grade-separated interchanges
- Parking facilities
- Street signs

The following sections discuss each of the above facilities.

5.2 FREEWAYS (OR EXPRESSWAYS)

The simplest way of defining a freeway (or expressway) is that it is a road-section of substantial length with little or no cross-traffic, vehicular or otherwise. For such roadway-sections, the only consideration in their design is the capacity they are supposed to provide. The capacity obtained from a freeway-section is dependent on the following factors:

- Projected or estimated peak demand (for example, peak hour volume or peak flow rate) for the road
- The design level-of-service
- The number of lanes
- The width of the lanes
- Shoulder width and quality
- Curvature of the road
- Gradient of the road

The projected demand and design level-of-service are the inputs to the design process. The design variables are, the number of lanes, the width of lanes, and the quality and width of shoulders. Often, the curvature and the gradient of the road are determined from other considerations such as sight-distance requirements and the topography of the area where the road is being built. For the purposes of this chapter, it suffices to say that determining a road layout (which includes horizontal curves and gradients) is a complex task and depends on a variety of factors like (i) natural lay of the land, (ii) cut-and-fill costs, (iii) presence of natural and man-made obstructions, and (iv) environmental impact of building a road. It is felt that this topic is better left for an advanced text when the reader has a good understanding of transportation engineering. In this section, focus is therefore on determining the number of lanes, the width of lanes, and the quality and width of shoulders.

Determining these design variables is a simple matter since the only objective is to select these variables such that the design level-of-service can be maintained for the projected or estimated demand for the road. Generally, the information on service flow rate[1] (or volume) for roads with different characteristics is derived from empirical observations and included in the manuals or codes published by different organizations. For example, the relevant information for Indian conditions may be found in IRC:64–1990 [83] and IRC:106–1990 [84]; similarly for US conditions, the *Highway Capacity Manual* [103] houses the relevant information. It may be noted here that the Indian specifications in traffic-related areas are not always complete in all respects. The reader is, therefore, advised to become familiar with the *Highway Capacity Manual* since it is a very comprehensive document and even though the specific details may not be the same for Indian conditions, the manual is in itself a good reference in the area of capacity and level-of-service analysis.

The following example illustrates how we may determine the design variables for freeway-sections.

EXAMPLE 5.1

An existing rural road going through a plain terrain with hardly any curves and bends is to be upgraded so as to form a part of an expressway. The annual average daily traffic (AADT)[2] on this section is expected to be 27,000 vehicles per day with the following mix: motorized two-wheelers (18%), passenger cars (35%), light commercial vehicles and tractors (30%), trucks and buses (17%). Design the freeway-section such that level-of-service (LOS) B can be maintained.

Solution

Whenever the flow consists of a variety of vehicles, it is only logical that they be converted to a single vehicle type (generally a passenger car) for the purposes of uniformity and ease of computation (otherwise we would have to list capacity and service flow values for every type of vehicle mix—which is not feasible). The conversion is done using passenger car equivalence factors obtained from empirical data on vehicle size and performance. These factors, for Indian conditions, are provided in IRC codes (for example, see IRC:64–1990 [83]).

Based on Table 1 of IRC:64–1990 [83], the AADT in passenger car units is

$$(0.5 \times 0.18 + 0.35 \times 1.0 + 0.3 \times 1.5 + 0.17 \times 3.0) \times 27{,}000 = 37800$$

passenger cars per day.

Based on Section 11 of IRC:64-1990, a four-lane (lane width of 3.75 m) divided expressway with 1.5 m wide hard shoulders has a service volume of 40,000 passenger

---

[1]This is the maximum allowable flow for a given level-of-service.

[2]This is the average daily traffic obtained over a year.

cars per day for LOS B. Hence, such a design will be able to provide a LOS B for the given demand. Therefore, the design is four-lane divided expressway, lane width of 3.75 m and 1.5 m wide hard shoulders.

Note that if we had to design the road for a level-of-service other than LOS B, we would not be able to do so since the IRC does not suggest the service flows for each LOS. Further, the IRC does not even indicate how the service volume or capacity would change if lane width, shoulder width, shoulder quality, etc. were to be changed.

5.3 INTERSECTIONS

Whenever two or more roads meet there is an intersection where different flows (vehicles moving in different directions) compete for the use of the same physical space. If no control over these flows exists, various points of conflict are created in the common area. These points of conflict are potential causes of accidents. These points are not only issues of safety concerns but their presence also adversely effects the movement of traffic as drivers become more circumspect and careful—which generally causes a drop in speed. Figure 5.1 shows a typical four-legged intersection and illustrates the concepts of conflict points and conflict zones.

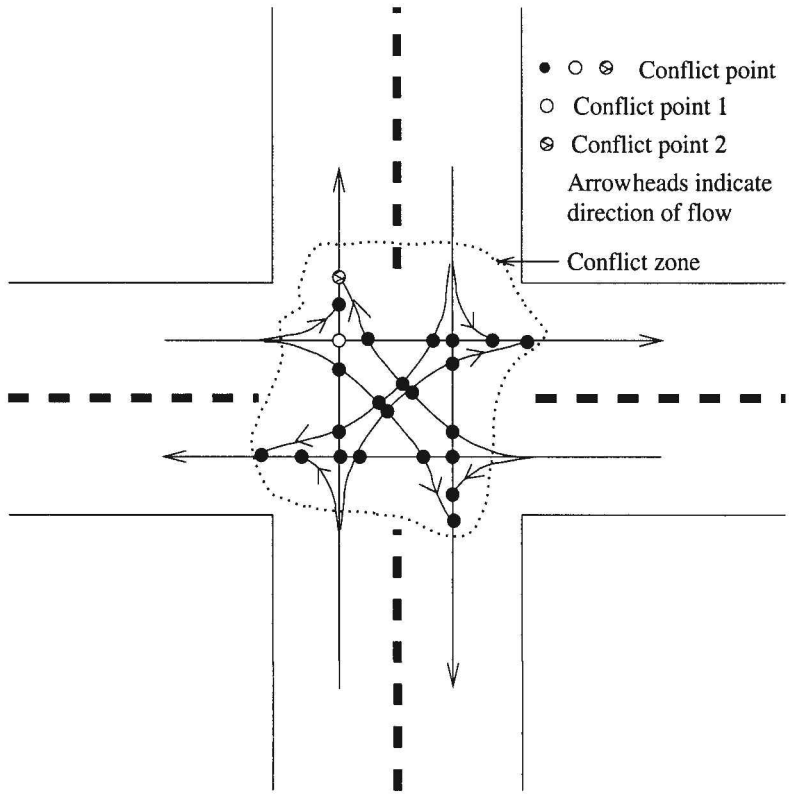


Figure 5.1 Conflict points and conflict zone.

The purpose of the intersection design is primarily to modify the zone of conflict (by either removing it or reducing it to a large extent) with the twin objectives of improving the safety and efficiency (in terms of delay to vehicles) of the intersection. Simplistically, the basic objectives of intersection design are achieved either by using static signs like **Stop** and **Yield** (or **Give Way**), or by using dynamic signs like **traffic signals** or by separating the levels (**grade separation**) of the intersecting roads. Intersections which primarily use static signs are referred to as **unsignalized intersections**, traffic signals are referred to as **signalized intersections**, and grade separations are called **interchanges**. In the following subsections, design aspects of unsignalized and signalized intersections are discussed. Interchanges are discussed separately.

5.3.1 Unsignalized Intersections

In this subsection, the following design aspects related to unsignalized intersections are discussed:

- Use of **Stop** and **Yield** control signs
- Channelization
- Lengths of auxiliary lanes
- Rotaries

Use of stop and yield signs

Unsignalized intersections are generally controlled by **Stop** and **Yield** signs. If an approach to the intersection has a stop sign then all vehicles coming on that approach need to stop before proceeding, irrespective of whether or not the conflicting stream has a vehicle. On the other hand, if an approach has a yield sign then all vehicles coming on that approach can continue to move and cross the intersection with caution, if there are no vehicles in the conflicting stream; however, if there are vehicles in the conflicting stream then the vehicles on the yield-controlled approach need to come to a stop before crossing the intersection.

Stop signs are generally employed under the following scenarios:

- On the minor street of a 'minor street'–'major street' crossing. In this case, the major street generally does not have any sign or may sometime have a yield sign.
- On all approaches to a 'minor street'–'minor street' crossing.

In both the cases mentioned above, the stop signs may not be appropriate if the volumes on the roads are such that the average delay to a vehicle on the stop-controlled approach becomes large. Although the real concern is delay to vehicles, specifications

by the codes, in this regard, are generally in terms of traffic volumes which if exceeded, signalization needs to be considered. In India, the relevant conditions are given in IRC: 93–1985 [93].

Yield signs are generally employed on a major street of a 'minor street'–'major street' intersection when the volume on the major street though higher than the minor street volume is not large in absolute terms. Such signs are recommended especially when sight distances on the major street approach are restricted.

Channelization design

Channelization refers to delineation of preferred paths for vehicles through road markings, islands, and other such static control measures. Channelization is often used at unsignalized intersections to control the movement of vehicles so as to reduce one or more of the following: (i) the area of the conflict zone, (ii) the complexity of the conflict zone (too many conflict points close to one another increase the complexity of the conflict zone since drivers have to worry about too many points within a short span of time, this increases the chance of an accident at a conflict point), (iii) the number of conflict points, and (iv) the severity of the conflict points (the severity of a conflict point may be thought of as the severity of the accident which may occur at that conflict point; for example in Figure 5.1, Conflict Point 1 is more severe than Conflict Point 2). Certain common types of channelizations are now described below:

CHANNELIZATION TO PROHIBIT OR DISCOURAGE CERTAIN MOVEMENTS

Figure 5.2 shows some examples of channelization which prohibit or discourage certain movements. By doing this, the number of conflict points at the intersection reduce (this fact can be understood by studying Figure 5.1 carefully). Note that in Figure 5.2, the dashed arrows show the movements which get prohibited and the dotted arrows show the movements which get discouraged by building the channelization aids, which are raised median dividers or islands [see Figure 5.2(a) and (b)] or particular road alignments [(see Figure 5.2(c)].

CHANNELIZATION TO PROMOTE DESIRABLE SPEEDS AND DELINEATE DESIRABLE PATHS

The channelizations which promote desirable or safe speeds and delineate desirable paths help reduce the severity of the conflict points and the complexity of the conflict zone. Figure 5.3(a) shows a case where a raised island 'A' is built to add curvature to the path of the right-turning vehicles each of which needs to stop at the intersection. The addition of curvature to the path reduces the speed of the right-turning vehicles coming on this path. This helps them to come to a stop without any problem, thereby reducing the chances of (i) rear-end collisions on this path, and (ii) right-turning vehicles getting into the major traffic stream without stopping. The figure also shows how another island 'B' and the corner radius help in promoting higher speeds for the left-turn movements. This helps in reducing the chances of rear-end collisions (conflict points) on the major street due to slowing down of the left-turning vehicles. Such channelization also reduces

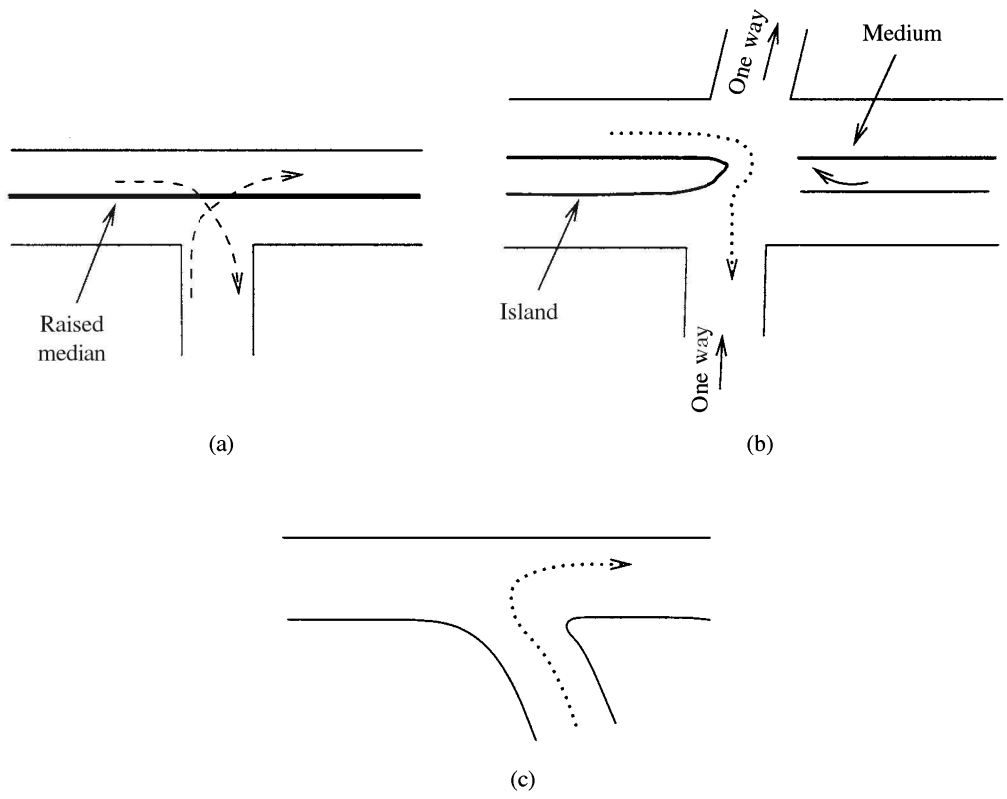


Figure 5.2 Channelization to prohibit or discourage certain movements.

the severity of a conflict point by reducing the collision angle (in this case the collision angle at the conflict point shown is very small, and hence chances of severe accidents at this point are less). Figure 5.3(b) shows another example where the geometry of the intersection is modified so as to facilitate the heavy left-turning movements—this again helps in reducing the chances of rear-end collisions as the vehicles in the heavy stream no longer have to slow down. Figure 5.3(c) shows another example where the left-turn movements are channelized so as to reduce the complexity of the conflict zone by building some separation between the conflict points shown in the figure. The reader should note that without the channelization these two points of conflict would have been closer.

CHANNELIZATION TO REMOVE STOPPED VEHICLES FROM A TRAFFIC STREAM

Channelizations which remove stopped vehicles from the main travel path help reduce accidents by eliminating the rear-end collisions which arise whenever there are stopped vehicles on a travel path. The stopped vehicles are removed into separate lanes referred to as auxiliary lanes. Figure 5.4 shows an example of such channelization. In this figure,

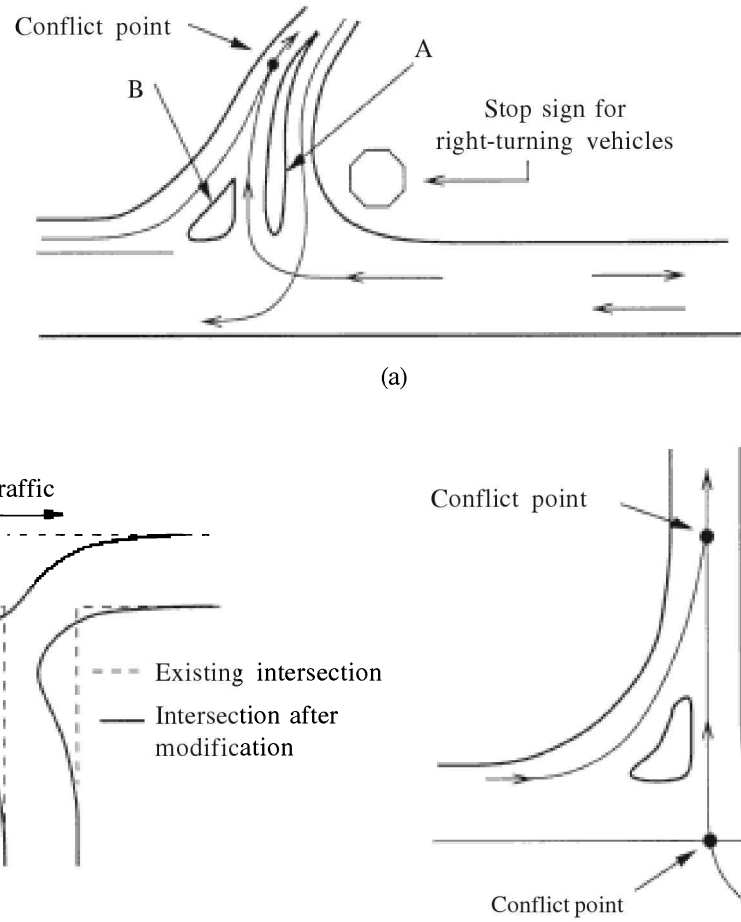


Figure 5.3 Channelization to promote desirable speeds and delineate desirable paths.

the turning movements (the turning vehicles may have to slow down or stop) are separated from the through movement lane near the intersection. Since auxiliary lanes form an important type of channelization, certain of its design aspects are discussed separately in a subsequent section.

CHANNELIZATION TO STREAMLINE FLOWS

Channelizations which streamline the lows often reduce the severity and the number of conflict points at an intersection since in general such a channelization reduces the types of movements present at an intersection. The reduction is achieved not by disallowing certain movements but by streamlining them. A very good example of such a channelization is the rotary. Figure 5.5 shows an example of a rotary intersection. Note

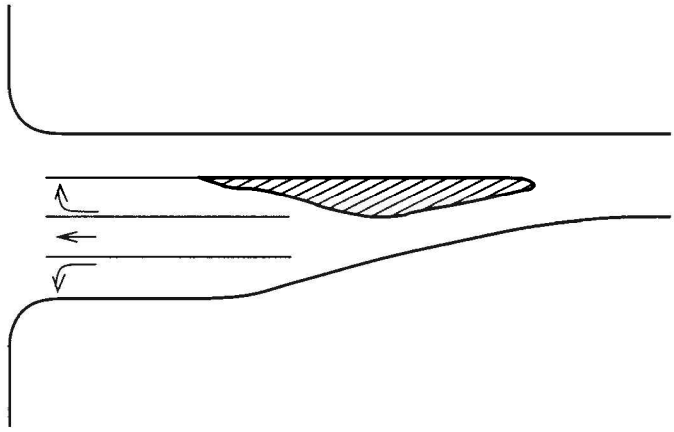


Figure 5.4 Channelization to remove stopped vehicles.

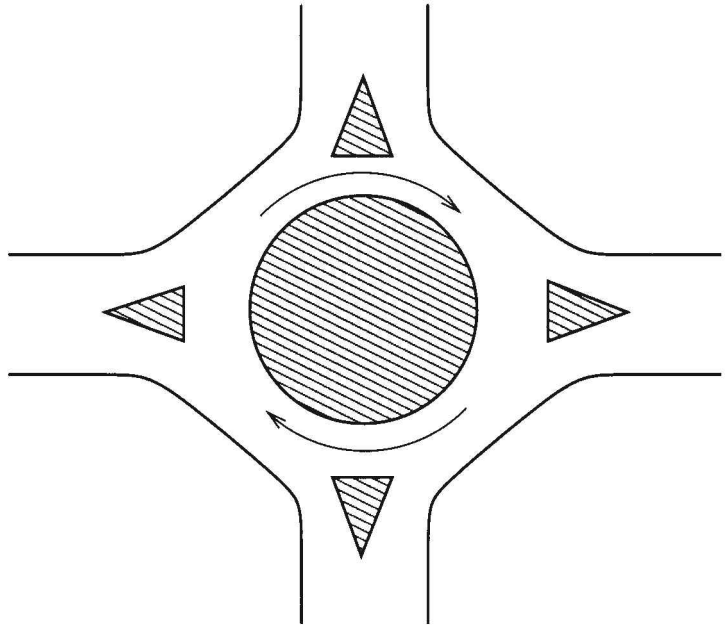


Figure 5.5 Channelization to streamline flows (rotary intersection).

that all movements at the intersection are now streamlined in the sense that irrespective of what the movement direction (like through, right-turning, left-turning) of a vehicle is, the vehicle has to merge into the traffic around the rotary and diverge from the traffic at the mouth of the appropriate road. This, not only reduces the number of conflict points but also reduces the severity of a conflict point since all the conflict points are now either

a *merge* or a *diverge* point. Since rotaries form an important type of channelization, certain of its design aspects are discussed separately in one of the subsequent sections.

Auxiliary lane

In the earlier discussions, it was stated how auxiliary lanes (used either for storage of vehicles[3] or for acceleration to merging speeds[4]) help in reducing the number of conflict points at an intersection by separating the slow moving or stopped vehicles from the mainstream flow.

The primary design variable in the case of an auxiliary lane is its length. Various countries follow various similar rules of thumb while determining the length. All such rules of thumb are based on the principle of providing adequate length so that (i) in the case of storage lanes vehicles can decelerate, come to a stop and queue up, and (ii) in the case of acceleration lanes enough distance is available for the vehicles to reach the requisite speed. Determining the length of an acceleration lane is simple and can be obtained using equations of motion. The IRC Special Publication 41 [86] also gives a table for minimum lengths of acceleration lane. In the following discussion, therefore, focus is on lengths of auxiliary lanes used for storage (or storage lanes).

In general, we can divide the entire length of an ideal storage lane into three parts: (i) the storage space, i.e., the space for stopped vehicles, (ii) the deceleration length, i.e., the length over which vehicles can decelerate to a stop, and (iii) the taper length, i.e., the length of the lane where the width (of the storage lane) changes from zero to a standard lane width. Sometimes, however, the entire length cannot be provided due to space restrictions and designers are therefore forced to provide lengths which are less than the design specifications.

The IRC guidelines regarding lengths of auxiliary lanes are contained in IRC Special Publication 41 [86]. However, the guidelines are not based on any deep understanding of the queueing process at intersections and are therefore, somewhat ad hoc. For example, the IRC provides that the storage length should be approximately 1.5 (no indication is given in the IRC publication as to why the number is 1.5) times the average number of vehicles (by vehicle type) that would queue up in the storage lane during the peak hour. It must be understood that the number of vehicles that will queue up is an outcome of a stochastic process and hence any guideline for storage length must be based on a probabilistic model of the intersection. Attempts have been made in the past to develop such guidelines and the interested reader may refer to Harmelink [98],

---

[3]Referred to as storage lanes.

[4]Sometimes a limited-length extra lane is provided on the curb side of a road so that vehicles turning left into the road can use it to accelerate and merge with the main traffic stream at more or less the same speed as that of the main stream. Such lanes are called *acceleration lanes*.

Kikuchi and Chakroborty [134], and Chakroborty, et al. [35] for a further study of related matters.[5]

Rotaries

It must be understood that around a rotary intersection, there are a lot of weaving movements generated by new vehicles merging into the stream around the rotary and there are vehicles in the stream diverging out of the rotary. Hence, the parameters of the weaving sections are important design variables for rotaries. These parameters are the width w of the roadway around the rotary and the length l between successive road entry points into the rotary over which weaving movements can take place (see Figure 5.6). Other design variables such as the average width of the roadway at the entry and exit points (average of g and h in Figure 5.6) of the rotary also play an important role. Guidelines for designing a rotary are provided in IRC:65–1976 [193].

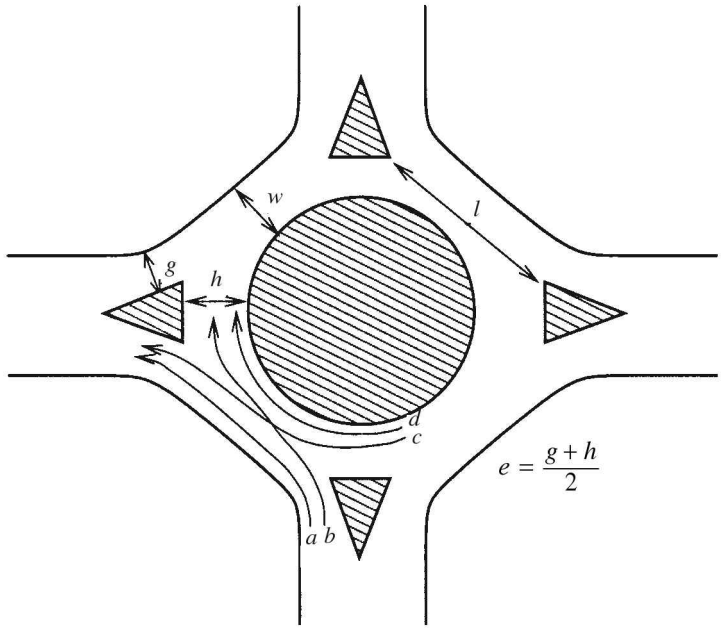


Figure 5.6 Design variables at a rotary intersection.

It should be noted that, in general, intersections are unsignalized when the volume of the conflicting flow across the intersection is not very high, and delays to vehicles even without signalization are small. As will be seen later that when these and certain

[5]The reader should note that these publications are from the USA, where vehicles follow a "keep right" policy as opposed to the Indian "keep left" policy—this means that from a traffic flow standpoint a right-turn lane in India is equivalent to a left turn-lane in the USA.

other conditions are not met, the intersections are signalized. Rotaries, however, are an exception and can handle moderate volume effectively. Hence, in the case of rotaries (as in the case of signalized intersections to be discussed later) the capacity becomes an important design consideration. The IRC:65–1976 [193] provides the following formula for determination of the capacity c_{rot} of a rotary intersection. The variables in Eq. (5.1) are defined in Figure 5.6.

$$c_{rot} = \frac{280w\left(1+\frac{e}{w}\right)\left(1-\frac{p}{3}\right)}{1+\frac{w}{l}} \tag{5.1}$$

Here all distance measures are in metres, all flows are in passenger car units per hour, and p is the proportion of weaving traffic in the flow around the rotary, i.e. $p = (b + c)/(a + b + c + d)$.

Limits on the distance ratios in Eq. (5.1) and several correction factors are also provided in the IRC code [193]. The details of these limits and factors are skipped here. It should, however, be borne in mind that given the requirements at the intersection, in terms of the total flow it should handle, the design variables should be chosen such that the capacity is adequate.

5.3.2 Signalized Intersections

As stated earlier, signalized intersections use traffic signals to control the use of the intersection. Before discussing the design aspects of a signalized intersection, various terminologies associated with a signalized intersection are described.

All legs of a signalized intersection are controlled by signals. A signal for any approach goes through three successive periods—the green period, the amber period (during which the vehicles can still move), and the red period—repeated in a cyclic manner. The duration of time that elapses between the start of one green period (or red period) for a given approach and the start of the next green period (or red period) for the same approach is referred to as the *cycle length*.

During an entire cycle length, all the approaches get their share of green time. Two or more approaches may get the green, amber and red indications at the same time. Such approaches are said to belong to the same phase. A signalized intersection generally has two to six phases during any given cycle time. The length of each phase is referred to as the *phase length*. The concept of signalization explained here is diagrammatically represented in Figure 5.7 with the help of a four-legged intersection.

Figure 5.7(a) shows the intersection and the movements allowed on each of the approaches. Each approach is indicated by an uppercase alphabet. In all there are ten approaches, marked A through J. Figure 5.7(b) shows the signal timings for each approach with reference to a time axis. Figure 5.7(c) shows, what is called the "phasing

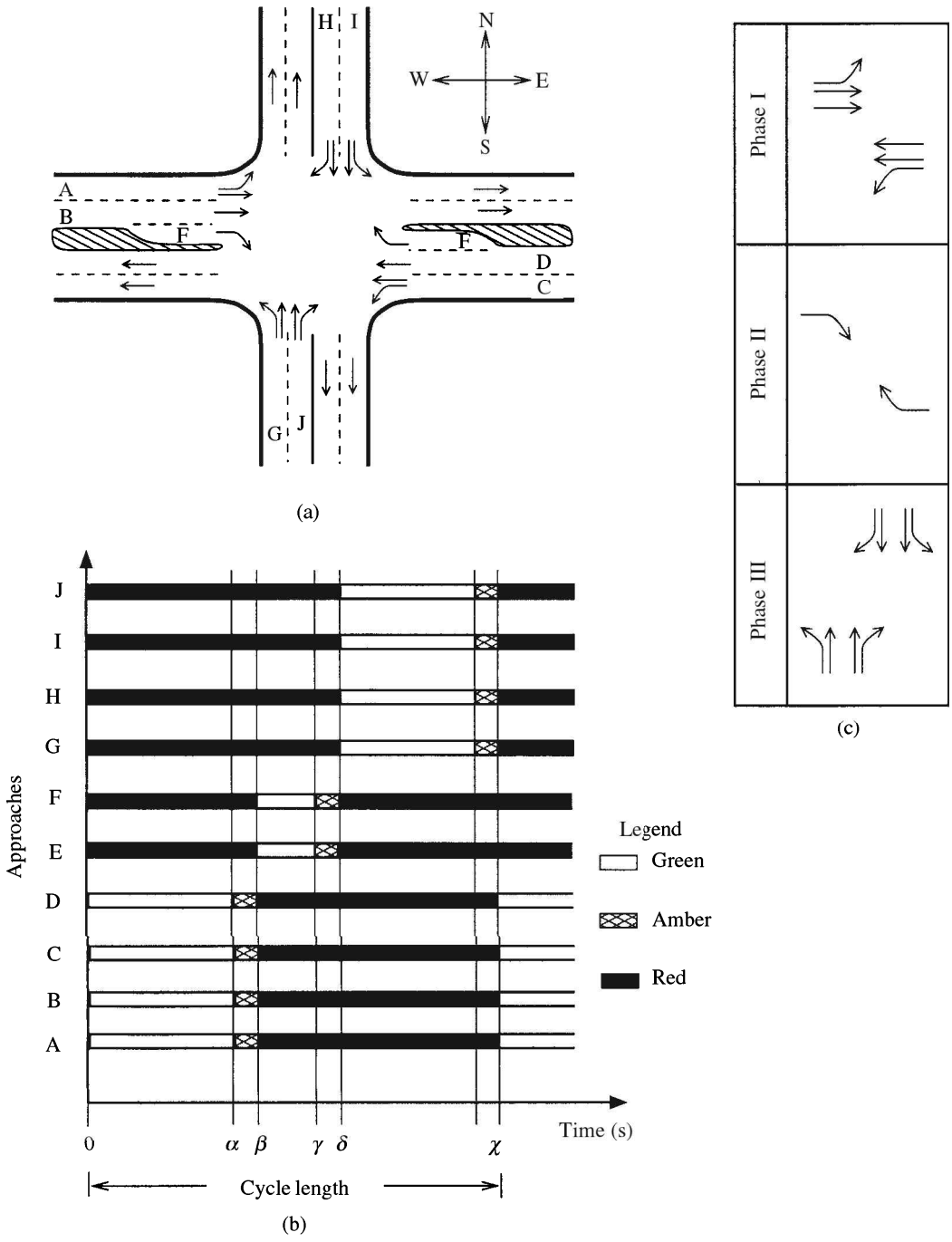


Figure 5.7 Diagrammatic representation of signalization at an intersection: (a) a typical four-legged intersection, (b) the signal indications over time for all the approaches, and (c) the phasing scheme diagram for the intersection.

scheme diagram" for the intersection. Note that the intersection signal system operates in three phases. In the first phase, approaches A, B, C, and D get green (compare with Figure 5.7(b)) and amber (the amber time can be included as part of the green or more correctly the non-red time) while all the other approaches get red. In the second phase, approaches E and F get green and amber while the rest get red. In the third phase, approaches G, H, I, and J get green and the rest red. The duration for which the approaches get green (plus amber) in a particular phase is called the *phase length*; for example the phase lengths in seconds for Phases I, II, and III are [see Figure 5.7(b)] β, $(\delta - \beta)$ and $(\chi - \delta)$ respectively. The cycle length in this example is obviously χ seconds or equivalently, the sum of all the phase lengths: $\beta + (\delta - \beta) + (\chi - \delta) = \chi$ seconds.

Signalized intersections where the signal timings (i.e. phase lengths and cycle length) remain constant over a period of time are said to employ *fixed-time* signals. However, the signals could also be *semi-actuated* or *fully-actuated*. *Semi-actuated* signals are often used when the intersection is between a major street and a minor street. In this kind of signalization, the philosophy is to always give green to the major street (or movement) unless a vehicle is detected on the minor street (or movement). Hence, this type of signalization requires the use of vehicle detectors on the minor street (or movement). In order to guard against the rare malfunction of the detector, generally the green given to the major street (or movement) is restricted to some maximum value after which for a very short duration the minor street (or movement) gets green irrespective of whether a vehicle is detected or not. Conversely, in order to guard against very quick shifts from green to red and back to green on the major street (or movement) on those rare occasions when vehicles on the minor street (or movement) seem to arrive frequently, the green on the major street (or movement) is not allowed to be less than a pre-specified minimum value. *Fully-actuated signals* use detectors on all legs of the intersection and are employed at intersections which have similar but fluctuating flows on all the approaches.

Actuated signals are sometimes referred to as demand responsive systems since their signal timings change with demand from one cycle to another. On the other hand, the signal timings of fixed-time signals do not change more than three to four times during the day. This change, however, occurs at pre-specified times and the changed signal times are also pre-determined. The changes are done in order to address the fact that demand varies significantly during different times of the day, like the morning peak hours, afternoon off-peak hours, the evening peak hours, etc. However, for heavy traffic intersections with more or less steady traffic (like what happens at major intersections during peak hours), it is better to use *fixed-time* signals rather than *actuated* signals. This is because when the traffic is heavy and steady, even if *actuated* signals are used, in theory, the apportionment of green times to the different movements would on an average become equal to the *fixed-time* signal apportionment; in such cases, the fixed-time signals are better since they reduce confusion and can also become a part of a coordinated system of signals (signal coordination and its benefits are explained later in this section). Further, in India most signalized intersections use *fixed-time* signals as they

are cheaper than the *actuated* signals. In this text, the discussion on signal timings and related matters, therefore, is with reference to *fixed-time* signals.

Having described the basic terminology and concepts related to signalization, the rest of this section focuses on the following design aspects of signalized intersections.

- Warrants for signalization (i.e. under what conditions should an intersection be signalized)
- Signal timing design
- Coordination of signals
- Auxiliary lane lengths

Warrants for signalization

Signals should be placed at an intersection only when certain conditions exist. Generally speaking, a signalized intersection is required when the volume on the intersecting roads is so high that leaving the intersection unsignalized would create flow and safety problems. Signalization may also be required if (irrespective of the volume on the roads) the number of accidents at the intersection is high; this may be due to a variety of reasons like inadequate sight distances, steep grades, etc. Sometimes signalization is provided in order to aid pedestrians to cross a road; obviously this is done when both pedestrian (across a road) and traffic volume (on the road) are on the higher side.

These conditions which justify the installation of a signal are generally formalized and referred to as the warrant (or justification) conditions for installing a signal. Various countries have devised their own set of warrant conditions for signalization of an intersection. In India, the IRC:93–1985 [93] lists the warrant conditions (or simply warrants) for signalization of an intersection. In this section, these warrants are described only in general terms. However, while discussing the warrants given in IRC:93–1985 [93], the warrants given in the *Manual on Uniform Traffic Control Devices for Streets and Highways* (MUTCD) [152] published in the USA[6] are also referred to.

The IRC:93–1985 suggests five warrant conditions. These can be classified as (i) the volume-related warrants and (ii) the accident-related warrants. The volume-related warrants are:

1. **Total traffic volume warrant** (referred to as minimum traffic volume warrant in the IRC code). This warrant basically lays down the minimum volume which must exist on the intersecting roads for at least eight hours for signalization to be justified. This warrant is in acknowledgement of the fact that if signalization is not provided in such heavy volume intersections then the hazards for flow and safety problems are quite high.

[6]It may be mentioned here that the IRC warrants seem to have been adopted in large parts from the MUTCD.

2. **Minimum delay warrant** (referred to as, 'interruption to continuous traffic warrant' in the IRC code). It may often happen that volume on one or more legs of the intersection does not satisfy the total traffic volume warrant and yet the traffic on the other legs of the intersection is much higher than the minimum traffic volume. In such cases the delay to vehicles on the low volume legs of the intersection becomes intolerably large if the intersection is unsignalized. This warrant tries to cater to such a situation and specifies the minimum volume combinations (on the different legs of the intersection) for which delays may become very large if the intersection is unsignalized. That is, like in the previous warrant, this warrant specifies (although from a different perspective) the minimum volume combinations which must exist for at least eight hours of a day for signalization to be justified.
3. **Pedestrian volume warrant** (referred to as 'minimum pedestrian volume warrant' in the IRC code). This warrant lays down the minimum pedestrian volume (across a road) and the road traffic volume (on the same road) that must exist concurrently for at least eight hours for signalization to be warranted. This warrant tries to address the issue that if the pedestrian volumes across a high volume road are high then the pedestrians face large delays to cross the road.
4. **Combination of warrants.** This is not a separate warrant condition. This warrant states that even if none of the above warrants are satisfied, signalization may still be justified if prevailing conditions at an intersection are close to more than one of the above warrant conditions.
5. **Accident related warrant** (referred to as the 'accident experience warrant' in the IRC code). It basically states that if at an intersection the frequency of *certain* kinds of accidents is higher than a minimum value then signals should be used. The word 'certain' is used to encompass only those types of accidents which the engineer feels are serious and can be corrected through signalization.

Although the IRC:93–1985 [93] includes only the above five warrants, it is felt that there are other conditions which can also justify the use of signals. For example, a minor intersection (which otherwise does not satisfy any of the five warrants stated above) in between two signalized intersections (on an arterial) could be signalized in order to perpetuate a coordinated movement of traffic on the arterial. Sometimes, the flow pattern at an intersection may be highly peaked with heavy volumes observed only during four to five hours of a day. Even, in such cases, it is felt that signals must be installed for the sake of smooth traffic flow and safety during peak and adjacent hours. The MUTCD [152], in this respect provides a more comprehensive set of eleven warrants (of which five are used in the IRC code) which cater to the above two example situations and various other similar situations.

Signal timing design

Signal timing design includes the selection or design of (i) a phasing scheme, (ii) cycle

length, and (iii) phase lengths of a signalized intersection. Each of the above aspects of signal timing design is discussed below.

PHASING SCHEME

The greater the number of phases, the better separated are the conflicting flows. Hence, safety improves. However, as stated in Chapter 4, for every phase there is some time lost (which is start-up lost time + movement lost time). Hence, the more the number of phases, the greater is the time lost, implying less time for vehicle movement. This in turn implies higher average delays for vehicles. Thus, it can be said that, the more the number of phases, the greater is the safety but lower is the efficiency. Further, a large number of phases may also cause confusion among the drivers. A good rule of thumb which can be followed while designing a phasing scheme is to start with a simple two-phase operation and increase the number of phases only if turning volumes are high and need to be separated. Further, a separate turning phase should be provided only if the turning traffic has dedicated lanes. In conclusion, it may be said that there are no clear rules of selecting a phasing scheme—the selection must be guided by experience alone.

CYCLE LENGTH

The primary concern in the determination of cycle length is the average delay to vehicles. In the discussion on delay analysis in Chapter 4, it was seen that cycle length does affect the average delay. Figure 5.8 shows the general nature of average delay versus cycle length for increasing volumes[7]. As can be seen from the plots, the nature

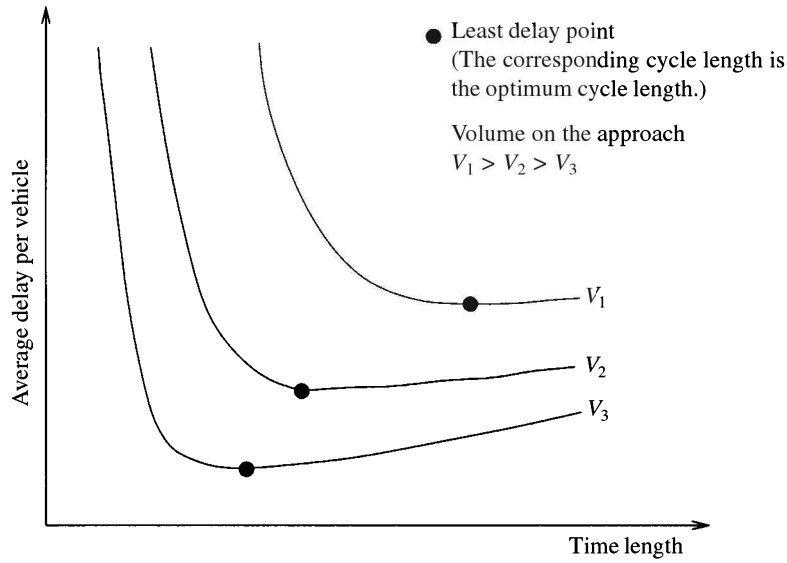


Figure 5.8 Nature of average delay per vehicle versus cycle length variations for different approach volumes.

[7]The general nature shown here is derived from the plots developed by Webster [259].

of cycle length's effect on delay is not monotonous and there exists a cycle length for which delay to vehicles is the least. Further, the sensitivity of average delay to cycle length variations around the optimum cycle length is very small. Another interesting observation is that the sensitivity of average delay to changes in cycle lengths when the cycle lengths are greater than the optimum cycle length is much less than the sensitivity of average delay to changes in cycle lengths when the cycle lengths are smaller than the optimum cycle length. These observations, as will be seen here, play a vital role in the selection of a good cycle length.

Based on the Webster's delay equation, Webster [259] developed an approximate relation for the optimum cycle length. This relation is possibly the most widely used formula for obtaining a first estimate of the cycle length (IRC:93–1985 [93] also suggests its use). As will be seen here, often the cycle length needs to be slightly changed from the value obtained from the equation because of considerations not accounted for in the equation.

The equation for determining the optimum cycle length C_o as suggested by Webster [259], is

$$C_o = \frac{1.5L + 5}{1 - \sum_{i=1}^{p} (V/s)_{cr}^{i}} \tag{5.2}$$

where

L is the total time lost per cycle

p is the total number of phases in the cycle

$(V/s)_{cr}^{i}$ is the critical flow ratio for phase i

V is volume of a particular movement

s is the saturation flow for the movement.

The procedure to obtain L and $(V/s)_{cr}^{i}$ is now described.

Here, L is the total time lost per cycle. Every time a light turns green, vehicles take a small time to react and start moving; this time is referred to as the *start-up lost time*, l_s (see Chapter 4 for a detailed discussion). Further, some part of the amber time is also unutilized, which also contributes to the lost time per phase; this time is referred to as the *movement lost time*, l_m (see Chapter 4 for a detailed discussion). In addition to these, lost times, sometimes, for safety reasons, once the amber on an approach ends, the green on another approach is not started immediately; that is, there is a time when all approaches get a red indication. This is referred to as the *all-red time*, l_r. The total time lost then is given as

$$L = \sum_{i=1}^{p} (l_s^i + l_m^i + l_r^i) \tag{5.3}$$

The term $(V/s)_{cr}^{i}$ is the critical flow ratio for phase i. In order to determine the critical flow ratio for a particular phase, first the flow ratios for all the approaches getting green in that phase are calculated. The critical value is then determined as the maximum of all the flow ratios obtained for that phase.

The actual cycle length used is generally kept a multiple of 5 and if C_o is not a multiple of 5 then the next highest integer which is a multiple of 5 seconds is used as the cycle length. Note that, as discussed earlier, increasing the cycle length does not have much impact on average delay of vehicles as the sensitivity of the latter to cycle length increments, beyond the optimal value, is quite low.

PHASE LENGTH

Phase lengths are determined by allocating the available green time in the ratio of the critical flow ratios for different phases. The available green time is determined by deducting the amber times (of all the phases) and the all-red times (if provided) from the cycle length obtained earlier.

The amber time for each phase should, in general, be determined from the dilemma zone analysis discussed later in this section. Further, the phase lengths should be checked for adequacy against pedestrian crossing times. The issues of pedestrian crossing times are also discussed later in this section. If the phase lengths are found to be inadequate from the pedestrian crossing time standpoint, then the cycle length should be increased in steps of five seconds up to the point where the phase lengths become adequate. Note that, as discussed earlier, increasing the cycle length beyond C_o does not have much impact on average delay of vehicles.

AMBER TIME DETERMINATION THROUGH DILEMMA ZONE ANALYSIS

When a light turns amber, the driver has two choices—the driver can either go and clear the intersection before the light turns red or come to a stop. The amber time should be such that at least one of these choices is always available to the driver irrespective of his or her position (on the approach) when the light turns amber. In the following, a simple analysis is done to determine the minimum amber time required to ensure the availability of one of the options.

The minimum distance required to come to a stop i.e. d_s, from the design speed of v m/s is (see also Chapter 2)

$$d_s = t_r v + \frac{v^2}{2d} \tag{5.4}$$

where t_r is the perception–reaction time and d is the comfortable deceleration rate. Hence, anyone who is a distance d_s or more away from the intersection, can stop without entering the intersection.

The distance that a vehicle can travel at the design speed during the amber time duration of τ, is $v\tau$. Thus, any vehicle (of length L) which is less than the distance, $d_c = v\tau - (w + L)$, from the intersection when the light turns amber can cross the intersection (of width w) before the light turns red.

There is a problem if $d_c < d_s$. In this case, if a vehicle is x distance away from the intersection when the light turns amber and if $d_c < x < d_s$, then the vehicle can neither clear the intersection safely (note, $x > d_c$) nor stop safely before the intersection (note, $x < d_s$). The zone from d_c to d_s (when $d_c < d_s$) is referred to as the *dilemma zone*—where the drivers cannot exercise either of the options safely. Clearly, this situation should be avoided. The minimum amber time $\tau_{\min}$, which will eliminate the dilemma zone can be easily computed by equating d_c and d_s. Thus:

$$v\tau_{\min} - (w + L) = t_r v + \frac{v^2}{2d}$$

or

$$\tau_{\min} = t_r + \frac{v}{2d} + \frac{w + L}{v} \tag{5.5}$$

The amber time obtained from Eq. (5.5) should be provided if there is no all-red phase. If, however, there is an all-red phase then the amber time plus the all-red time should exceed $\tau_{\min}$.

PEDESTRIAN CROSSING TIME

At any intersection the pedestrians may need to cross the different legs of the intersection. Pedestrians generally utilize the green time of one approach to cross the other legs (which have red indications). Unlike vehicles, pedestrians cross roads in a bunch. Hence, the crossing time of all the pedestrians could be calculated as the sum of the actual crossing time of the road and the perception–reaction time (which is the time that elapses between the start of the red indication and the time at which the pedestrians actually start crossing the road). The actual crossing time of the road is simply the quotient of the width of the road divided by the average walking speed.

The IRC:93–1985 [93] suggests the use of the following equation for determining t_p, the time to be alloted for pedestrian crossing.

$$t_p = 7 + \frac{w}{1.2} \tag{5.6}$$

where w is the width of the road (in m) which the pedestrians have to cross, and t_p is the crossing time in seconds.

EXAMPLE 5.2

Consider the signalized intersection shown in Figure 5.7(a). Also assume that a three phase signal system [as shown in Figure 5.7(c)] is selected for the intersection. The width of each approach lane is given as 3 m. The volume on each of the legs of the intersection is as follows: right-turn volume on the eastbound leg is 400 pcu/h, through and left-turn

volume on the eastbound leg is 850 pcu/h; right-turn volume on the westbound leg is 325 pcu/h, through and left-turn volume on the westbound leg is 900 pcu/h; total volume on the northbound leg is 540 pcu/h; total volume on the southbound leg is 400 pcu/h; volumes of right-turning vehicles on the northbound and southbound legs are negligible and do not affect the saturation flow on these approaches. The width of the intersection in the north-south direction is 17 m and in the east-west direction the width is 13.5 m. Since commercial vehicle traffic at this intersection is very low, assume a design vehicle length of 5 m and an approach speed on all legs of the intersection to be 45 kmph (or 12.5 m/s). Assume a perception–reaction time of 1.0 s for responding to green to amber change in signal indications and a comfortable deceleration of 4 m/s^2. Finally, assume that the start-up lost time is 2 s, the movement lost time is about half the amber time, and that there is no all-red time. For the above intersection, determine all aspects of signal timing.

Solution

We first determine the amber time duration using the dilemma zone analysis. From this analysis, the amber time for Phase I, t_a(I), and Phase II, t_a(II), can be obtained from Eq. (5.5) as

$$t_a(\text{I}) = 1.0 + \frac{12.5}{2\times 4} + \frac{13.5+5}{12.5} = 4.04 \text{ s} \approx 4 \text{ s}$$

$$t_a(\text{II}) \approx 4 \text{ s}$$

Note that, it is assumed that although the amber time is being actually determined for Phase I movements, it is also applicable to Phase II because the time spent by the right-turn and through movements while crossing the intersection is approximately the same.

The amber time for Phase III i.e. t_a (III), can be obtained as

$$t_a(\text{III}) = 1.0 + \frac{12.5}{2\times 4} + \frac{17+5}{12.5} = 4.3 \text{ s} \approx 4 \text{ s}$$

Next the pedestrian crossing time required during Phase I, t_p(I), and that during Phase III, t_p(III) need to be calculated. It is assumed that during Phase II the pedestrians are not allowed to cross the intersection (the general practice is not to allow pedestrian movements during the protected turning phases). Thus,

$$t_p(\text{I}) = 7.0 + \frac{13.5}{1.2} = 18.25 \text{ s} \approx 18 \text{ s}$$

$$t_p(\text{III}) = 7.0 + \frac{17}{1.2} = 21.2 \text{ s} \approx 21 \text{ s}$$

Next the optimal cycle length C_o should be obtained. In order to obtain C_o, first, the saturation flows (as discussed in Chapter 4) should be calculated and then the

critical flow ratios should be obtained. The saturation flows are calculated using the provisions in IRC:93–1985 [93]. However, the reader is strongly urged to also pursue the discussion on the IRC approach for determination of saturation flows given in Chapter 4.

Approach	*Width* (m)	*Saturation flow* (in pcu/(h of green))
Eastbound through and left-turn approaches	2 × 3 = 6	525 × 6 = 3150
Eastbound right-turn approach	3	1850
Westbound through and left-turn approaches	2 × 3 = 6	525 × 6 = 3150
Westbound right-turn approach	3	1850
Northbound approaches	2 × 3 = 6	525 × 6 = 3150
Southbound approaches	2 × 3 = 6	525 × 6 = 3150

Based on the above saturation flows and the phasing scheme, the critical flow ratios are calculated as follows:

Phase	*Flow ratios*	*Critical flow ratio*
I	$\frac{850}{3150} = 0.27, \frac{900}{3150} = 0.29$	max{0.27, 0.29} = 0.29
II	$\frac{400}{1850} = 0.22, \frac{325}{1850} = 0.18$	max{0.22, 0.18} = 0.22
III	$\frac{540}{3150} = 0.17, \frac{400}{3150} = 0.13$	max{0.17, 0.13} = 0.17

The total time lost,

$$L = \sum_{i=1}^{p} (l_s^i + l_m^i + l_r^i)$$

or

$$L = (2 + 0.5 \times 4 + 0) + (2 + 0.5 \times 4 + 0) + (2 + 0.5 \times 4 + 0) = 12 \text{ s}$$

Hence the optimal cycle time is

$$C_o = \frac{1.5 \times 12 + 5}{1 - (0.29 + 0.22 + 0.17)} = \frac{23}{1 - 0.68} = 71.9 \text{ s}$$

Thus the first estimate of C_o is 75 s.

The final step is to determine the phase lengths and check for their adequacy with regard to pedestrian crossing times (if green time plus amber time plus all-red time for a phase is greater than or equal to the pedestrian crossing time required for that particular phase, then the green time for that phase is considered to be adequate). Note that, the total green time (without the amber times) available for distribution among the three phases is $(75 - 3 \times 4) = 63$ s.

Phase	*Green* (s)	*Amber* (s)	t_p	*Is green adequate from t_p consideration?*
I	$\frac{0.29}{0.68} \times 63 \approx 27$	4	18	Yes (since 27 + 4 > 18)
II	$\frac{0.22}{0.68} \times 63 \approx 20$	4	0	Yes
III	$\frac{0.17}{0.68} \times 63 \approx 16$	4	21	No (since 16 + 4 < 21)

The green time for Phase III is not adequate; hence, we increase the cycle length in steps of 5 s till the green time for Phase III becomes adequate. So in the first iteration, we make the cycle length = 80 s and recalculate the phase lengths as shown below:

Phase	*Green* (s)	*Amber* (s)	t_p	*Is green adequate from t_p consideration?*
I	$\frac{0.29}{0.68} \times 68 \approx 29$	4	18	Yes
II	$\frac{0.22}{0.68} \times 68 \approx 22$	4	0	Yes
III	$\frac{0.17}{0.68} \times 68 \approx 17$	4	21	Yes

Thus the final signal timings for the intersection are as follows:

Cycle length	80 s
Phase (I) green time	29 s
Phase (I) amber time	4 s
Phase (II) green time	22 s
Phase (II) amber time	4 s
Phase (III) green time	17 s
Phase (III) amber time	4 s
No all red-time for any phase	

Signal coordination

Sometimes, especially on urban arterial and collector roads, there are frequent signals (frequency of about 1 every 2 km or higher). In such cases, it is advantageous to have coordinated signals along the corridor so as to allow platoons of vehicles to move unhindered (i.e. without stopping at each and every intersection) for extended lengths along the corridor. In this section, the procedure to design such coordinated signals for fixed time signalized intersections is described.

Consider the eastbound direction of the urban corridor shown in Figure 5.9. All the three intersections (marked A, B, and C in the figure) are signalized with a cycle length of T seconds. At all the three intersections, the through movement in the eastbound direction has a green plus amber time of $0.25T$ s. The speed of the through movement vehicles on this corridor is v m/s.

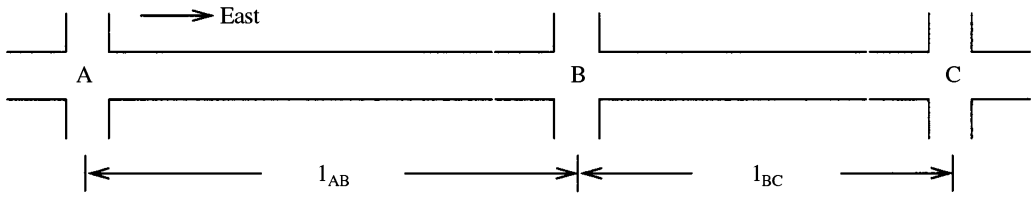


Figure 5.9 A typical urban corridor with frequent signalized intersections.

For this corridor, consider the time–distance diagram shown in Figure 5.10. Lines 1 and 2 in the figure represent the time–distance diagram (slope of the lines equals the speed of v m/s) of the first and last vehicle to leave A during the green indication, respectively. The figure also shows a possible signalization pattern at the three intersections. In this particular example, the signals at all the three intersections show the green indication to the through movement in the eastbound direction at the same time. The fall out of this is that the vehicles moving from A get a red indication at B and have to stop. Again after moving from B the same vehicles get stopped at C. The figure helps to highlight the fact that given the signalization pattern on the corridor, vehicles have to stop repeatedly while moving on the corridor leading to increased delay and driver dissatisfaction. It can be said that the signals on the corridor in their present design are completely uncoordinated.

For the same corridor, consider the time–distance diagram shown in Figure 5.11. Again, in this figure, Lines 1 and 2 have the same meaning as before. The signal timings are the same as before, the only differences are that the green at B starts $0.25T$ s after the start of the green at A, and the green at C starts $0.75T$ s after the start of the green at A. The effect of this staggering of the green times on the flow of traffic is extremely good. As can now be seen, the vehicles moving from A do not get stopped at the intersections B and C. This then leads to very little or no delay for the vehicles. Such staggering of signal timings at the intersections of a long corridor is referred to as *signal*

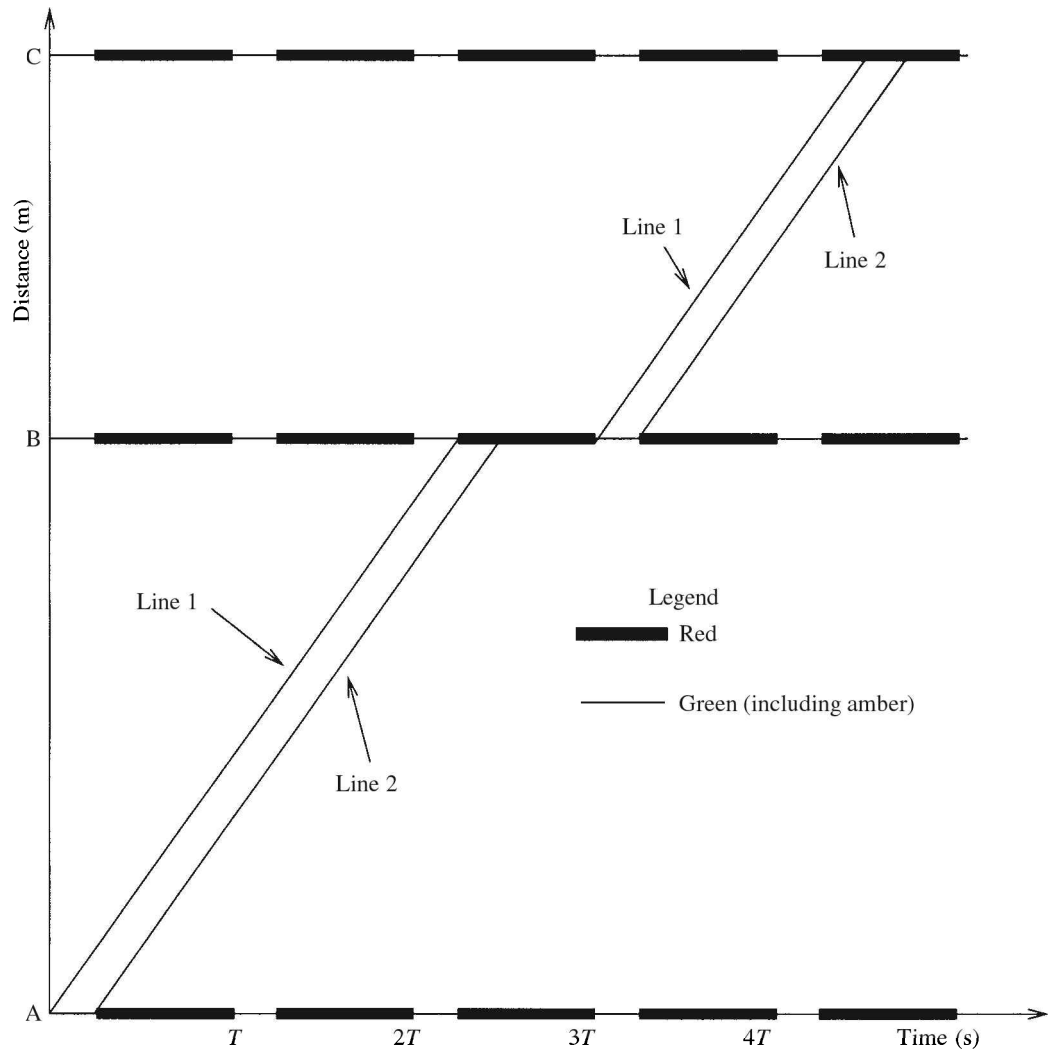


Figure 5.10 Effect of no signal coordination on traffic flow along a corridor.

coordination. The amounts by which the signals at the intersections are staggered with respect to the first intersection are referred to as the *offsets*. For example, the offset at B is $0.25T$ s and the offset at C is $0.75T$ s. Sometimes, in order to maintain uniformity in description, it is stated that the offset at A is zero (note that it is always zero), since A is the first intersection. The maximum horizontal distance between the lines (representing time–distance diagrams of the vehicles) which can cross all the intersections unhindered is referred to as the *through-band*. In the example shown here, the through-band for the through movement in the eastbound direction is $0.25T$ s. The maximum through-band possible for a given movement is always equal to the minimum green time for that movement among all the intersections being coordinated. Incidentally, in this case the maximum through band for the through movement in the

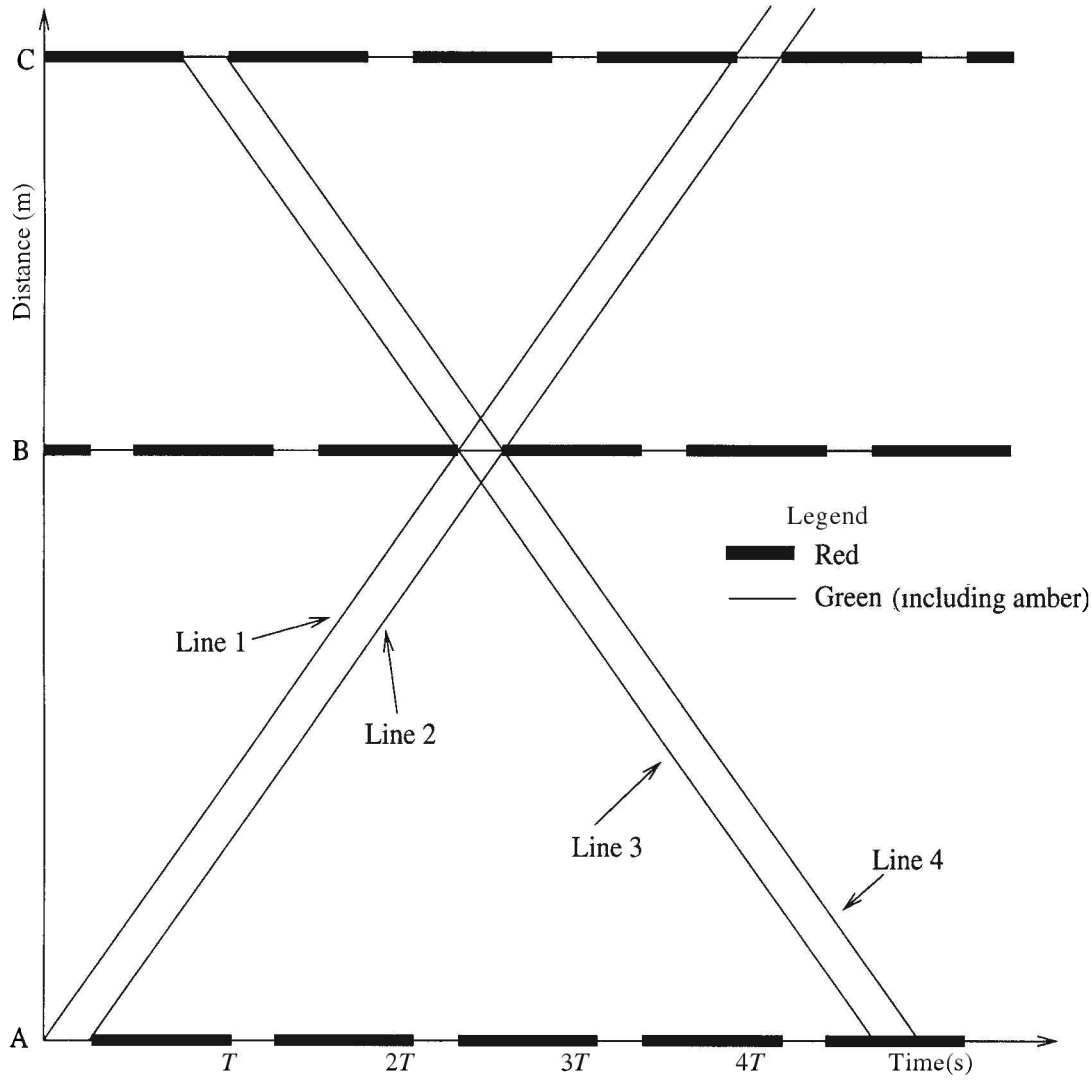


Figure 5.11 Achieving signal coordination through staggering of signal timings.

eastbound direction is $0.25T$ s. Obviously, it is not always possible to coordinate signals for the maximum through-band.

Figure 5.11 also helps to clarify another concept, that of *preferential* versus *balanced* design for signal coordination. Consider the Lines 3 and 4. These lines represent the westbound through movement on the corridor (from C to A). Note that, in general the green for the eastbound through movement will coincide with the green for the westbound through movement. The figure shows that the vehicles that leave C get stopped at A. That is, the through-band for the westbound through movement is zero. This type of signal coordination which has a very good through-band in one direction and very poor through-band in the opposing direction, is called *preferential*

coordination. On the other hand, if there is a balance between the through-bands in both the directions then the design is called a *balanced design*.

Another point which needs to be remembered is that, all the signals being coordinated should have the same cycle length (otherwise the coordination cannot be achieved over a long period of time). However, the signals may have different phase lengths.

Example 5.3 given below illustrates the design process of signal coordination and the calculation of through-bands.

EXAMPLE 5.3

The fixed time signals at the intersections of a one-way street have been coordinated. The relevant data on these intersections are given below:

Inter-section	*Green* (s)	*Amber* (s)	*Red* (s)	*Offset* (s)	*Distance from A* (m)
A	35	5	40	0	0
B	45	5	30	40	610
C	30	5	45	10	1520

The operating speed on the street is 48 kmph (or 13.33 m/s). Determine the (a) width of the through-band and the (b) offset pattern which will improve the width of the through-band.

Solution

(a) The cycle length (green + amber + red) at all the intersections is 80 s. In order to determine the through-bandwidth, the first task is to determine which is the first vehicle which can cross without being stopped. Determination of this is essentially a trial and error process. To start the process, we begin with the first vehicle at A that is, the first vehicle to leave A. This vehicle leaves A at time $t = 0$ (assuming that vehicles are waiting at the intersection when light turns green).

The vehicle reaches B at time $t = 0 + (610/13.33) = 45.75$ s. In order to determine the state of the signal at that time, the remainder of (45.75 – offset) divided by cycle time is obtained; if the remainder is less than the green time then the vehicle has reached the intersection during the green time and therefore can pass. In this case, (45.75 – 40) mod 80 = 5.75. Since 5.75 s is less than 50 s (the green + amber time at B), the vehicle can pass. Note that by obtaining the remainder we are basically determining how much time from the start of a new cycle has elapsed when the vehicle reaches the intersection.

The vehicle reaches C at a time $t = 0 + (1520/13.33) = 114$ s. Again, (114 – 10) mod 80 = 24. Since 24 s is less than 35 s (the green + amber time at C), the vehicle can pass.

Thus, the first vehicle that can pass all the three intersections is the vehicle which leaves A at time $t = 0$. The next task is to determine the last vehicle that can pass without stopping at any of the intersections. In order to determine this, the residual green times (i.e. the green time left after the arrival of the first vehicle) at all the intersections are determined. In this case, the residual green times at A, B, and C are 40 – 0 = 40 s, 50 – 5.75 = 44.25 s, and 35 – 24 = 11 s, respectively. The minimum residual green + amber time is 11 s. Hence, a vehicle which leaves A at $t = 11$ s (a vehicle can leave A at $t = 11$ s since A has a green + amber time of 40 s) after the first vehicle, will be able to cross B (since it will arrive at B at $t = 5.75 + 11 = 16.75$ s into a new cycle, when it is still green) and will be able to just cross C (since it will arrive at C at $t = 24 + 11 = 35$ s into a new cycle, when it is about to turn red). Any vehicle leaving A after $t = 11$ s will get stopped at C. Hence, the through-bandwidth is 11 s.

(b) In order to achieve an improvement in the through-bandwidth, we must first find out the signal which offers the least residual green time (and hence controls the through- bandwidth) and see if the offset can be changed to improve the residual green time. In this case, it is Signal C which has the least residual green time.

On analysis, it is seen that the residual green time at C is small not because its green time is very small but because the first vehicle arrives at C, 24 s, after the green time has started. This then shows that by increasing the offset of C, so as to allow the first vehicle to reach C at the beginning of the green time, the residual green time at C can be increased.

Hence, we change the offset of C from 10 s to 10 + 24 = 34 s. This change in offset would imply that the first vehicle from A will reach C at (114 – 34) mod 80 = 0 s, i.e. at the beginning of the green time. The residual green time at C will then change to 35 – 0 = 35 s.

On comparing with the residual green + amber times at the other intersections, it is seen that 35 s is still the least and hence, the through-bandwidth with the changed offset is 35 s.

Two points may be noted here. First, in this case, the through-bandwidth cannot be increased further, since it is now equal to the minimum green time of all the intersections being coordinated. Second, and more importantly, there could be cases where one iteration may not be enough to obtain the maximum through-bandwidth. For example, if in this case, after increasing the residual green time of C some other residual green time became the least, then one would need to find out (in the next iteration) the offset timings of this other intersection in order to improve the through-bandwidth.

For this example, the reader should verify the results by drawing the time–distance diagrams of the kind shown in Figures 5.10 and 5.11.

Auxiliary lane lengths

The purpose of auxiliary lanes was described in Section 5.3.1 on unsignalized intersections. The design aspects of lengths of auxiliary lanes were also discussed in that section. However, design considerations for the storage lengths of auxiliary lanes at

signalized intersections are slightly different. At signalized intersections, the storage length of auxiliary lanes (generally the right-turn lanes require storage) should not only be enough to store the waiting vehicles most of the time (i.e. overflow should not occur) but also be enough so that most of the time the entrance to the lane is not blocked by the through vehicles waiting at the intersection on a red. That is, the length should be equal to the maximum of the length required from overflow consideration and the length required from blockage consideration.

The relevant IRC publication (IRC Special Publication 41 [86]) recognizes these two design aspects of the storage lengths. However, the design procedure suggested by IRC is basically the same as that mentioned in Section 5.3.1 on unsignalized intersections. According to the IRC guideline, the length should be maximum of 1.5 times the average number of vehicles (by vehicle type) that would store in the auxiliary lane per cycle during the peak hour, and the average number of vehicles (by vehicle type) that would store in the adjacent through lane per cycle during the peak hour.

Some of the problems with such simplistic viewpoint of the storage requirement were mentioned in Section 5.3.1 on unsignalized intersections. The same criticism also applies to the viewpoint taken by the IRC on blockage considerations. The reader is referred to Kikuchi et al. [136] and NCHRP synthesis report 225 [144] for a better understanding of the requirements of auxiliary lane lengths at signalized intersections.

5.4 INTERCHANGES

As stated earlier, interchanges are grade separated intersections (commonly referred to as flyovers in India) where the conflict in traffic flow is resolved by duplicating the intersecting space at various heights. An example of an interchange on NH-8 in India is shown in Figure 5.12. Grade separated intersections obviously need to have roads which connect the intersecting roads. These connecting roads are called *ramps*. The important design aspects related to interchanges are—to know when to use them (*warrant conditions*), and the geometry of the interconnecting roads or ramps. There are other issues like which road should be on top, whether all movements should be grade separated, etc. However, these issues are not discussed here keeping in mind the thrust of this textbook. Only the warrant conditions and the geometric design of the ramps, and the intersecting roads are discussed here. The interested reader may refer to IRC:92–1985 [90] or AASHTO [3] for a more detailed description of the design aspects of interchanges.

5.4.1 Warrants for Interchanges

[8]The reader should note that these publications are from the USA, where vehicles follow a "keep right" policy as opposed to the Indian "keep left" policy—this means that from the traffic flow standpoint a right-turn lane in India is equivalent to a left-turn lane in the USA.

Figure 5.12 An interchange on NH-8 in India. (Source: *Realizing a Dream*, Ministry of Road Transport and Highways)

Interchanges require large capital investments and hence must be used when absolutely necessary. In order to determine when an interchange may be justified, certain warrant conditions have been evolved. Both AASHTO [3] and IRC:92–1985 [90] provide some warrant conditions. These warrant conditions are similar in nature and can be classified into the following classes: (i) design designation warrant, (ii) volume warrant, (iii) accident related warrant, and (iv) topography warrant. These warrants are briefly described here.

The **design designation warrant** states that if a road is fully access controlled, like an expressway, then all intersections on that road should be grade-separated. That is, the design designation of the road makes it imperative that all intersections become interchanges irrespective of whether at-grade intersections can handle the expected traffic on the roads.

The **volume warrant** on the other hand states that if the volume at an intersection is so high that the capacity provided by an at-grade intersection will be insufficient then interchanges should be used. The IRC:92–1985 [90] suggests that such situations may arise when the sum of flows on all the legs of the intersection exceeds 10,000 passenger cars per hour.

The **accident related warrant** states that if an intersection has a disproportionate rate of serious accidents, and if analysis of the intersection suggests that the accident hazard cannot be reduced by possible and inexpensive traffic control measures, then an interchange should be provided at the intersection. This warrant is quite subjective with neither the IRC nor the AASHTO suggesting what constitutes a "disproportionate rate of serious accidents."

The **topography warrant** states that in some cases the topography of the area may be such that the only feasible, or sometimes cheaper, alternative is an interchange; in such cases, an

interchange is definitely justified.

5.4.2 Design of Interchanges

In this section, some aspects related to the design of the grade-separated roads and the connecting ramps are discussed. The primary design features of concern here are those related to geometric design. The only design features of the grade-separated roads (and some of the ramps which may be going over one another or over one of the intersecting roads) that need to be looked into are the same as those of vertical curves discussed in Chapter 3 on geometric design. Hence, they are not repeated here.

However, the design feature specific to interchanges alone is that of the layout of the ramps. There are various kinds of layouts which are commonly used. Some of them are: (i) trumpet interchange, (ii) diamond interchange, (iii) partial clover-leaf interchange, and (iv) full clover-leaf interchange. These layouts are briefly described below. For a detailed description of these and other interchanges, the reader may refer to IRC:92–1985 [90] or AASHTO [3]; the description in AASHTO [3] is more detailed and complete.

Trumpet interchange

Figure 5.13 shows a typical trumpet interchange. The lines in the figure show movements (with the arrows beside them showing the directions); the curved lines represent movements on ramps except when they are mentioned to be a part of an at-grade intersection. The dotted portion of a line indicates that that particular movement is the one which is at a lower level at the grade separated part of the intersection. The trumpet interchange is appropriate at locations where a major road terminates at another major road. As can be seen from the figure, all movements are separated and can proceed without any interruptions.

Diamond interchange

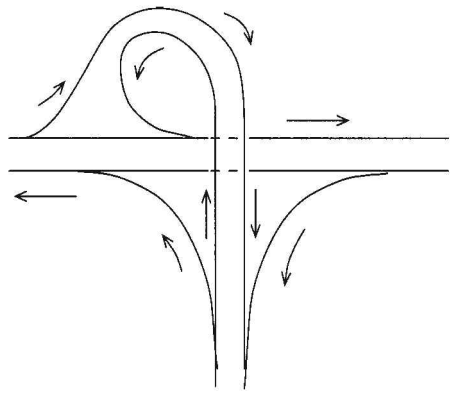


Figure 5.13 Trumpet interchange.

Figure 5.14 shows a typical diamond interchange. The representation style used in this figure is the same as that for Figure 5.13. The diamond interchange is appropriate at locations where a major road intersects a minor road (in the figure this is the road in the north-south direction). Such interchanges help avoid interruptions to traffic flow on the major road with very little additional space requirement. Their reduced space require-ment makes them suitable particularly for urban road networks. However, such interchanges necessitate the use of two at-grade intersections on the minor road as shown in the figure.

Partial clover-leaf interchange

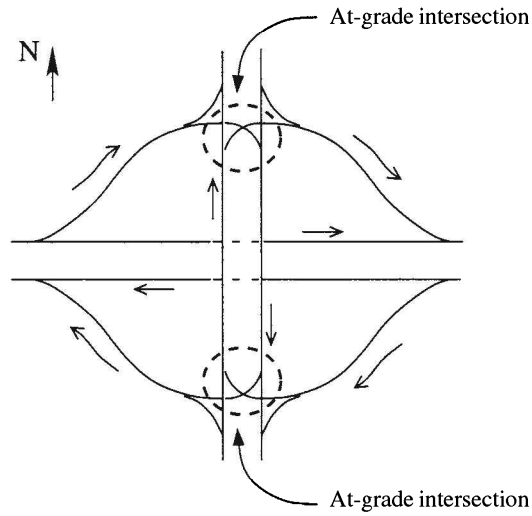


Figure 5.14 Diamond interchange.

Figure 5.15 shows a typical partial clover-leaf interchange. The representation style used in this figure is the same as that for Figure 5.13. This is a modification of the diamond interchange with the critical turning movements into the minor road being replaced by merging movements with the help of looping ramps. Such interchanges can be used when some simple at-grade intersections (i.e. with very few conflicting movements) can be tolerated on the minor road (in the figure this is the road in the north-south direction). These interchanges, however, require more space than the diamond interchanges.

Clover-leaf interchange

Figure 5.16 shows a typical clover-leaf interchange. The representation style used in this figure is the same as that for Figure 5.13. Such interchanges are appropriate when two major roads intersect one another and all conflicting movements need to be separated. These interchanges provide a high level-of-service as there are no interruptions to traffic flow due to conflicting movements. However, such interchanges require a large amount of space and often cannot be

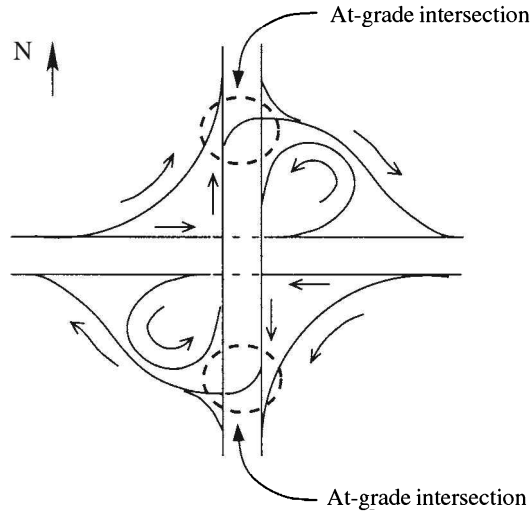


Figure 5.15 Partial clover-leaf interchange.

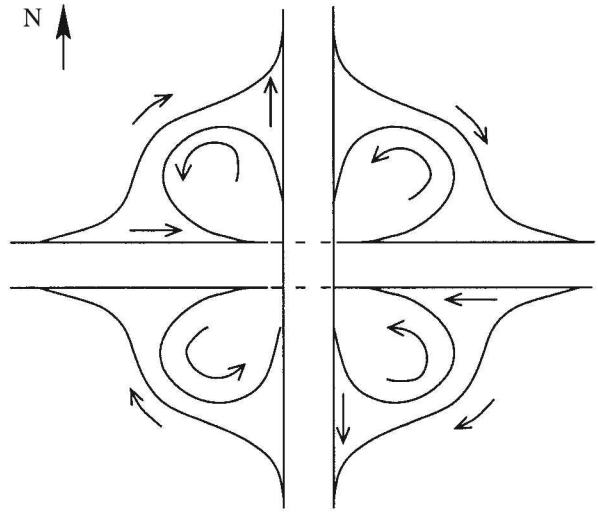


Figure 5.16 Clover-leaf interchange.

provided in urban locations. As can be seen from the figure, they do not require any at-grade intersections.

5.5 PARKING FACILITIES

As discussed in Chapter 1, one of the elements of any transportation system is the terminal. In the case of roadway transportation system, parking facilities form the terminal. Hence, for

a highway transportation system to function effectively, proper parking facilities must be built. In this section some of the design aspects of a parking facility are described. Further, parking is an issue of concern in urban areas where space is less and vehicle ownership density (per unit area) is high. Hence, much of the discussion here is based against a backdrop of parking facility design in urban areas.

Parking facilities can be either on-street or off-street. On-street parking facilities are basically the spaces near the sides of the roads where vehicles are allowed to park. Off-street parking facilities are parking spaces away from the main thoroughfare and connected to it through a service road; these spaces, unlike on-street parking spaces, are developed solely for the purposes of parking. In the following, some of the design issues related to on- and off-street parking are discussed. Before discussing the design issues, however, a discussion on the parking demand is necessary.

5.5.1 Parking Demand

The demand for parking (in terms of say, the number of vehicles wanting to park in a given hour) in an area is generated by the land use pattern in the vicinity. For example, the demand for parking in an area replete with shops will be high and the demand for parking in an area without much development either commercial or residential will be low. Another issue related to the demand for parking is the duration of parking. This is an important variable because, the average duration of parking gives an idea as to how frequently the same parking space can be used for parking different vehicles. The duration of parking in an area, as expected, is also a function of the land-use in that area. For example, the average vehicle parking duration in front of a post office will be much less than that in front of a restaurant. The average demand for parking and the average duration of parking (observed during an appropriate time of the day and over a period of time) together can give a reasonable idea about the requirement of parking spaces in a given area.

The IRC Special Publication 12 [231] provides an idea of the requirement for parking spaces (based on empirical observations and past experience) for various types of residential and commercial developments. For example, it suggests that for shops and markets, one parking space should be provided for every 80 sq. m of floor area; for apartment houses (flats), one parking space for every two flats of 50 to 99 sq. m area, and so on. It should be borne in mind here that these space requirements were determined under the socio-economic conditions prevailing before 1988 (when these recommendations were published); since then the conditions have changed substantially especially with respect to automobile ownership and hence the space requirements now will definitely be much more than those provided in the IRC publication.

5.5.2 On-street Parking

There are four issues related to on-street parking which are relevant to a transportation engineer. These are: (i) whether requirement for on-street parking exists in a particular location, (ii) whether the capacity of the roadway will be enough (after on-street parking is provided) to cater to the traffic on the road, (iii) whether on-street parking will increase safety hazards substantially, and (iv) what kind of on-street parking should be provided. Each of these issues is discussed in the following text; issues (ii) and (iii) are discussed concurrently.

Requirement for on-street parking

Given the land-use in an area, we can determine the total parking space requirement for that area. If, however, adequate (for the given requirement) off-street parking facilities are not available in the vicinity then the requirement for on-street parking would exist. We can assume that the difference between the total parking requirement and the total available off-street parking space is the requirement for on-street parking. It may, however, be pointed out that if on-street parking is available, drivers may be biased towards its use (as opposed to off-street parking), because on-street parking typically would offer less walking distances to the intended destination of the drivers.

Effect of on-street parking on capacity and safety

On-street parking adversely affects the capacity of the roadway and safety of personnel. The reduction in capacity occurs primarily due to two reasons: (i) physical roadway space is occupied by parked vehicles; this reduces the carriageway width and hence the capacity, and (ii) parking manoeuvres on the road cause frequent interruptions to traffic flow and thus reduce capacity. Although the *Highway Capacity Manual* [103] and the AASHTO [3] recognize the fact that on-street parking reduces capacity (and attempt to quantify the effect), none of the IRC codes address this issue. This lack of addressal does not mean that parking does not adversely affect the capacity in Indian conditions—it only means that this issue has not been looked into so far.

On-street parking also increases proneness to accidents on a road. This is primarily due to the manoeuvres of the vehicles which are either parking or getting out of a parked position or are looking for a parking spot.

Types of on-street parking

If the requirement for on-street parking exists and it is deemed that provision of on-street parking will not substantially affect the traffic flow (due to capacity reduction and increase in safety hazards) then on-street parking may be provided. Before proceeding further, it must be pointed out that on-street parking, except possibly on local residential roads, should never be the first choice but always a last resort.

Two types of on-street parking are possible. One, as shown in Figure 5.17(a) is referred to as *parallel parking* and two, as shown in Figure 5.17(b) is referred to as *angle parking*. Both

have their positive and negative aspects. In parallel parking, the requirement for lateral width is less (about 2.5 m from the kerb) and hence, less of the carriageway width is occupied by the parked vehicles. However, over a given road length, less number of vehicles can be parked (about 7 m should be set aside for a single space). Parallel parking also involves difficult driving manoeuvres (especially when a vehicle has to be parked between two parked vehicles) and hence cause flow interruptions on the thoroughfare. Angle parking, on the other hand, occupies more carriageway width, but by the same token more vehicles can be parked over a given road length (note that even in angle parking the same space of 2.5 m by 7 m is provided for each space but at an angle to the kerb). Driving manoeuvres required in angle parking are also less complicated; however, it has been observed that angle parking causes a lot more accidents than parallel parking. This is possibly due to the fact that drivers are sometimes blinded by other parked vehicles when backing out on to the road from a parked position. Also, large vehicle lengths cause special problems in angle parking because these vehicles may project onto the carriageway (notice this is not a problem in parallel parking as vehicles with longer lengths at the most will intrude into following parking space and not the carriageway). Hence, it is generally suggested that parallel parking should be first considered and only in special cases angle parking should be used. If, however, the carriageway width is large then angle parking may be a better choice since more vehicles can be parked over a smaller length.

5.5.3 Off-street Parking

Off-street parking facilities are facilities built solely for the purpose of parking vehicles. Various

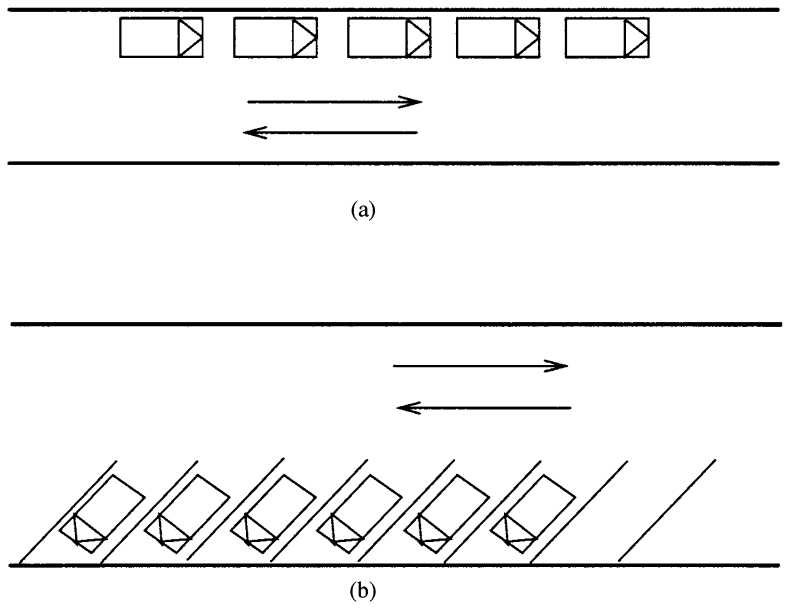


Figure 5.17 (a) Parallel on-street parking and (b) angle on-street parking.

types of such facilities, which vary in their design details, can be built. For example, off-street parking may be an open-paved space (surface parking garage), or a multistoreyed parking garage, or a park-and-ride facility, and so forth. However, any such facility must concentrate on providing space so as to (i) allow easy and independent (i.e. each vehicle can park and de-park without regard to any other parked vehicle) parking, (ii) allow easy vehicle circulation in the parking area, and (iii) utilize the space most effectively (i.e. provide as large a number of parking spaces as possible). In addition to these, each of the different types of off-street parking facilities has some special requirements which must be met. For example, a multistoreyed parking facility must have elevators for drivers and passengers to come down to ground level (or any other level) after parking and go up to the required level for picking up the vehicles.

Although it is not possible to discuss all possible aspects of an off-street parking facility in an introductory textbook of this nature, some typical designs of arrangement of parking spaces (or stalls), vehicle circulation in surface parking, and multistoreyed parking garages are described. The interested reader may refer to O'Flaherty [177] for a more detailed discussion.

5.5.4 Parking Stalls

Various kinds of arrangements of parking stalls[9] are possible which allow independent and easy access to vehicles and also utilize the available space effectively. Some of these stall designs are shown in Figure 5.18. Figure 5.18(a) also gives some of the definitions associated with parking stalls and their arrangement. In general, the stall width can be taken as about 2.5 m, stall length as about 6 m, and aisle width as (i) about 4 m for 45° and 60° parking stalls on one-way aisles, (ii) about 6 m for 90° parking stalls on one-way aisles, and (iii) about 7 m for two-way aisles.

5.5.5 Vehicle Circulation

Vehicle circulation within the parking area (be it surface or multistoreyed) is another important aspect of an off-street parking facility. The circulation is designed primarily based on access facilities to the parking garage, size and shape of the parking facility and also based on factors like pedestrian considerations, orientation of the parking stalls, fee collection system, etc. In Figure 5.19, some of the typically used vehicle circulation patterns are shown. Note that the parking stalls shown are only for the sake of better visualization of the parking facility; it does not imply that the stalls should always be at 90° for the kind of circulation shown in the figure. Figures 5.19(a), (b), and (c) show the circulation patterns for vehicles on a level parking area—they are equally applicable to either surface parking or to any particular level (or floor) of a multistoreyed parking facility. Figures 5.19(d) and (e) show some examples of vehicle circulation between the floors of a multistoreyed parking garage. Figure 5.19(d), for example, shows a system where the ramps for going up or down are within the parking floor space, w h e r e a s

[9]Space for parking a vehicle is called a *parking stall* or simply a *stall.*

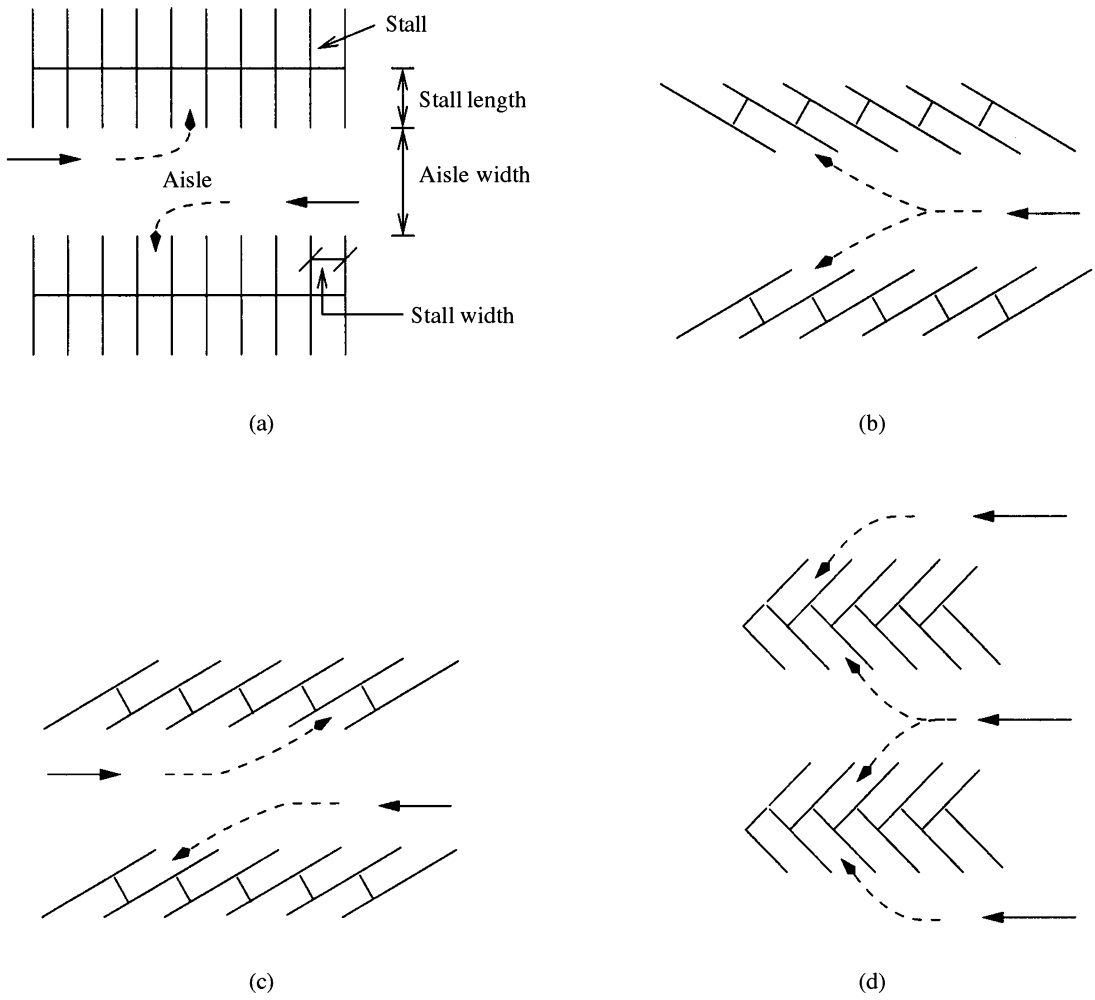


Figure 5.18 Arrangement of parking stalls: (a) 90° parking stalls with two-way aisles, (b) 60° drive-through parking stalls with one-way aisle, (c) 60° parking stalls with two-way aisles, and (d) 45° herringbone stalls with one-way aisles.

Figure 5.19(e) shows a system where the ramps are external to the parking floor space.

5.6 ROAD SIGNS

Road signs are an important part of traffic facilities design. Proper road signs aid the drivers in reaching their destinations safely and efficiently. In spite of its effect on safety and efficiency

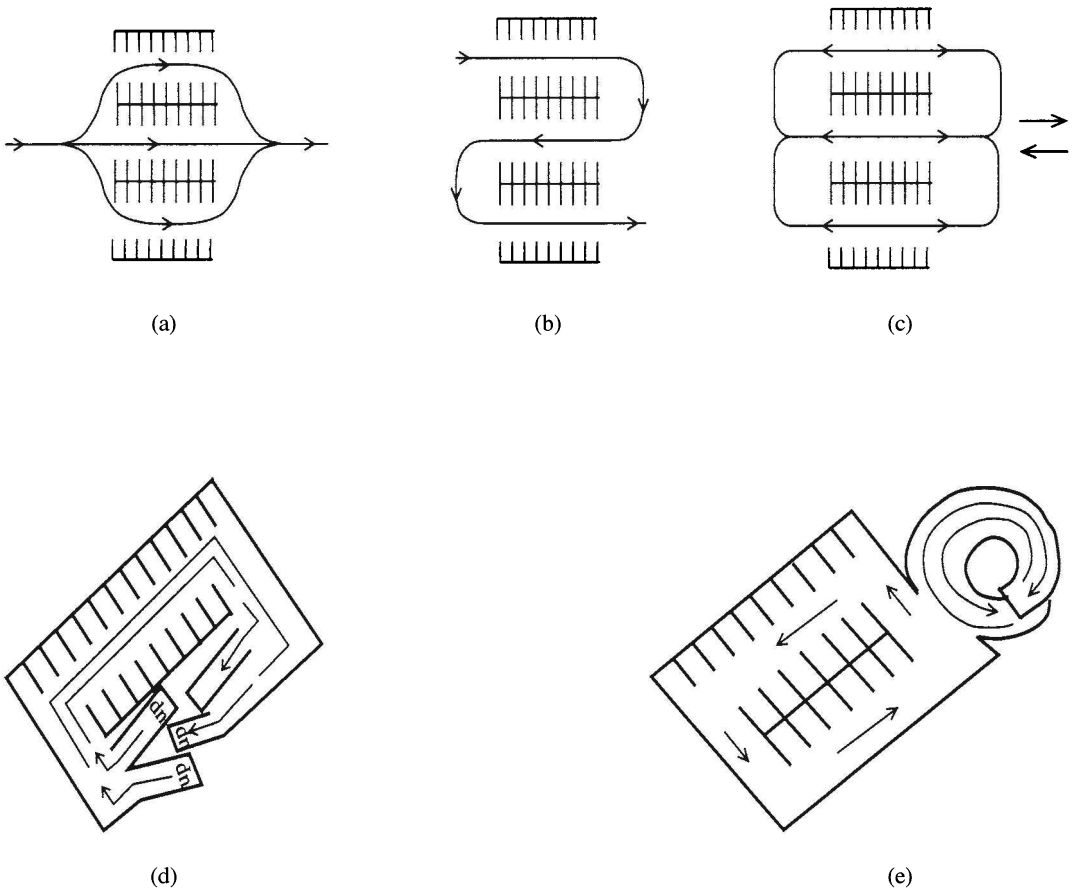


Figure 5.19 Various examples of vehicle circulation: (a) one-way circulation, (b) one-way circulation, (c) two-way circulation, (d) one-way ramp circulation with adjacent parking for multistoreyed garages, and (e) clearway, external spiral ramp circulation for multistoreyed parking garages.

of a trip, the design of road signs, especially in India has been neglected. This, however, is not the case in most other countries where substantial attention is paid to proper design of road signs.

Properly designed road signs improve safety by (i) instructing drivers on safe speeds (for example, signs like *curve ahead; reduce speed* to 40 kmph), (ii) informing drivers on impending changes in road geometry (for example, signs like *merge left—lane ends* in 500 m), and (iii) reducing driver confusion through clear signs on allowable traffic movement patterns (for example, signs like *no entry*, or *no U turn*, etc.). Road signs improve efficiency through well-designed directional signs which help drivers reach their destinations safely.

Road signs, irrespective of whether they are static (like the painted signs) or dyna-mic (like the electronic message signs) have three design elements. These are: (i) the text of the sign,

(ii) the lettering, letter sizes and the colour combination of the sign, and (iii) the placement of the sign. These design elements are discussed in the following text.

5.6.1 Text of the Sign

The details of the most common road signs are specified by some code of practice specific to a country. In India, the IRC:67–1977 [40] is the relevant code. In general, the road signs should not use text and should convey the relevant message through pictograms and shapes of the road signs. This is because the text has to be read (which requires time) whereas pictograms and shapes convey the message much faster and thereby require much less attention of the driver towards the sign. The Indian code gives a detailed description of all the pictograms that should be used.

However, certain signs like speed limit signs, directional signs, definition plates (which are attached to some signs in order to define the pictogram, especially when the pictogram is not often used or needs to be qualified in some manner) and some other special informative signs need to use text. The text for speed limit signs is for obvious reasons very much standardized. Even directional signs are reasonably standardized. The only question that arises is their letter height and colour. These form the subject matter of the next section. As regards to a design guideline for the text in a sign, when no codal suggestion exists, the rule of thumb that needs to be followed is that text should be brief and to the point—a driver should not be required to spend more than a second or two to read the sign. All signs which require a long text (say more than 3 to 4 words) should be discarded.

5.6.2 Lettering, Letter Sizes, and Colour

All countries generally have a standardized lettering scheme for the road signs. The IRC: 30–1968 [219] describes the standard lettering to be used in Indian road signs. The colour of the base and the colour of the text (or pictogram) are also specified in IRC:67–1977 [40]. The main concern while deciding on the colour is visibility (under a variety of light conditions) and clarity (i.e. how clearly the sign conveys the message).

The IRC:67–1977 [40] also specifies the range of letter sizes that should be used in the signs. The particular letter size to be used in a sign, however, is a matter of design. The letter height should be so chosen that the design driver is able to read the sign from a distance as required by the placement of the sign. Here, only the principle of choosing a letter size is mentioned. The way this principle is used in design is described in the next section together with the placement of the sign. The reason for doing this is that the size of the letter would depend on where the sign is placed and vice versa.

The size of a letter affects its readability. In general, as described in Chapter 2, a normal vision (or 6/6 vision) person can see a letter of height 8.5 mm from

approximately 6 m. As the distance increases, the size of the letter which is readable to a normal vision person increases proportionately. Also not all drivers have normal vision, for example, some may have 6/9 or 6/12 vision. (note that a 6/9 vision person can read a letter from only 6 m whereas the same letter can be read by a normal vision person from 9 m.) A more complete description on vision of drivers is provided in Chapter 2. All these factors need to be considered while choosing a letter size.

5.6.3 Placement

Signs are generally placed slightly away from the main carriageway (or roadway) at about right angles (93–95°) to the direction of travel. The lateral offset distance of the sign from the carriageway should not be too small so as to pose a hazard to the traffic nor should it be too large so that the driver's line of sight diverges to a large extent from the straight-ahead position. It is generally suggested that the sign should be placed within a 10° cone of vision (also see Chapter 2). When a sign cannot be posted on the side of the road within a reasonable angle (for example on multi-lane highways) or when the driving environment is so competitive that the driver should at all times maintain a straight-ahead line of vision, then signs are provided overhead.

The lateral positioning of a sign is only one aspect of placement. The other aspect of placement is the longitudinal positioning of a sign—the distance of the sign from the feature or point of action that the sign indicates. This positioning must take into account two factors—safety and clarity. The latter factor implies that the sign should not be placed much ahead of the arrival of the feature or point of action so that the driver is not confused when the feature or point of action arrives. The factor of safety implies that the sign should be placed far enough from the feature or point of action so that the driver can take necessary actions safely. The IRC: 67–1977 [40] suggests distances at which certain signs should be placed. However, these are only suggestions and the designers must often use their engineering skills to determine the ideal location of a sign. The following example illustrates how the placement and size of letters of a sign can be determined.

EXAMPLE 5.4

On a freeway (expressway) a sharp horizontal curve exists. The speed limit on the curve is 40 kmph. The speed limit on the expressway is 75 kmph. A sign is to be posted, warning drivers of the impending curve and advising them to slow down to the speed limit. Determine the longitudinal placement of the sign and the letter size for the sign. Assume that the perception–reaction time is 1.5 s, the coefficient of friction is 0.3, the road has 0% grade, and a design driver has 6/9 vision. Also assume that the perception–reaction time includes the time taken to read the sign.

Solution

Let the letter height used be h mm. Hence the sign can be read by a 6/6 vision person from a distance of ($6h$/8.5) m. Therefore, a 6/9 vision person can see the sign from

$$\frac{6}{9}\frac{6h}{8.5} = 0.377h \text{ m}$$

Now, for a driver to reduce the speed safely from 75 kmph (or 20.83 m/s) to 40 kmph (or 11.11 m/s) the distance required, d, is (refer to discussion on braking distance requirement in Chapter 2)

$$d = v_i t_r + \frac{v_i^2 - v_f^2}{2g(f_r + G)} = 20.83 \times 1.5 + \frac{20.83^2 - 11.11^2}{2 \times 9.81 \times (0.3 + 0)} = 84 \text{ m}$$

Therefore, the total distance required between the point at which the sign becomes legible to the driver to the start of the curve should be 84 m. If x is the distance (in m) between the sign and the start of the curve, then

$$0.377h + x = 84$$

or

$$x = 84 - 0.377h$$

This is a relation between the location of the sign and the letter height which must be maintained for safe reduction of speed. Alternatively speaking, it gives the designer a variety of choices (infinite, in fact) on the placement and letter height. Instead of leaving the problem at this stage, some more calculations (taking into account certain codal provisions) are done as these give valuable insight into the design process.

The code (Item 11.4 of IRC:67–1977 [40]) stipulates that letter sizes for expressways should not be greater than 25 cm. Thus, if a letter size of 250 mm is chosen then the distance x is –10.25 m. This indicates that the sign can be placed about 10 m after the curve has started. This, however, cannot be allowed since (i) visibility will be restricted because of the road bending and (ii) it is not a sound practice to place a sign concerning the driving restrictions on the curve and that too after the curve has started. This then implies that x should not be allowed to become negative.

Generally, the letter heights on such roads are not less than 80 mm. Hence, if 80 mm is used then the sign can be placed at a distance of 53.84 m (or 54 m) before the start of the curve. Note that what is tacitly implied here (and should therefore be checked) is that the sign is legible to the driver for a period at least equal to the time taken to read the sign. If it is assumed that out of the 1.5 s of perception–reaction time, 1 s is required to read the sign then it implies that the sign should be legible from a distance of about 20.83 m. Since the sign is legible from a distance of 0.377 × 80 = 30.16 m, it is OK. Another point which should be kept in mind, if the sign is placed on the roadside, is that the divergence in the line of sight should not exceed about 10° throughout the time

it is being read.

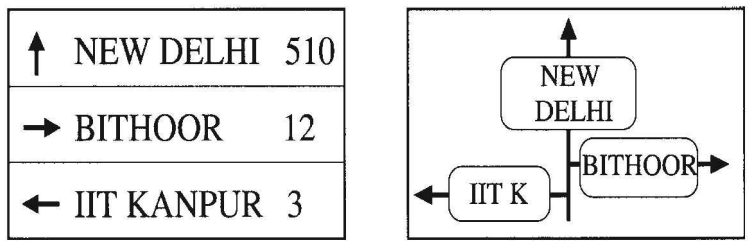
Discussion. As can be seen from this example, the time taken to read a sign (especially when the sign contains text) is an important consideration in design. Results from studies done in the UK regarding reading time (Agg [5]) as reported by Bell et al. [11] are provided here. Note that the results are from a population whose mother tongue is English and hence should be applied with caution in India.

The reading time T_s of stack type signs [see Figure 5.20(a)] in seconds is given by

$$T_s = 2.135 + 0.333 \times (\text{number of words in the sign})$$

The reading time T_m of map type signs [see Figure 5.20(b)] in seconds is given by

$$\ln (T_m) = 1.043 + 0.054 \times (\text{number of words in the sign})$$



(a) (b)

Figure 5.20 (a) Stack type sign and (b) map type sign.

For both the equations, only 1 in 1000 requires a reading time more than that specified.

Also the same source gives an empirical equation, which is based in parts (especially if lateral offset is zero) on the same procedure as followed in the previous discussions, for determination of letter height in a sign. The letter height, h, in mm is given by

$$h = 0.464vT + 9.44s$$

where v is the 85th percentile approach speed in kmph, s is the offset distance of the centre of the sign in metres, and T is the time required to read the sign in seconds.

EXERCISES

1. Visit two or more signalized intersections during the morning peak hour (say around 10:00 a.m.) and collect data on cycle length, phasing scheme, and green and amber times for all the approaches to the intersection.
2. Visit the same intersections as in Exercise 1 and collect the same set of data during the (i) afternoon off-peak period and (ii) evening peak period. Is there any difference in the values of the data collected? Comment on your findings.
3. Consider the signalized intersection shown in Figure 5.7(a). Also assume that a three phase signal system [as shown in Figure 5.7(c)] is selected for the intersection. The width of each

approach lane is given as 3 m. The volume on each of the legs of the intersection is as follows: right-turn volume on the eastbound leg is 200 pcu/h, through and left-turn volume on the eastbound leg is 650 pcu/h; right-turn volume on the westbound leg is 225 pcu/h, through and left-turn volume on the westbound leg is 1000 pcu/h; the total volume on the northbound leg is 650 pcu/h; the total volume on the southbound leg is 500 pcu/h; the volumes of right-turning vehicles on the northbound and southbound legs are negligible and do not affect the saturation flow on these approaches. The width of the intersection in the north-south direction is 16 m and in the east-west direction 12.5 m. Assume a design vehicle length of 7 m and approach speed on all legs of the intersection as 35 kmph (or 9.7 m/s). Assume a perception–reaction time of 1.0 s for responding to green to amber change in signal indications and a comfortable deceleration of 4 m/s^2. Finally, assume that the start-up time lost is 2 seconds, the movement-time lost is about half the amber time, and that there is no all-red time. For the above intersection, determine all aspects of signal timing.

4. Graphically determine the through-bandwidth for vehicles travelling from signalized intersection A towards signalized intersection D. The intersections A, B, C, and D are all on the same road and 1 km away from the previous intersection (i.e. the distance from A to B is 1 km, from B to C is 1 km and from C to D is 1 km). The operating speed on this road is 30 kmph. The signal timing data is as follows:

Inter-section	*Effective green* (s)	*Effective red* (s)	*Offset* (s)
A	40	40	0
B	50	30	40
C	40	40	0
D	40	40	60

5. For Exercise 4, can the through-bandwidth be increased by changing the offset pattern? If so, what should be the new offset times at the intersections?
6. Is the signal coordination given in Exercise 4 a balanced design? If not, how should the offsets be designed in order to have a balanced coordination?
7. For Example 5.2 on signal timing design, determine the length of the right-turn lane (on Approach E). Assume that Approach B carries half the traffic using Approaches A and B. Also, assume that only passenger cars use the intersection.
8. Given the position of a sign warning drivers to slow down to a particular speed before entering the curved section of a road just about works for 6/6 vision drivers, determine by how much should the sign be moved and in which direction (away from or towards the starting point of the curve) so that the sign can work even for 6/9 vision drivers. Assume the reading time to be negligible.
9. Determine the 99.9 percentile reading time for the signs shown in Figure 5.20.

PART II

PUBLIC TRANSPORTATION

Transit System Operations

6.1 INTRODUCTION

Transit systems are public transportation systems which move a large number of passengers to their destinations. Public transportation systems are of various types; they vary in their operational style and purpose. Some of the different types of public transportation systems are introduced below:

6.1.1 Para-transit Systems

Para-transit systems consist of small capacity vehicles which ply on more or less fixed routes but according to no pre-specified schedule. Generally, such services (i) use vehicles which shuttle between points of high demand within an urban area or within two closely-spaced urban or semi-urban areas, (ii) use the existing road network (as opposed to a separate right of way), and (iii) do not have pre-specified stop locations. A good example of such a service is the 'share-taxis' which run in many Indian cities or between two nearby cities like Kanpur and Lucknow, Siliguri and Darjeeling, and so forth.

6.1.2 Street Transit Systems (or Transit Systems)

Street transit systems (or simply transit systems) consist of medium capacity vehicles which ply on fixed routes and follow fixed schedules. Generally, vehicles on such services (i) operate on routes with pre-defined stops within an urban area or within two nearby urban or semi-urban areas and (ii) use the existing road network (as opposed to a separate right-of-way). Good examples of transit systems are the numerous city and intercity bus services. Owing to the fact that these services use the existing roads, street transit systems are the most widely used public transportation systems around the world. For this reason, in this chapter the focus is on street transit system operations.

6.1.3 Rapid Transit Systems

Rapid transit systems generally consist of more than one large capacity vehicle moving as a train on fixed routes and following fixed schedules. Generally, vehicles on such services (i) operate on routes with pre-defined stops within a greater metropolitan area or within a region consisting of a big city and associated suburbs and (ii) use dedicated paths and guided technology like trains on rails (as opposed to steered technology like buses on streets) to achieve higher operating speeds. Good examples of such services are the underground metro of Kolkata (see Figure 1.6), the regional rail network of Mumbai and Kolkata (see Figure 1.7), and so forth. Many big cities of the world use such public transportation systems. The analysis of operation of such systems generally does not vary much from the analysis of street transit systems.

The rest of this chapter is devoted to presenting analysis and design procedures for the key features of the (street) transit system. The key features of a transit system include (i) the development of an efficient set of routes, (ii) the development of an efficient stopping policy and stop locations, and (iii) the development of efficient schedules. In the next few sections, each of these topics is discussed.

6.2 ROUTE DEVELOPMENT

A transit system typically consists of many routes (referred to here as a *route set*) on each of which different number of transit units (for example, buses, trams, street cars, and the like) ply. A route is a path that a transit unit follows during its journey from the origin terminus to the destination terminus and back. Along the route there can be many stops at which the transit units halt to let passengers alight and board. The subject of this section is to discuss what constitutes a good set of routes and how we can determine such a set. The next section describes how we can determine the location of stops on a route.

6.2.1 Properties of a Good Route Set

A transit system's aim or for that matter any public transportation system's aim is to provide transport to a large portion of the population efficiently. In order to achieve this goal, the following properties of a route set are desirable.

Ridership. The percentage of the potential transit users who are served by the designed route set should be as high as possible. Given a route set and a statement of the origins and destinations of the potential transit users, we can determine the above quantity easily. An example is shown later to illustrate how this can be done.

Riding time. The time spent by any passenger on the transit unit should not be more than the time the passenger would have to spend going from his/her origin to his/her destination directly.

Transfers. Often in a transit network, we may have to change routes at some intermediate stops to go from the origin to the destination. This changing of routes is referred to as a *transfer*. The route set should be designed such that passengers are not required to make too many transfers while travelling from their origins to their destinations.

EXAMPLE 6.1

For the set of three routes on the road network shown in Figure 6.1 and the origin–destination matrix of potential transit users (in hundred persons per day) given in Table 6.1 (the origin nodes are in the first column and the destination nodes are in the first row), determine (i) the daily ridership on Route 1, (ii) the total daily ridership of the route set, (iii) the number of passengers who can travel without transferring, (iv) the number of passengers who have to transfer once, (v) the riding time of passengers going from Point C to Point F, and (vi) the shortest time in which a passenger can go from Point C to Point F. Assume that a passenger will avoid the transit system if required to make more than

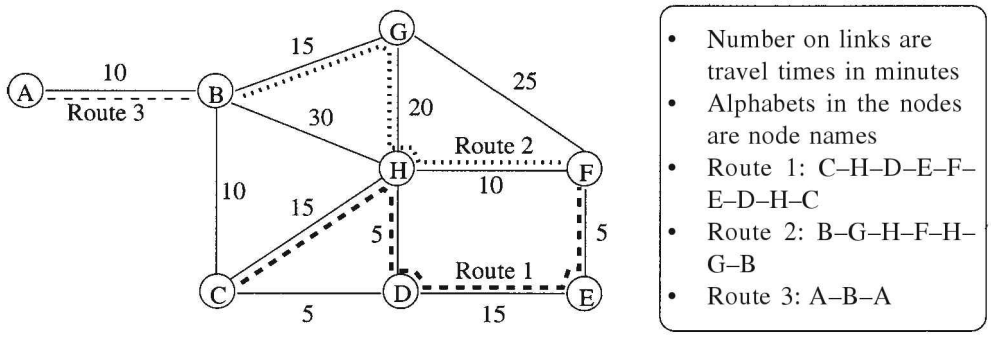


Figure 6.1 The road and route network for Example 6.1.

Table 6.1 The origin–destination matrix of potential transit users of Example 6.1

	A	B	C	D	E	F	G	H
A	0	2	1	3	2	4	1	1
B	2	0	1	1	4	2	3	1
C	1	1	0	1	3	2	4	1
D	3	1	1	0	1	4	2	3
E	2	4	3	1	0	3	2	4
F	4	2	2	4	3	0	1	2
G	1	3	4	2	2	1	0	2
H	1	1	1	3	4	2	2	0

one transfer to reach the destination from the origin. Further assume that a passenger will choose that transit route (or routes) which will take him/her to the destination in the shortest possible time; for this assume that each transfer requires a passenger to wait for 5 minutes. If two options have equal times, then the passengers choose the one which requires less number of transfers.

Solution

(i) The number of persons who use Route 1 can be obtained by realizing that (a) all passengers who go between nodes C, D, H, E, and F use Route 1 (except H to F and F to H since passengers between these points will use Route 2 which takes less time than Route 1) and (b) all passengers who travel from one of the nodes C, D, and E to one of the nodes B or G and vice versa use Route 1 (in order to transfer to or from Route 2 which takes them to their destination). Adding the demands (obtained from the table) between all the pairs of nodes formed from the above combinations, we obtain the ridership for Route 1. Note that, it is assumed (as stated in the problem) that passengers required to make more than one transfer avoid the transit system and hence do not contribute towards the ridership.

The daily ridership thus obtained on Route 1 is 7600 passengers per day.

(ii) On listing all the $8 \times 7 = 56$ origin-destination (O–D) pairs for the given problem and studying the routes which a passenger can use to travel from a given origin to a given destination it is found that passengers going from node A to either node C, D, or E (and vice versa) have to make two transfers. Hence, as per the assumptions of the problem they will not use the transit system. Adding the demands between A–C (and C–A), A–D (and D–A), and A–E (and E–A), we obtain the total demand which does not use the transit system as 1200 passengers per day.

The total number of passengers wanting to use the transit system (i.e. the sum of all the pairwise demands given in the origin–destination matrix) is 12,200 per day.

Therefore, the number of passengers who use the system (or the total ridership of the route set) is $12{,}200 - 1200 = 11{,}000$ passengers per day.

(iii) The number of passengers who can travel without transferring are those for whom the direct route between their origin and destination provides the smallest riding time. Again, here we should list all the 56 O–D pairs and determine how we would travel between each pair of points. For example, from B to G passengers will use Route 2 as it is the only route, from H to F passengers will use Route 2 as it offers a riding time of 10 minutes as opposed to Route 1 which offers a riding time of 25 minutes, and from D to F passengers will take Route 1 as it offers a direct connection and a riding time of 20 minutes while the other option (which is D to H on Route 1, transfer to Route 2 at H and H to F on Route 2) will take $5 + 5 + 10 = 20$ minutes with one transfer (recall that a passenger has to wait 5 minutes for a transfer).

On listing all the possibilities and deciding who will take direct routes (in this case they are A to B, B to G, B to H, B to F, G to H, G to F, H to F, C to H, C to D, C to E,

D to H, D to E, D to F, E to F, H to E and vice versa), the total number of passengers who can travel without transferring is obtained as 6600 passengers per day.

(iv) The number of passengers who will transfer once is obtained in much the same way as in (iii), only this time the demands for those node pairs which take one transfer will be added. In this case these node pairs are A to G, A to H, A to F, C to G, C to B, D to G, D to B, E to G, E to B, C to F, and vice versa. On adding their demands the total number of passengers who will transfer once is obtained as 4400 passengers per day.

(v) In order to determine the riding time of passengers going from node C to node F we first need to determine the routes the passengers would use to go from C to F. In this case, there are two possible ways of going: (a) we could take Route 1 to go directly from C to F or (b) we could take Route 1 from C to H, transfer to Route 2 at H, and then take Route 2 from H to F. If we use the former route then the riding time is equal to 15 + 5 + 15 + 5 = 40 minutes; however, if we uses the latter option then the riding time (including the 5 minutes of waiting time for a transfer) is 15 + 5 + 10 = 30 minutes. Obviously then (as per the assumptions of the problem) a person will use the latter option with a riding time of 30 minutes.

(vi) Determining the shortest path between any two nodes is, in general, a difficult problem for any reasonable-sized network. We generally use *minimum path* algorithms like Djkastra's algorithm or Floyd's algorithm to obtain the shortest path (see Teodorovic [233] for a good description of these algorithms). In this case, however, we can simply compare the possible paths (roads not routes) between C and F to determine the shortest among them. Here, the shortest path is C–D–H–F with a travel time of 20 minutes.

In Example 6.1, all the desired quantities were easily determined but in reality when the network is big and there are many routes the process is not so simple primarily for the following two reasons: (i) the number of possible ways of going from one point to another increase and the determination of quantities such as passengers going directly, passengers not being served, etc. cannot be done manually and the process has to be computerized, and more importantly (ii) the demand for public transport is generally not concentrated at nodes but distributed spatially over the entire network; this implies that decisions on who chooses which route depends on a variety of other factors such as distances to the stop locations of various routes, the mode of transport used to go to the stops, the number of routes which go through a particular stop, and so forth. Again, models for these decisions exist and are used to assign demands to different routes. These models are not discussed here given the scope of the book.

6.2.2 Determination of a Good Route Set

Determining a good route set is an optimization problem where conflicting objectives such as (i) maximizing the ridership, (ii) minimizing the travel time, and (iii) minimizing the number of transfers, must be balanced. In addition, these objectives could be subject

to numerous constraints such as (i) the length of a route cannot be more than some specified maximum, (ii) the number of transit units plying on a route or the entire route set must be less than the *fleet size* available with the operator (this constraint has effect on the maximum number of passengers a route or route set can carry), and so on. Further, as will be seen later, the effectiveness of a route set may also depend on the schedule of operation.

Although the problem is multi-objective in nature, often these objectives are put together in a single function. Under such a mathematical construction, the problem may be written as shown in Eq. (6.1).

Develop a set of routes by maximizing

$$W_r \sum_{k=0}^{k=k_{\max}} w_k D_k - W_{TT} \sum_{\forall i,j,j\neq i} d_{i,j} t_{i,j} - W_T \sum_{k=1}^{k=k_{\max}} D_k \tag{6.1}$$

subject to

$$\sum_{\forall m} \ell_m^n \leq L \qquad \forall n$$

$$\sum_{\forall n} \frac{CT_n}{f_n} \leq N$$

$$\frac{Q_n}{f_n TC} \leq LF \qquad \forall n$$

where

D_i is the percentage of total demand satisfied with i transfers

w_i is the weight given to demand satisfaction with different number of transfers (typically, $w_0 > w_1 > w_2 > \cdots$)

$d_{i,j}$ is the number of passengers who use the transit system to travel from i to j

$t_{i,j}$ is the shortest time in which the transit system can transport passengers from i to j

ℓ_m^n is the length of the mth link on the nth route

L is the pre-specified maximum allowable route length

CT_n is the round trip time (including the rest time at the termini) of the nth route

f_n is the frequency (in number of transit units per unit time) of operation on route n

N is the total number of transit units available with the operator

Q_n is the total demand for the nth route during a time period T

C is the capacity of a transit unit running on the system

load factor LF is the maximum proportion of the total capacity which can be utilized on an average on any route

W_i is the weight of the three distinct terms (explained later) in the objective function.

The first term in the objective function of the above formulation [Eq. (6.1)] gives the weighted sum of the ridership on the transit system, the sum is weighted (with $w_0 > w_1 > w_2 > \cdots$) because it is more desirable to have a greater percentage of the demand being satisfied directly, rather than the percentage being satisfied through one transfer, and so on. The second term gives the total riding time of all the passengers who use the transit system. The third term gives the sum of all passengers who make one or more transfers (note $k = 1$ or more). Maximizing the objective function would, therefore, imply making the first term as large as possible while making the next two terms as small as possible—this is what is desired of a good route set.

The first constraint says that the length of any route cannot be more than a maximum allowable value of L; this restriction is often necessary because of concerns about driver fatigue and labour laws. The second constraint states that the number of transit units running on the system has to be less than the available number of transit units; note that CT_n/f_n gives the number of transit units that need to be used for a route which has a round trip time of CT_n and operates at a frequency of f_n. The last constraint in a way states the maximum number of passengers that a route (operating with a frequency of f_n) can carry; what it states is that the total demand for route n over a time of T divided by the total number of transit units plying on route n during T (which is f_nT) should be less than $LF \times C$ (where C is the capacity of a transit unit).

Although the above problem is written in the format of a mathematical programming (optimization) problem, it is not a complete formulation of the problem and cannot be as such used to obtain a route set for a real-world problem. Equation (6.1) is presented here in order to put forward to the reader how the problem may be formulated for solving by traditional optimization techniques, modern optimization techniques, or heuristics.

In the past, several attempts have been made to solve the problem using traditional techniques with the realization that the general routing problem cannot be completely specified as a mathematical programming problem. Many traditional procedures, however, exist for an idealized routing problem with a large number of simplifying assumptions. The interested reader may refer to Holroyd [105], Byrne and Vuchic [27], and Byrne [28] for such procedures. However, none of them can be used for designing the actual routes on any given road network.

There are many heuristics which have been employed to solve realistic vehicle routing problems. Some of them are by Lampkins and Saalmans [142], Rosello [198], Mandl [149, 150], Dubois et al. [55], Ceder and Wilson [29], and Baaj and Mahmassani [8, 9]. Chakroborty and Dwivedi [32] have used genetic algorithms to obtain near optimal route sets for real-world networks. These algorithms are not discussed here as they are far beyond the scope of this book.

6.3 STOP LOCATION AND STOPPING POLICY

In this section, some issues related to the number of stops (and their locations) to be provided on a route and the stopping policy to be followed by the transit units plying on a route are discussed. The first subsection is on *stopping policy* and the next on *stop locations*.

The contents of this section are based on the development by Kikuchi and Vuchic [272]. For a more detailed description the reader may refer to the above source.

6.3.1 Stopping Policy

Stopping policy relates to the policy followed by the transit system as regards to stopping on the route. There are three different stopping policies of which any one may be used by a transit system. These are: (i) **All stop**, (ii) **On-call stop**, and (iii) **Demand stop**.

In the **all stop** case, the transit unit stops at all the stops on the route irrespective of whether there is any demand for that stop. For a given transit unit, the demand for a stop exists if someone wants to board the transit unit from the stop or someone on the transit unit wants to alight at that stop. In the **on-call stop** case, the transit unit stops at a stop if and only if there is a demand for that stop. In the **demand stop** case, the transit unit stops anywhere along the route where it needs to stop to pick up or drop off passengers. In this case, although the route is fixed, the stop locations are not fixed or alternatively all points on the route are viable stop locations (although the transit unit does not stop at most of them).

In order to decide which is the best policy for a transit system (or a route in the transit system) a simple analysis, based on the expected number of places at which a transit unit on a given route may have to stop, is done. The analysis procedure is shown below.

Assume that the transit units are operating on a route with n stops (i.e. stop locations). Note that n could be infinitely many. Further, assume that a transit unit has to stop at a stop only if someone is waiting at the stop or someone on the transit unit wants to get off at the stop. Also assume that the passenger demand for boarding/ alighting at any stop follows a Poisson distribution with a rate of λ. Now if h is the time headway at which transit units operate on the route being analyzed, then the probability that k persons demand to use a stop is given by

$$\text{Prob. } (k \text{ persons demand to use a stop}) = P(k) = \frac{(\lambda h)^k \exp(-\lambda h)}{k!} \tag{6.2}$$

Hence,

$$\text{Prob. (transit unit stops at a STOP)} = 1 - P(0) = 1 - \exp(-\lambda h) \tag{6.3}$$

Therefore, on the entire route of n stops, the probability the transit unit stops s times, $\Pi(s)$, is given by

$$\Pi(s) = \binom{n}{s} [1 - \exp(-\lambda h)]^s [\exp(-\lambda h)]^{n-s} \tag{6.4}$$

Hence, the expected value of s or the average number of times a transit unit stops, $E(s)$, is given by (consult any introductory book on Probability Theory to see how the following relation can be obtained from the earlier one)

$$E(s) = n[1 - \exp(-\lambda h)] \tag{6.5}$$

The rate λ, at which passenger demand for boarding/alighting arises at any stop, can be written in terms of p—the total passenger demand per unit time for the entire route. If p is the total passenger demand per unit time, then the total demand for boarding or alighting per unit time is $2p$. If n is the total number of stops, then the total passenger demand per unit time for boarding/alighting at a particular stop is $2p/n$ (assuming all stops have, on an average, equal demand). Hence, $\lambda = 2p/n$. Therefore,

$$E(s) = n[1 - \exp(-2ph/n)] \tag{6.6}$$

If the demand is very large, that is, $ph \to \infty$, then from the above relation,

$$\lim_{ph \to \infty} E(s) = n$$

The above relation implies that when the demand is large then on an average, a transit unit will have to stop at all the stops. Hence, in this case it is better to use an *all-stop* stopping policy.

If, on the other hand, demand ph is small compared to the total number of stops, n, then by expanding the exponential term and ignoring all higher order terms of (ph/n) in the expansion, we can obtain $E(s) = 2ph$. This implies that the average number of places where the transit unit will have to stop is twice the number of passengers who use the transit unit during the period h. That is, no special benefit is obtained by providing a fixed number of stops and stop locations. Hence, in the low demand case it is better to follow a *demand stop* stopping policy with no predefined stop locations on the route. For the intermediate values of demand, $E(s) < n$ [as per Eq. (6.6)] and therefore following the *on-call stop* stopping policy makes sense.

The above analysis, in effect, brings out two aspects of stopping policy: (i) it relates demands to stopping policy (see Figure 6.2) and (ii) it shows that for the entire range of demand for a route, the expression $n[1 - \exp(-2ph/n)]$ describes the average number of times a transit unit will have to stop.

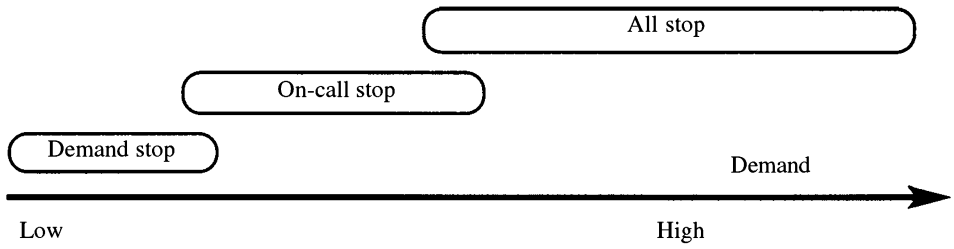


Figure 6.2 Stopping policy as a function of demand.

6.3.2 Stop Location

Decisions on the number of stops to be provided on a route and where to put those stops effect the efficiency of operation on the route as well as the ridership on the route. The number of stops affect the efficiency of operation because every stop adds to the travel time on the route and hence to the round trip time. The number and location of stops also affect the ridership as too few and poorly located stops increase (i) the access time of passengers from their origins to a stop and (ii) the egress time of passengers from a stop to their destination; and too many stops increase the riding time of passengers as the operating speed falls with the increasing number of stops.

Hence in deciding the number of stops and stop locations, we have to strike a compromise between the operating efficiency and passenger convenience. Now, one of the simple models for deciding the number of stops is described.

A good way of determining the number of stops on a route (or the stop spacing) would be to determine the number of stops (or stop spacing) which minimizes the sum of the passenger travel time cost C_p and the operator cost C_o. The passenger travel time should include the access/egress time plus the in-vehicle riding time of passengers. The operator cost is related to the number of transit units the operator has to operate in order to maintain the required frequency of service (or the required time headway between the successive transit units). Mathematically, the total cost per unit time C_T is given by

$$\begin{aligned} C_T &= C_p + C_o \\ &= pT_u k_p + Nk_o \end{aligned} \tag{6.7}$$

where

p, as before, is the number of passengers who use the route per unit time

T_u is the average passenger travel time

k_p is the cost of unit interval of time to a passenger

N is the fleet of transit units running in order to maintain the required time headway of h

k_o is the cost of operating a single transit unit per unit time.

However, before attempting to minimize the total cost, we need to determine T_u and N as these are functions of the number of stops n.

Determination of T_u

Let ℓ be the average distance a passenger (or user) travels on a route and let L be the total one way length of the route. Let v be the operating speed of the transit unit on the route and let T_v be the total travel time of the transit unit over length L.

Every time a transit unit stops at a stop location, it needs to decelerate from v to zero speed, remain stopped till all boarding and alighting is over, and then accelerate to v again. In this process, the transit unit loses time due to acceleration and deceleration and

also due to remaining stationary. Every time a transit has to decelerate from v to zero and accelerate from zero to v, it loses a time, t_{loss}. If a is the acceleration rate and b is the deceleration rate, then it can be easily shown that the extra time it requires to cover the distance of $(v^2/2a) + (v^2/2b)$ (which is equal to the distance it covers during acceleration and deceleration) during acceleration and deceleration (as opposed to moving at v) is $v[(1/2a) + (1/2b)]$. Hence,

$$t_{loss} = \frac{v}{2}\left(\frac{1}{a} + \frac{1}{b}\right) \tag{6.8}$$

Since this time is lost every time the transit unit stops and according to Eq. (6.6) on an average the transit unit stops at $n[1 - \exp(-2ph/n)]$ locations; the total time lost, T_{loss} due to the process of stopping is

$$T_{loss} = t_{loss}\, n\left[1 - \exp\left(-\frac{2ph}{n}\right)\right] \tag{6.9}$$

As stated earlier, the transit unit loses some time when it remains stationary for boarding and alighting. Since demand for the route is p passengers per unit time and if h is the headway at which transit units operate, then each transit unit caters to ph passengers. Now each passenger boards once and alights once. Hence, the total number of boardings and alightings for a transit unit is $2ph$. If μ is the rate at which boardings and alightings take place, then the total time T_s required to complete $2ph$ boardings and alightings is

$$T_s = \frac{2ph}{\mu}$$

Thus the travel time of the transit unit T_v, for the length L is given by

$$T_v = \frac{L}{v} + t_{loss} n\left[1 - \exp\left(-\frac{2ph}{n}\right)\right] + \frac{2ph}{\mu} \tag{6.10}$$

Thus the riding time of passengers on the route i.e. T_r, is given by

$$T_r = \ell\frac{T_v}{L} = \frac{\ell}{L}\left\{\frac{L}{v} + t_{loss} n\left[1 - \exp\left(-\frac{2ph}{n}\right)\right] + \frac{2ph}{\mu}\right\}$$

In addition to the riding time, the travel time of users includes the access/egress time and the waiting time. The waiting time is purely a function of headway, which is assumed to be a constant (or pre-specified) in this analysis. Hence, the waiting time is ignored here as it plays no role in determining the optimal number of stops. The access/egress time T_a, however, is dependent on the number of stops. This quantity is determined as explained below.

Assume that passengers go to their nearest stop and that the points of origin (or destination) are distributed uniformly along the route. Also assume that the speed of the mode (like walking, or bicycling) which passengers use to access the stop or egress from

the stop is v_a. Under the first assumption, it can be said that the maximum distance a passenger has to travel to reach the stop from his/her origin (or reach the destination from the stop) is half the inter-stop distance. Assuming that n stops on the route of length L are placed uniformly, the maximum distance a passenger has to travel is $(L/2n)$. Obviously, the minimum distance a passenger has to travel is zero. Since passenger origins and destinations are assumed to be distributed uniformly, on an average a passenger travels $(L/4n)$ to or from the stop. However, each passenger has to access the stop once and egress from the stop once. Thus, each passenger travels a total of $2 \times (L/4n) = (L/2n)$ distance. The access/egress time therefore is given as

$$T_a = \frac{L}{2nv_a}$$

Therefore, the average passenger (or user) travel time T_u, is given by

$$T_u = T_r + T_a = \frac{\ell}{L}\left\{\frac{L}{v} + t_{\text{loss}} n\left[1 - \exp\left(-\frac{2ph}{n}\right)\right] + \frac{2ph}{\mu}\right\} + \frac{L}{2nv_a} \tag{6.11}$$

Determination of N

If transit units are to be run on a route at a time headway of h and if T_t is the terminal *lay-off* time for the drivers, then it follows that a total of $2(T_v + T_t)/h$ transit units need to be run at any given time. Thus,

$$N = \frac{2(T_v + T_t)}{h}$$

The total cost function given in Eq. (6.7), can now be written as

$$C_T = pk_p\left[\frac{\ell}{L}\left\{\frac{L}{v} + t_{\text{loss}}\, n\left[1 - \exp\left(-\frac{2ph}{n}\right)\right] + \frac{2ph}{\mu}\right\} + \frac{L}{2nv_a}\right] + k_o\left[\frac{2}{h}\left\{\frac{L}{v} + t_{\text{loss}} n\left[1 - \exp\left(-\frac{2ph}{n}\right)\right] + \frac{2ph}{\mu} + T_t\right\}\right] \tag{6.12}$$

We can differentiate Eq. (6.12) with respect to n (by assuming n to be a real variable rather than an integer variable) and set the resulting expression to zero to obtain the optimal value of n, which is, say, n^*. Alternatively, Eq. (6.12) can be written in terms of the inter-stop spacing s (where $s = L/n$) as shown in Eq. (6.13) and differentiated with respect to s to obtain the optimal value, say, s^*. This could then be converted to n^* by using $n^* = L/s^*$. Thus,

$$C_T = pk_p\left[\frac{\ell}{L}\left\{\frac{L}{v} + t_{\text{loss}}\frac{L}{s}\left[1 - \exp\left(-\frac{2phs}{L}\right)\right] + \frac{2ph}{\mu}\right\} + \frac{s}{2v_a}\right] + k_o\left[\frac{2}{h}\left\{\frac{L}{v} + t_{\text{loss}}\frac{L}{s}\left[1 - \exp\left(-\frac{2phs}{L}\right)\right] + \frac{2ph}{\mu} + T_t\right\}\right] \tag{6.13}$$

However, in neither of the above two cases the equation obtained by equating the differentiated expression to zero yields a closed form solution for n^* (or s^*) and hence it needs to be solved numerically. Figure 6.3 shows a typical plot of n^* versus passenger demand p at various levels of k_o/k_p with all other parameters constant. It can be seen from the figure that for low values of p the optimal number of stops are very large, but as p increases the value of n^* converges to a constant value n_c. From Eq. (6.12), it can be easily derived that this constant value will be

$$n_c = \frac{L}{\sqrt{2t_{\text{loss}}v_a\left\{\ell + \dfrac{2k_oL}{pk_ph}\right\}}} \tag{6.14}$$

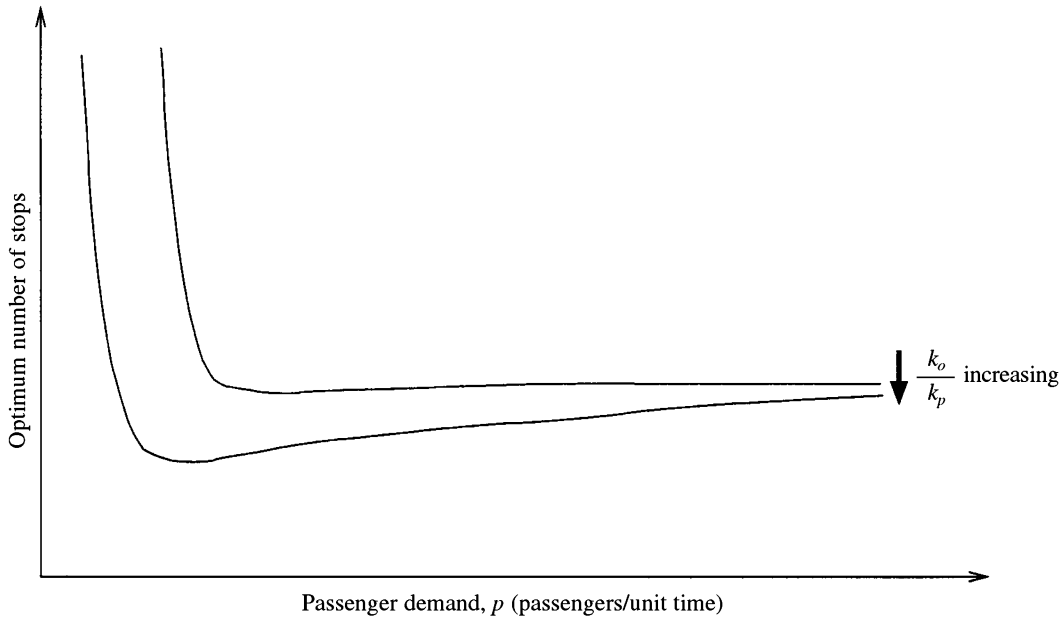


Figure 6.3 Typical plot of the optimum number of stop locations versus demand.

From Eq. (6.14) and the nature of the curves shown in Figure 6.3, the following points emerge: (i) as the length of the route L increases the optimum number of stops increase, (ii) as the length of the journey ℓ increases the optimum number of stops decrease (hence transit systems catering to long-distance passengers should have less number of stops), (iii) as the speed of access/egress v_a increases, the optimum number of stops decrease, and (iv) as the time lost due to the stopping process t_{loss} increases, the optimum number of stops decrease.

Another feature which comes out from the plot is that for low demands, the value of n^*

is large; this is in keeping with the understanding obtained in stopping policy analysis where for low demands, a policy of stopping anywhere on the route was found to be better. In fact, on plotting the expected number of stops $E(s)$ of Eq. (6.6) with the value of n^* obtained here the typical plot that develops is shown in Figure 6.4. From the figure, the three stopping policies discussed in Section 6.3.1 also emerge naturally. In the low end of the demand spectrum, the average number of stops increase almost linearly with p (compare with Eq. (6.6) for $E(s)$ for low demands) while the number of stop locations are almost infinitely many; this is a clear scenario for *demand stop* stopping policy. On the higher end of the demand spectrum, the average number of stops is almost equal to the total number of stop locations. This again indicates (as discussed in Section 6.3.1) that for high demands the best stopping policy is *all stop*. For medium demands, the average number of stops is less than the total number of stops and the total number of stops are not too many—in this case, *on-call* stopping policy seems to be the ideal choice (again as indicated in Section 6.3.1).

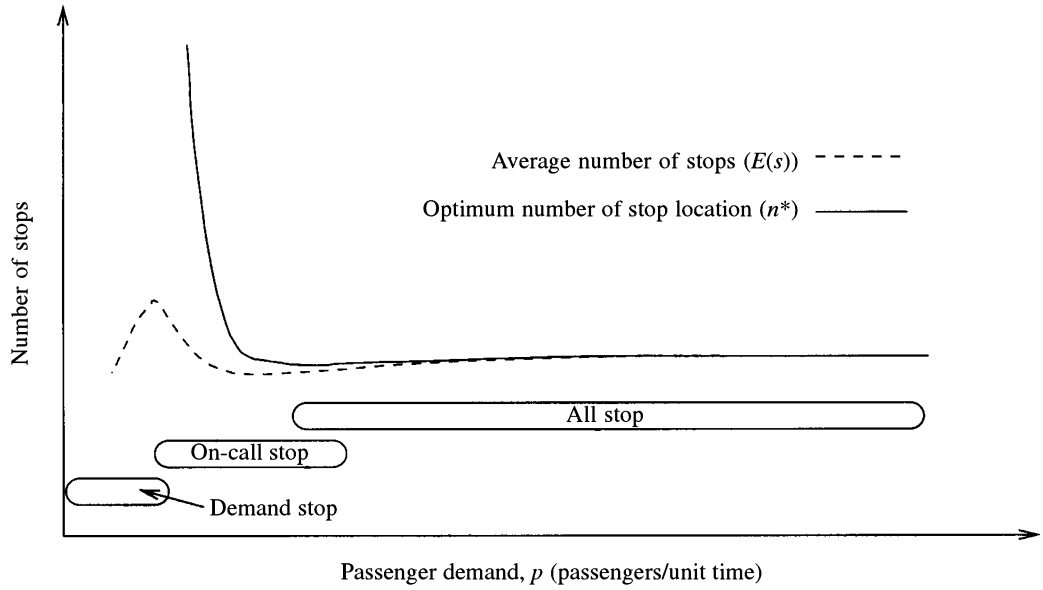


Figure 6.4 Typical plot of average number of stops versus demand.

6.4 SCHEDULE DEVELOPMENT

Transit systems run on specified routes and schedules. A schedule is a statement of times at which transit units on different routes arrives at (and departs from) a particular stop over a period of time. It is also the statement of times at which a single transit unit of a given route arrives at (or departs from) different stops on its route. The latter information can always be derived

once the schedule of arrivals and departures at all stops is known. This section analyzes how good schedules can be developed for efficient transit system operations. Before proceeding with such analysis, however, the properties of a good schedule are discussed.

6.4.1 Properties of a Good Schedule

A good schedule is one which distributes the available number of transit units in such a way that the following properties are satisfied.

Initial waiting time. The time between the arrival of a passenger at its origin stop for a particular route and the arrival of the next transit unit on that route is referred to as the initial waiting time. This should be as small as possible.

Transfer time. The time spent by passengers at an intermediate stop in order to transfer to another route (from the route on which the passenger came to the intermediate stop) is referred to as *transfer time*. This time should be small. Further, the transfer time of any passenger should not be greater than some allowable maximum value.

Policy headway. The time headway between the transit units of a given route should not be greater than a pre-specified maximum value, referred to as the *policy headway*.

6.4.2 Determination of a Good Schedule

Determining a good schedule is an optimization problem with the objective of minimizing (i) the sum of all initial waiting times and (ii) the sum of all transfer times. In addition to these objectives, there are numerous service related and physical constraints such as (i) stopping time at a stop cannot be less than a stipulated minimum, (ii) stopping time at a stop should not be more than a specified maximum, (iii) transfer time of passengers should not be more than an allowable maximum, and (iv) time headway between transit units should not be more than the policy headway.

Unlike the transit routing problem, the problem of scheduling can be reasonably formulated as a mathematical programming problem. However, before presenting the complete formulation, the problem of scheduling a single route transit system is presented. Next the problem of determining the fleet size requirement for a route is discussed. Later, the general problem with multiple routes is described.

Schedule determination of a single-route transit system

The determination of schedules is the determination of headways at which transit units of the route (or routes) should arrive at a particular stop over a period of time given the available fleet size.

Say a schedule has to be determined for a transit system operation between the time zero and time H and the total number of transit units that can be used during this time is N. Let h_i be the headway (time gap) between the arrival times of $(i-1)$th and ith transit units

(the arrival time of the zeroth transit unit is taken as the start of the time period of scheduling which in this case is zero). Further, let the arrival rate of passengers for the ith transit unit at the stop be defined by the function $v_i(t)$ as shown in Figure 6.5. The waiting time of passengers at the stop (which is the stop of their origin) for the ith transit unit, IWT_i, can be obtained by evaluating the integral[1] in the following Eq. (6.15). Note that Eq. (6.15) assumes that the stopping times of successive buses are the same and hence the difference in departure times is also equal to the headway.

$$\text{IWT}_i = \int_0^{h_i} v_i(t)(h_i - t)\,dt \tag{6.15}$$

Hence, the total initial waiting time for all the passengers arriving at the stop, IWT, is given

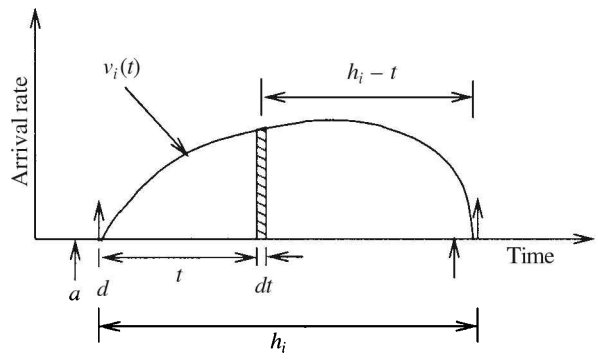


Figure 6.5 Arrival pattern of passengers at a stop and calculation of IWT_i.

by

$$\text{IWT} = \sum_{i=1}^{i=N} \int_0^{h_i} v_i(t)(h_i - t)\,dt \tag{6.16}$$

The problem therefore is

$$\text{minimize} \sum_{i=1}^{i=N} \int_0^{h_i} v_i(t)(h_i - t)\,dt \tag{6.17}$$

$$\text{subject to} \sum_{i=1}^{i=N} h_i = H$$

[1]Note that in this case, only the initial waiting time is applicable as the question of transfers does not arise.

It can be shown that for most practical arrival rate functions, $\int_0^{h_i} v_i(t)\,dt$ (which indicates the total number of passengers which arrive for the ith transit unit) is proportional to h_i, and IWT_i is proportional to h_i^2. For example, if the arrival rate is constant as shown in Figure 6.6(a), or triangular as shown in Figure 6.6(b), then $\int_0^{h_i} v_i(t)\,dt$ is $A_i h_i$ or $0.5A_i h_i$, respectively, and the IWT_i is $(A_i/2) \times h_i^2$ or $(A_i(2 - \alpha)/6) \times h_i^2$, respectively. Hence for most practical arrival rate functions, the problem stated in Eq. (6.17) reduces to

$$\text{minimize} \sum_{i=1}^{i=N} k_i h_i^2 \tag{6.18}$$

$$\text{subject to} \sum_{i=1}^{i=N} h_i = H$$

where k_i is a constant which depends on the arrival rate function.

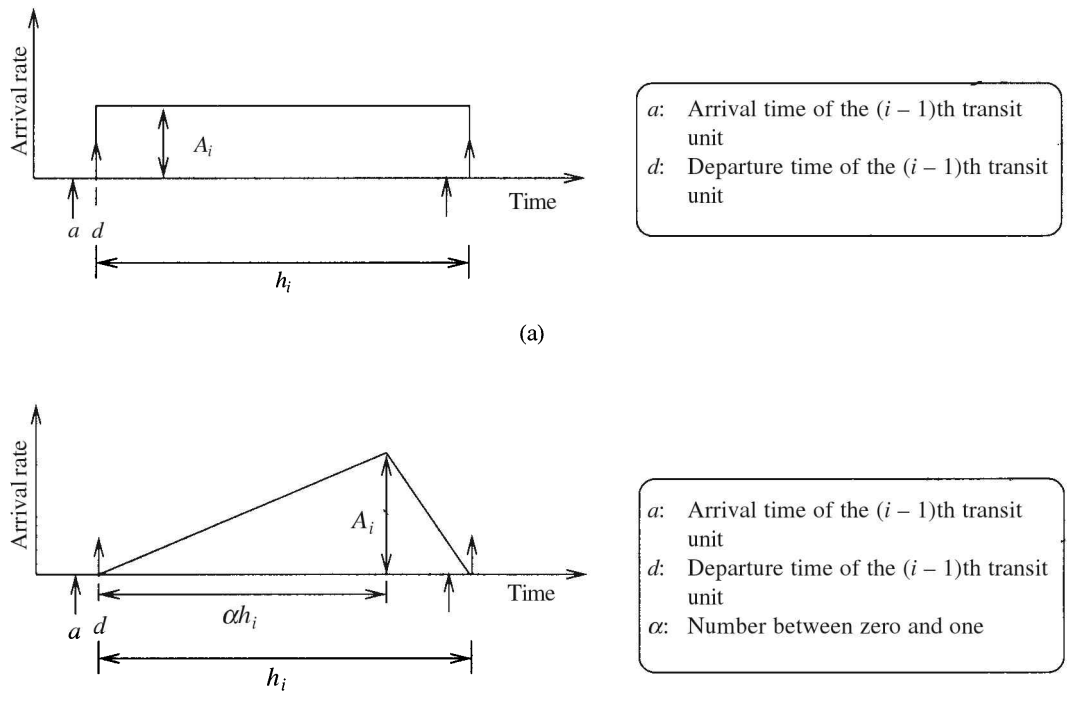


Figure 6.6 Different types of arrival patterns of passengers at a stop: (a) constant pattern, (b) triangular pattern.

Often, separate scheduling is done for periods of similar demands. For example, the schedule for the morning peak hour is done separately from the schedule for the afternoon off-peak hour period. Hence, during a scheduling period we may assume demand to be of similar nature. Under this assumption, we may assume that $k_1 = k_2 = \cdots = k_N = k$. With this assumption and the fact that $\sum_{i=1}^{i=N} kh_i^2$ will be minimum when $\sum_{i=1}^{i=N} h_i^2$ is minimum, the formulation shown in Eq. (6.18) can be further simplified as

$$\text{minimize} \sum_{i=1}^{i=N} h_i^2$$
$$\text{subject to} \sum_{i=1}^{i=N} h_i = H \tag{6.19}$$

Equation (6.19) can be solved relatively easily using the Lagrangean method (see any introductory book on optimization for a description of this method) by writing the Lagrangian ℓ as

$$\ell = \sum_{i=1}^{i=N} h_i^2 - \lambda\left(H - \sum_{i=1}^{i=N} h_i\right)$$

and differentiating it with respect to h_i and λ and setting the resulting $N + 1$ equation to zero. The optimal value of h_i, h_i^*, obtained in this manner is

$$h_i^* = \frac{H}{N} \tag{6.20}$$

This basically states that under the conditions and assumptions of this analysis, the best schedule is to equally distribute the transit units over the entire period H—a solution which also makes intuitive sense. Further, once h_i^* is obtained it can be quickly checked whether it is less than the policy headway, if not, the fleet size has to be increased. Since, this only gives the optimal headway there is no restriction on the stopping times of transit units and an appropriate value can be chosen.

EXAMPLE 6.2

Passengers arrive uniformly at a stop during the morning peak hour period of 4 hours. The only bus route going through this stop requires 80 minutes to go from its origin terminus to its destination terminus. The *lay-off* time at each terminus is 5 minutes. The number of buses available with the operator is 10. Determine the headway at which the

buses should be run in order to minimize the waiting time of the passengers at this stop. If the policy headway is 15 minutes, what should be the size of the fleet?

Solution

The round trip travel time of a bus in this route is $2 \times (80 + 5 + 5) = 180$ minutes. That is, after every 180 minutes the same bus can be used. This therefore means that for a period of 180 minutes (which is less than the scheduling time period of 240 minutes), 10 buses can be used.

Now, since for minimum waiting time of passengers, headways should be equal, the optimal headway can be obtained as $180/10 = 18$ minutes. Thus, buses should arrive at the stop every 18 minutes for 4 hours. The following table illustrates to the reader how with 10 buses a headway of 18 minutes can be maintained. Note that the scheduling time period ends at approximately 240 minutes. (It may not be exactly equal to 4 hours; the reader should try to figure out why this is so.)

Arr. Time (min)	0	18	36	54	72	90	108	126	144	162	180	198	216	234	252
Bus Num.	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5

If the policy headway is 15 minutes (which is less than 18 minutes), we have to increase the fleet size. The same bus, can only be used only after 180 minutes. Hence to maintain 15 minutes headway during the 180 minutes, we need to use a fleet of $180/15 = 12$ buses.

Fleet size allocation

In Example 6.2, the fleet size for a given route was assumed to be known. Now, consider the scenario where an operator (often the government) has a total fleet of F vehicles and wants to allocate it among the R different routes it runs in an equitable and efficient manner. This is the problem of fleet size allocation (that is, the problem of determining N for each route).

The number of transit units to be pressed into service on a given route should in general depend on the total number of passengers who will use that route during the scheduling time period. If it is assumed that the occupancy of a transit unit should be $Z\%$ of the total capacity C of the transit unit, the total demand during the scheduling time period of H for the route is D, and the round trip travel time for the route is RT then we can quickly see that the minimum number of transit units $N_{\min}$ required on the route, is

$$N_{\min} = \frac{100 \times (D/H) \times \min[RT, H]}{Z \times C}$$

This only gives a bound on the fleet size requirement of a route and does not really answer the

question of how F could be distributed efficiently among the routes.

A simple analysis, as described in the following, provides considerable insight into this problem. Consider the reasonably realistic arrival pattern of passengers shown in Figure 6.6(b). Also assume that for a particular Route j, the values of all A_is [see Figure 6.6(b)] are the same (during the scheduling time period) and equal to A_j. However, for different routes the values of A_j are different. Further, assume that the round trip travel time for Route j is RT_j. As stated earlier, the IWT (j) for all the passengers of Route j is given by

$$\text{IWT}(j) = \sum_{i=1}^{i=F_j} \frac{A_j(2-\alpha)}{6} \times h_i^2$$

where F_j is the fleet size allocated to Route j. Further, as derived in the previous section [see Eq. (6.20) and Example 6.2] in order to minimize the waiting time of passengers, the headways on Route j should be equal to $\min[H, RT_j]/F_j$. For the purposes of simplicity of notation, the numerator of this expression is denoted as TW_j (note TW_j is H if $RT_j > H$, else it is RT_j). Hence, IWT(j) is

$$\text{IWT}(j) = \sum_{i=1}^{i=F_j} \frac{A_j(2-\alpha)}{6} \times \frac{TW_j^2}{F_j^2}$$

$$= F_j \times \frac{A_j(2-\alpha)}{6} \times \frac{TW_j^2}{F_j^2}$$

$$= \frac{A_j(2-\alpha)}{6} \times \frac{TW_j^2}{F_j}$$

Therefore, for all the R routes the total IWT, i.e. $\text{IWT}_{\text{total}}$ is given by

$$\text{IWT}_{\text{total}} = \sum_{j=1}^{j=R} \frac{A_j(2-\alpha)}{6} \times \frac{TW_j^2}{F_j}$$

$$= \frac{(2-\alpha)}{6} \sum_{j=1}^{j=R} \frac{TW_j^2 A_j}{F_j}$$

Now the problem is to find the values of F_j such that $\text{IWT}_{\text{total}}$ is minimized subject to the constraint, the sum of all F_j should be equal to F (note that though the sum can be less than F, there is no reason why some of the fleet will be left idle, especially because by providing more number of transit units, the waiting time improves). Mathematically, the problem can be written as shown in Eq. (6.21)

$$\text{minimize } \frac{(2-\alpha)}{6}\sum_{j=1}^{j=R}\frac{TW_j^2 A_j}{F_j} \tag{6.21}$$

$$\text{subject to } \sum_{j=1}^{j=R} F_j = F$$

Like before, we could write the Lagrangian ℓ for the above problem as

$$\ell = \frac{(2-\alpha)}{6}\sum_{j=1}^{j=R}\frac{TW_j^2 A_j}{F_j} - \lambda\left(F - \sum_{j=1}^{j=R} F_j\right) \tag{6.22}$$

Differentiating the above expression with respect to each of the F_js and λ, we obtain $R + 1$ equations and setting them equal to zero we can show that at the optimal,

$$\frac{F_{j1}}{F_{j2}} = \frac{TW_{j1}\sqrt{A_{j1}}}{TW_{j2}\sqrt{A_{j2}}} \tag{6.23}$$

From Eq. (6.23), we can write all fleet sizes in terms of any one fleet size (say F_1). Thus,

$$F_j = \frac{TW_j\sqrt{A_j}}{TW_1\sqrt{A_1}}F_1$$

Since $\sum_{j=1}^{j=R} F_j = F$, we can write

$$\frac{F_1}{TW_1\sqrt{A_1}}\left\{TW_1\sqrt{A_1} + TW_2\sqrt{A_2} + \cdots + TW_R\sqrt{A_R}\right\} = F$$

or

$$F_1 = \left(\frac{TW_1\sqrt{A_1}}{\sum_{j=1}^{j=R} TW_j\sqrt{A_j}}\right)F$$

or, in general,

$$F_j = \left(\frac{TW_j\sqrt{A_j}}{\sum_{j=1}^{j=R} TW_j\sqrt{A_j}}\right)F \tag{6.24}$$

Since the demand on Route j, i.e. D_j is (see Fig. 6.6(b)) $0.5HA_j$ (the reader should try to confirm why this is so) or $A_j = D_j/0.5H$, we can write Eq. (6.24) as

$$F_j = \left(\frac{TW_j \sqrt{D_j}}{\sum_{j=1}^{j=R} TW_j \sqrt{D_j}} \right) F \tag{6.25}$$

Equation (6.25) for fleet sizes has been derived for a specific arrival pattern. However, since the arrival pattern is applicable to a wide variety of real-world situations, Eq. (6.25) can be used to determine the fleet size allocation among various routes. Sometimes, as in many metropolitan cities in India, in addition to the government run transit units many private operators also run transit units on various routes specified by the government. In such cases, there is no effective limit on the total fleet size. The fleet running on any given route is primarily a function of the profitability of running transit units on that route.

EXAMPLE 6.3

A bus system operator has a fleet size of 50 buses which can be used during the morning peak hour period of 4 hours. There are four routes in the system. The round trip times for Routes 1, 2, 3, and 4 are 80 min, 120 min, 120 min, and 180 min, respectively. The demands for the routes in passengers per hour at a typical stop on these routes are 400 for Route 1, 225 for Route 2, 100 for Route 3, and 225 for Route 4. Determine how should the fleet size be distributed so as to minimize the waiting time of passengers. Also, give a schedule of times at which buses on each of these routes should leave their terminus given that the first bus should leave at 7:00 am.

Solution

In this example, all the RT_js are less than H (which is 240 minutes). Hence, $TW_j = RT_j$. $D_1 = 4 \times 400 = 1600$ passengers, $D_2 = 4 \times 225 = 900$ passengers, $D_3 = 4 \times 100 = 400$ passengers, and $D_4 = 4 \times 225 = 900$ passengers.

$$TW_1 \sqrt{D_1} = 80 \times 40 = 3200$$

$$TW_2 \sqrt{D_2} = 120 \times 30 = 3600$$

$$TW_3 \sqrt{D_3} = 120 \times 20 = 2400$$

$$TW_4 \sqrt{D_4} = 180 \times 30 = 5400$$

Hence from Eq. (6.25)

$$F_1 = 50 \times \frac{3200}{3200 + 3600 + 2400 + 5400} = 10.95 \approx 11$$

$$F_2 = 50 \times \frac{3600}{3200 + 3600 + 2400 + 5400} = 12.33 \approx 12$$

$$F_3 = 50 \times \frac{2400}{3200 + 3600 + 2400 + 5400} = 8.22 \approx 8$$

$$F_4 = 50 \times \frac{5400}{3200 + 3600 + 2400 + 5400} = 18.49 \approx 19$$

The headways at which these routes should operate are 80/11 ≈ 8 min (by making $F_1 = 10$), 120/12 = 10 min, 120/8 = 15 min, and 180/19 = 9 min (by making $F_4 = 20$). Table 6.2 shows how the buses of various routes are scheduled to leave the terminus assuming 8 min headway for Rt. 1, 10 min for Rt. 2, 15 min for Rt. 3, and 9 min for Rt. 4.

Table 6.2 Schedule of departure times of buses from terminus for Example 6.3 on fleet size allocation

Departure time				Bus number			
Rt. 1	Rt. 2	Rt. 3	Rt. 4	Rt. 1	Rt. 2	Rt. 3	Rt. 4
7:00	7:00	7:00	7:00	1	1	1	1
7:08	7:10	7:15	7:09	2	2	2	2
7:16	7:20	7:30	7:18	3	3	3	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7:56	8:10	8:45	8:03	8	8	8	8
8:04	8:20	9:00	8:12	9	9	1	9
8:12	8:30	9:15	8:21	10	10	2	10
8:20	8:40	9:30	8:30	1	11	3	11
8:28	8:50	9:45	8:39	2	12	4	12
8:34	9:00	10:00	8:48	3	1	5	13
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9:08	9:40	11:00	9:24	7	5	1	17
⋮	⋮	OVER	⋮	⋮	⋮	—	⋮
9:32	10:10		9:51	10	8		20
9:40	10:20		10:00	1	9		1
⋮	⋮		⋮	⋮	⋮		⋮
10:12	11:00		10:36	5	1		5
10:20	OVER		10:45	6	—		6
10:28			10:54	7			7
10:36			11:03	8			8
10:44			OVER	9			—
10:52				10			
11:00				1			
OVER				—			

Schedule determination of a multiple-route transit system

Figure 6.7 shows a typical transit system network. On each route there are several stops (not shown in the figure). Among these stops, some are transfer stops (points circled in the figure) at which passengers transfer themselves from one route to another. Schedule determination of such a multiple-route system is much more difficult than a single-route system because of transfer time considerations. A mathematical programming formulation of the problem is provided below. The formulation presented in the following Eq. (6.26) is a slight extension of the formulation first presented in Chakroborty et al. [31].

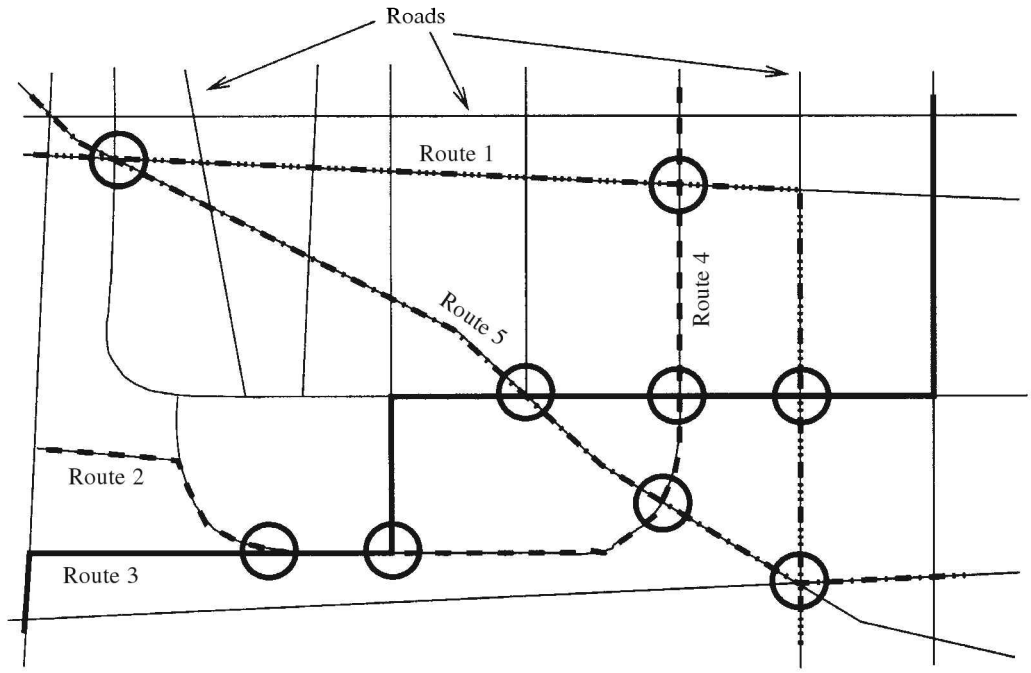


Figure 6.7 A schematic of a typical urban transit system route network.

MATHEMATICAL FORMULATION

Minimize

$$\sum_m \sum_i \sum_j \sum_k \sum_l \delta_{i,j,m}^{k,l} (d_{j,m}^l - a_{i,m}^k)\, \omega_{i,j,m}^k + \sum_m \sum_l \sum_i \int_0^{a_{i,m}^k - a_{i,m}^{k-1}} v_{i,k,m}(t)(a_{i,m}^k - a_{i,m}^{k-1} - t)\, dt \tag{6.26}$$

subject to

$$g_1 \equiv d_{i,m}^k - a_{i,m}^k \leq s_i^{\max} \qquad \forall i,\ m,\ k$$

$$g_2 \equiv d_{i,m}^k - a_{i,m}^k \geq s_i^{\min} \qquad \forall i,\ m,\ k$$

$$g_3 \equiv d_{j,m}^k - a_{i,m}^k + M(1-\delta_{i,j,m}^{k,l}) \geq 0 \qquad \forall i,\ j,\ k,\ l,\ m,\ j \neq i$$

$$g_4 \equiv \sum_l \delta_{i,j,m}^{k,l} = 1 \qquad \forall i,\ j,\ k,\ l,\ m,\ j \neq i$$

$$g_5 \equiv a_{i,m}^{k+1} - a_{i,m}^k \leq H_i \qquad \forall i,\ m,\ k$$

$$g_6 \equiv a_{i,m}^{k+1} - a_{i,m}^k > 0 \qquad \forall i,\ m,\ k$$

$$g_7 \equiv (d_{j,m}^l - a_{i,m}^k)\delta_{i,j,m}^{k,l} \leq T \qquad \forall i,\ j,\ k,\ l,\ m,\ j \neq i$$

$$g_8 \equiv a_{i,m}^k - d_{i,m-1}^k \leq t_{i,m-1,m}^{\max} \qquad \forall i,\ m,\ k$$

$$g_9 \equiv a_{i,m}^k - d_{i,m-1}^k \geq t_{i,m-1,m}^{\min} \qquad \forall i,\ m,\ k$$

In Eq. (6.26), the notation used is as follows:

$a_{i,m}^k$: Arrival time of the kth transit unit of the ith route at the mth transfer stop.

$d_{i,m}^k$: Departure time of the kth transit unit of the ith route from the mth transfer stop.

H_i : The policy headway for the ith route.

M : An arbitrarily large positive number.

$s_i^{\min}$: The minimum stopping time of a transit unit on the ith route at any stop.

$s_i^{\max}$: The maximum stopping time of a transit unit on the ith route at any stop.

$t_{i,m_1,m_2}^{\min}$: Minimum travel time of a transit unit on the ith route from transfer stop m_1 to m_2.

$t_{i,m_1,m_2}^{\max}$: Maximum travel time of a transit unit on the ith route from transfer stop m_1 to m_2.

T : Maximum transfer time for any transferring passenger.

$v_{i,\,k,\,m}(t)$: Arrival pattern of passengers for the kth transit unit of the ith route at the mth transfer stop.

$\delta_{i,j,m}^{k,l}$: Binary variable that is 1 if transfer from the kth transit unit of the ith route to the lth transit unit of the jth route at the mth transfer point is ideal to passengers; 0 otherwise.

$\omega_{i,j,m}^k$: Number of passengers who want to transfer from the kth transit unit of the ith route to the jth route at the mth transfer point.

The first term in the objective determines the total transfer time, TT, of all the passengers while the second term determines the total initial waiting time, IWT, (i.e. the sum of the waiting times of passengers at their origin stop) of all the passengers whose origin stop is one of the transfer stops. Note that the initial waiting time of the passengers who come at stops which are not transfer stops are not explicitly taken into consideration here. However, under the assumption that the arrival patterns of passengers at all the stops are similar, it can be said that minimizing the initial waiting time of passengers at any stop automatically minimizes the initial waiting time of passengers at all stops of the route.

The constraint sets g_1 and g_2 state that the stopping time of a transit unit at a stop must lie within a range—it cannot be too little or too large. The constraint set g_3 states that a person cannot transfer to a transit unit which has left the transfer stop before the arrival of that person on another transit unit. The constraint set g_4 states that a person can transfer to only one transit unit. These two constraint sets together with the objective function ensure that a person transfers to the next available transit unit (on the route for which the passenger is waiting) at the transfer stop (i.e. $\delta_{i,j,m}^{k,l} = 1$ only for such a combination of i, j, k, and l at the mth transfer stop). The constraint sets g_5 and g_6 state that headway on a given route must be positive and less than the policy headway for that route. The constraint set g_7 states that the scheduling should be such that no person has to wait more than T units of time for a transfer. The constraint sets g_8 and g_9 state that the arrival time of a transit unit at a transfer stop is dependent on the arrival time of the same transit unit at the previous transfer stop and the travel time between the transfer stops.

The above formulation is a nonlinear (constraint set g_6 and the objective function are nonlinear), mixed integer (the variables $\delta_{i,j,m}^{k,l}$ are integer variables while the others are real variables) programming problem. The traditional optimization methods are not adequate to solve such problems for any realistic size. Further, even though the formulation in Eq. (6.26) leads to a complicated optimization problem, the formulation is simplistic because it assumes that (i) the transit unit capacity is much greater than the demand and (ii) the arrival times of transit units are deterministic, whereas in reality it is stochastic. Incorporating these features into a mathematical programming formulation in a realistic manner is almost impossible. This is because (i) incorporating limited transit unit capacity would imply that queues be maintained at every stop (so that a correct picture of the waiting times can be obtained) and (ii) incorporating stochastic arrival times, especially in the presence of transfers, would imply that arrival time distributions be represented in the formulation. Although traditional methods may not be able to solve such optimization problems, some new optimization tools, like genetic algorithms have been successful in solving these problems. The interested reader may refer to Deb and Chakroborty [49] for a detailed description of these procedures.

There are also a number of heuristic- and simulation-based approaches which are used to develop schedules for transit systems. The reader may refer to Rapp and Gehner [189] and Bookbinder and Desilets [16] for detailed descriptions of such approaches.

EXERCISES

1. Show that the expression given in Eq. (6.25) is also valid for the type of arrival pattern shown in Figure 6.6(a) with all other assumptions remaining the same as those made in the derivation of Eq. (6.25).
2. Show that if the number of transit units on various routes going through a transfer stop are the same then the headway which minimizes the initial waiting time of passengers at the stop also minimizes the transfer time of the transferring passengers at that stop.
3. Show that the expression for the travel time of a transit unit obtained in Eq. (6.10) is valid only if the interstop spacing is greater than or equal to the $(v^2/2a)+ (v^2/2b)$, where v, a, and b have the same meanings as they do in Eq. (6.8).
4. Assuming that Route 2 in Figure 6.1 is B–G–F–G–B, calculate all the quantities determined in Example 6.1. Routes 1 and 3 are the same as those shown in the figure; use the same origin–destination matrix.
5. Write a computer program which can determine ridership, riding time, and transfers for a given network, route set, and origin–destination matrix.
6. Prove the relation given in Eq. (6.8).
7. Prove the claim made in Eq. (6.14).
8. Derive Eq. (6.23) from Eq. (6.22).
9. Redo Example 6.2 for a total fleet size of (a) 40 buses and (b) 60 buses. For each part develop a departure schedule of buses from their terminus.
10. Modify the formulation of the multi-route scheduling problem given in Eq. (6.26) under the assumption that passengers can board a bus if and only if the departure time of that bus is more than t minutes after the arrival of the bus which the passengers want to board.
11. Capacity conditions are such that passengers cannot board the next available bus but have to wait for the next to next bus. Under this assumption, modify the formulation of the multi-route scheduling problem given in Eq. (6.26).

Capacity of Transit Systems

7.1 INTRODUCTION

In Chapter 6 various types of public transportation systems are described. Of these, the transit systems and the rapid transit systems are of particular interest to the transportation engineer as they move a large number of people across a network. As indicated earlier, the transit systems use the street network and, in general, share it with other modes of transport. Rapid transit systems (RTSs), on the other hand, always use dedicated ways. This chapter discusses the determination of the capacity of such transit systems. In the following two sections, it briefly describes the capacity calculations for an RTS and for a street transit system. The contents of this brief chapter are adapted from Vuchic [256]. The reader may refer to this source for a detailed discussion of this topic.

7.2 CAPACITY OF RAPID TRANSIT SYSTEMS

An RTS may be thought of as a train which has various compartments or cars. The entire train is referred to as a transit unit, TU. The TU runs on a dedicated line punctuated with loading and unloading locations called *stations*. The capacity of an RTS may be defined as the maximum number of passengers the system can move across a point in a specified period of time. This section describes the various aspects of calculating this capacity under three subsections.

7.2.1 Line Capacity of RTS

The capacity of an RTS is often referred to as the *line capacity* of the RTS. The line capacity C_ℓ in passengers spaces[1] per hour is given by

$$C_\ell = nC_c \frac{1}{h_\ell^{\min}} \tag{7.1}$$

[1]A passenger space may be defined as the space which can be occupied or used by a passenger.

where

n is the number of cars or compartments per transit unit

C_c is the passenger holding capacity of each car, or the number of passenger spaces per car

$h_\ell^{\min}$ is the minimum headway (in hours) at which transit units can run in the system being considered.

Therefore, in order to calculate the capacity we have to determine what the minimum headway will be for a given RTS. As stated earlier, the transit unit in an RTS runs on a dedicated line or dedicated way with stations on them. The minimum headway at which TUs can run is determined from two considerations: (i) the minimum headway at which TUs can run safely on the dedicated line if there were no stations and (ii) the minimum headway at which TUs can run safely given that the TUs have to stop at the stations. The former minimum headway is called *minimum way headway*, $h_w^{\min}$, and the latter is called *minimum station headway*, $h_s^{\min}$. Obviously,

$$h_\ell^{\min} = \max\left\{h_w^{\min}, h_s^{\min}\right\} \tag{7.2}$$

The way in which $h_w^{\min}$ and $h_s^{\min}$ quantities are determined is discussed in the following two subsections.

Minimum way headway, $h_w^{\min}$, of RTS

Consider the schematic of two transit units shown in Figure 7.1. The figure shows the position of two TUs initially moving at a speed of v m/s and then coming to a stop. The problem at hand is to determine the value of the gap g between the TUs, such that if the leading TU, LTU, comes to a stop and remains stopped then the following TU, FTU, can also come to a stop at least a distance s_o behind LTU. Two points may be noted here: (i) the FTU can only react to the braking of the LTU after realizing that the LTU has started decelerating; the time taken during this perception–reaction process is referred to as the perception–reaction time, t_r; and (ii) the distance s_o is referred to as the buffer distance.

The minimum distance g that must be maintained such that the above criterion is met is given by

$$vt_r + \frac{v^2}{2b_2} - g = \frac{v^2}{2b_1} - s_o \tag{7.3}$$

or

$$g = vt_r + \frac{v^2}{2b_2} - \frac{v^2}{2b_1} + s_o \tag{7.4}$$

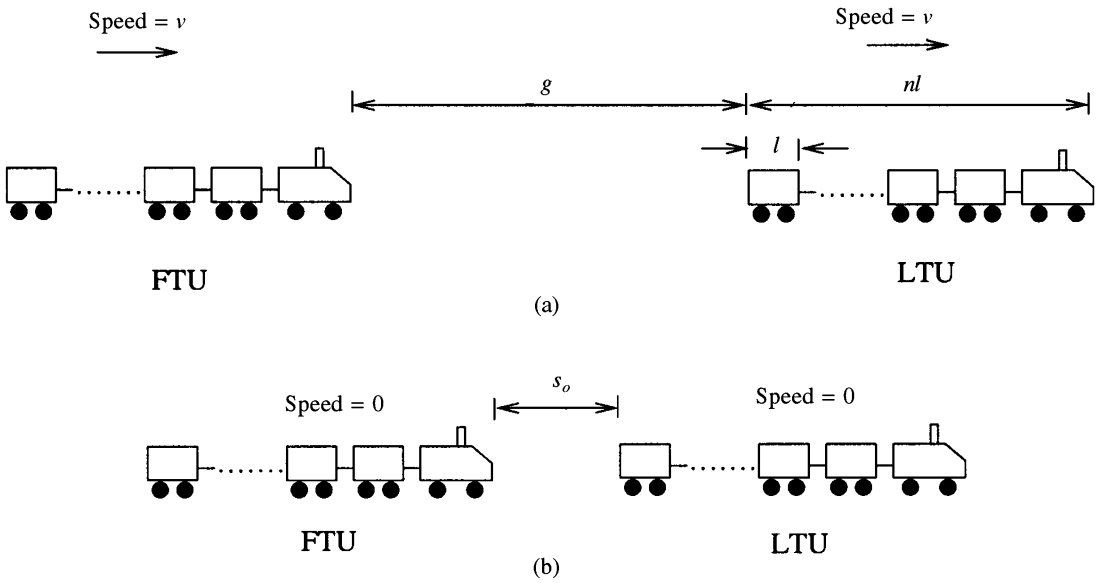


Figure 7.1 Schematic of two TUs: (a) FTU moving at a speed v and at a distance g from LTU, (b) both FTU and LTU stopped at a distance s_o apart.

where

v is the speed at which both the TUs are travelling when the LTU decides to stop

t_r is the perception–reaction time of the driver of FTU

s_o is the final buffer distance

b_2 is the deceleration rate of FTU

b_1 is the deceleration rate of LTU.

Now, we can determine the minimum distance headway (which is the distance between the corresponding points of the TUs) that must be maintained between the TUs. The minimum distance headway $d_w^{\min}$ is given by

$$d_w^{\min} = g + nl = vt_r + \frac{v^2}{2b_2} - \frac{v^2}{2b_1} + s_o + nl \tag{7.5}$$

This distance translated into the following minimum time headway, gives

$$h_w^{\min} = \frac{d_w^{\min}}{v} = t_r + \frac{v}{2b_2} - \frac{v}{2b_1} + \frac{s_o + nl}{v} \tag{7.6}$$

We now see that the minimum headway is a function of b_1 and b_2, the deceleration rates of LTU and FTU, respectively. Obviously, therefore, in order to determine $h_w^{\min}$ we need to make judicious assumptions about these values. In any transit system operation,

generally there are three different levels of deceleration rates that are considered: (i) the normal deceleration rate, b_n—this is the maximum deceleration rate which can be used under normal circumstances without causing discomfort to the passengers, (ii) the emergency deceleration rate, b_e—this is the maximum deceleration rate which can be used by the TU. The use of such a rate generally causes substantial discomfort to the standing passengers as well as bearable discomfort to the sitting passengers; it is used only under extreme conditions, and (iii) the infinite deceleration rate, or *stonewall*—this rate is considered only for LTU assuming that it may meet with an accident where it would stop almost immediately; obviously this is not a rate which can be consciously used for deceleration.

In order to analyze the $h_w^{\min}$ values, three different combinations of b_1 and b_2 are generally studied:

(a) **Safety Regime A.** In this case, the following assumptions are made: $b_1 = \infty$, $b_2 = b_n$. This assumption implies that a system operating under these conditions has absolute safety and comfort. However, the design made under this assumption often leads to inefficient use of the resources.

(b) **Safety Regime B.** In this case, the following assumptions are made: $b_1 = \infty$, $b_2 = b_e$. This assumption implies that a system operating under these conditions sometimes has to sacrifice comfort to get absolute safety.

(c) **Safety Regime C.** In this case, the following assumptions are made: $b_1 = b_e$, $b_2 = b_n$. This assumption implies that a system operating under these conditions has absolute comfort although the safety levels achieved are not foolproof.

The values of $h_w^{\min}$ for the safety regime i, $h_{w,i}^{\min}$, can be easily obtained from Eq. (7.6) as

$$h_{w,A}^{\min} = t_r + \frac{v}{2b_n} + \frac{s_o + nl}{v} \tag{7.7}$$

$$h_{w,B}^{\min} = t_r + \frac{v}{2b_e} + \frac{s_o + nl}{v} \tag{7.8}$$

$$h_{w,C}^{\min} = t_r + \frac{v}{2}\frac{b_e - b_n}{b_e b_n} + \frac{s_o + nl}{v} \tag{7.9}$$

Interestingly, and somewhat counter intuitively, it can be shown that if $b_e > 2b_n$ then $h_{w,C}^{\min} > h_{w,B}^{\min}$.

Minimum station headway, $h_s^{\min}$, of RTS

The minimum headway that needs to be maintained from the standpoint that TUs have to come to a stop at a station is described in this subsection through Figure 7.2. In the figure, the distance-time plots of the front and rear of both the LTU and FTU are shown.

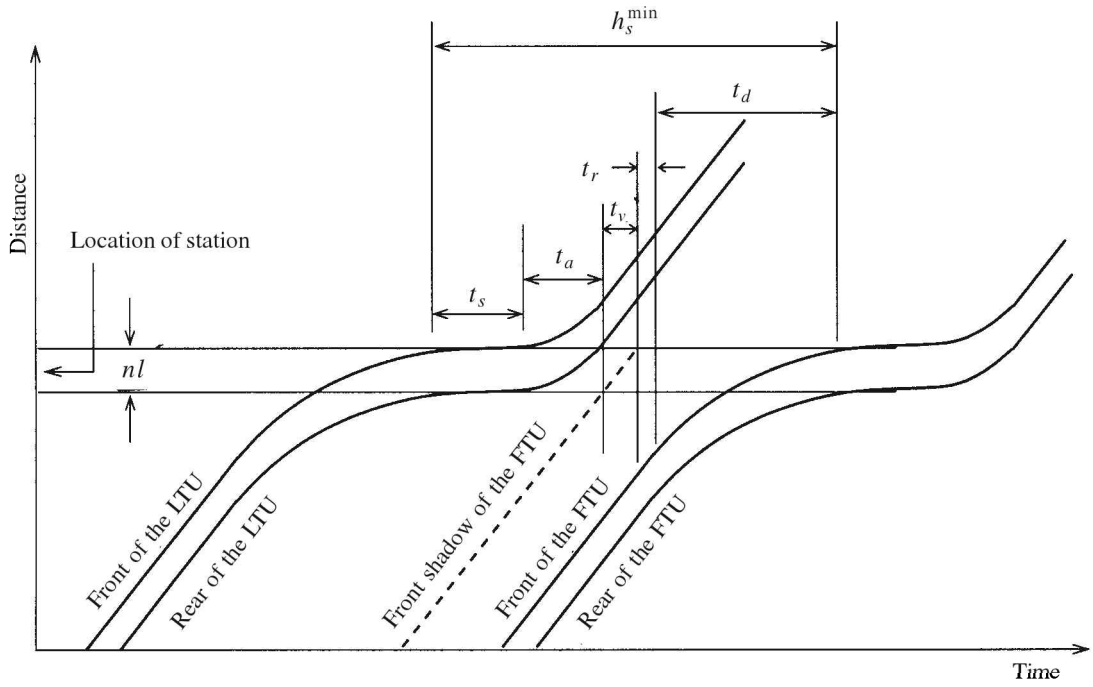


Figure 7.2 Distance–time diagrams of two consecutive TUs coming to a stop at a station.

The distance between the distance–time plots of the front and rear of either the LTU or the FTU is nl, the length of the TU, and the slope of the straight sections of the plots is v m/s, i.e. the speed at which the TUs are running.

Ahead (in time) of the front of the FTU is a broken line referred to as the *front shadow* of the *FTU*. The *front shadow* at any given time, say t, indicates the distance at which the front of the FTU will eventually come to a stop if the FTU is instructed to come to a stop at time t. This means that the distance of the *front shadow* from the distance–time plot of the front of the FTU will be equal to the stopping distance required by the FTU from an initial velocity of v m/s. That is, the above distance will be equal to $vt_r + (v^2/2b_2)$, where t_r is the perception–reaction time of the driver of the FTU and b_2 is the rate of deceleration employed by the FTU to come to a stop.

In order to determine the h_s^{min}, we need to find out, given the distance–time plot of the LTU, how close the distance–time plot of the FTU can be to the distance–time plot of the LTU. Since the two TUs should not occupy any portion of the station (or platform) at the same time, the *front shadow* can cross the platform (indicated by the lower horizontal line in Figure 7.2) only after the distance–time plot of the rear of the LTU has crossed the other end of the platform (indicated by the upper horizontal line in Figure 7.2). Therefore, the closest the two sets of distance–time plots can be is when the *front shadow* crosses one end of the platform at the same time as the distance–time plot

of the rear of the LTU crosses the other end of the platform. This situation is shown in Figure 7.2. Hence, the time difference between the time at which the LTU comes to a stop and the time at which the FTU comes to a stop indicates the value of the $h_s^{\min}$. Therefore,

$$h_s^{\min} = t_s + t_a + t_v + t_r + t_d \tag{7.10}$$

Here, t_s is the stopping time of a TU at the station, t_a is the time taken by the LTU to completely cross the station while accelerating from a stopped position at a constant rate of a m/s^2, t_v is the time duration as shown in the figure (and is calculated later from simple geometry and the fact that the slope of the dotted line is v), t_r is the reaction time of the driver of a TU required to initiate the stopping process, and t_d is the time required by the FTU to come to a stop from an initial speed of v m/s at a constant deceleration rate of b_2 m/s^2 (b_2 is generally assumed as b_n). Hence,

$$t_a = \sqrt{\frac{2nl}{a}} \tag{7.11}$$

$$t_v = \frac{nl}{v} \tag{7.12}$$

$$t_d = \frac{v}{b_n} \tag{7.13}$$

Using the above, the relation for $h_s^{\min}$ can be re-written as

$$h_s^{\min} = t_s + t_r + \sqrt{\frac{2nl}{a}} + \frac{nl}{v} + \frac{v}{b_n} \tag{7.14}$$

Before leaving the discussion on capacity of RTS, two additional points as enumerated below may be noted:

- Although in the derivation of $h_s^{\min}$ the deceleration rate is assumed as b_n, in general, the rate applied for stopping at a station is less than b_n.
- It can be easily shown that $h_s^{\min}$ will always be greater than $h_w^{\min}$. Thus, the former always controls the line capacity. Most of the time, therefore, a station has more than one parking bay or platform so that more than one TU can be parked at the same time. This improves the line capacity by making the minimum line headway close to the minimum way headway.

7.3 CAPACITY OF STREET TRANSIT SYSTEMS

The concepts of line capacity, minimum way headway, and minimum station headway should hold even for a street transit system. However, owing to the fact that street transit systems (i) do not require any special structures for boarding or alighting and (ii) stop for a

shorter period of time at stops than do RTSs, minimum station headways are no longer necessary to be determined as they become irrelevant. Further, because such transit systems operate on streets which generally offer no special privileges in terms of right-of-way to the transit system, the calculation of minimum way headway of the type done earlier is also meaningless.

The line capacity of a particular route of a street transit system is $nC_c f_{\max}$ [see Eq. (7.1)] where, n is generally unity, and $f_{\max}$ is the maximum frequency at which the transit units run on the route. In general, the value of maximum frequency $f_{\max}$ at which the transit units are operated on a route is a policy decision determined by the round-trip-travel time of the route and the available fleet size. Chapter 6 has discussed these quantities at length and, therefore, these are not repeated here.

EXERCISES

1. Show that if $b_e > 2b_n$, then $h_{w,C}^{\min} > h_{w,B}^{\min}$.
2. Show that the difference between the minimum way headway given by the safest and most comfortable safety regime and the minimum station headway is proportional to the operating speed v. Assume the buffer distance to be zero.
3. Determine the speed at which the line capacity of an RTS is maximum. Assume that the RTS is operating under Safety Regime A. Determine the relation for the maximum line capacity.
4. Determine the line capacity of an RTS given that (a) the operating speed is 60 kmph, (b) the operation is under Safety Regime B, (c) the emergency and normal deceleration rates are 4 m/s^2 and 2.5 m/s^2, respectively, (d) the effective length of a car is 18 m, (e) the usable floor space per car is 40 m^2, (f) the space allocated for sitting passengers is 1.5 times the space allocated for standing passengers, (g) the space required per seat is 0.4 m^2, (h) the space required per standing passenger is 0.2 m^2, (i) there are 8 cars per transit unit, (j) the stopping time at stations is 5 minutes, (k) the reaction of drivers is 2 s, and (l) the buffer distance between standing TUs is to be 5 m.
5. Determine an expression for the speed at which (a) the minimum way headway is the least under all the safety regimes and (b) the minimum station headway is the least.
6. Show that for realistic values of s_o, $h_w^{\min} < h_s^{\min}$ for all safety regimes.

PART III

TRANSPORTATION PLANNING

Transportation Planning Process

8.1 INTRODUCTION

Transportation planning, as the name suggests, deals with the development of a comprehensive plan for the construction and operation of transportation facilities. In order to develop a good and an efficient transport facility, it is necessary to have a proper planning procedure in place. The planning process should be continuous and dynamic i.e. it should be sensitive to the continuing changes in the socio-economic needs, technology, and financial status of a state and its people. Further, the planning process should be rational (i.e. the process should be methodical) rather than political (which often tends to be the case), where decisions on construction and operation of transportation facilities are taken by political functionaries to gain short-term popularity. In such cases, more often than not, the development is piecemeal and crisis-mitigation oriented leading to largely inefficient use of the limited resources.

In this chapter, the rational transportation planning process is outlined. It first describes the various elements of such a process and their interactions. Next, these elements are discussed in greater detail. Throughout the chapter an example case is followed in order to illustrate how the process works.

8.2 ELEMENTS OF TRANSPORTATION PLANNING

A rational transportation planning process begins with the definition of the goals and objectives which are to be achieved through transportation and ends with the development of an implementation strategy of a particular course of action. The various elements of this process are described below:

Statement of goals and objectives. In this stage, as will be described later, the policy making body (often the local, state or national government) defines a set of objectives which it seeks to achieve through the development of transportation facilities.

Identification of needs. In this step, we identify the components of the transportation

system needed to satisfy the goals based on the present conditions and forecasts of future conditions of travel demand, fuel costs, etc.

Generation of alternatives. In this stage, the alternative plans are developed which address the needs identified in the previous stage.

Evaluation of alternatives. Each alternative developed in the previous stage, will have technical, land-use and demographic, environmental, financial, and economic implications. In this stage, each alternative is evaluated based on these implications.

Implementation of alternatives. Once an alternative is found to be suitable, a plan is drawn up as to how it will be implemented. The planning process stops here, and the project is carried forward by the implementing agencies identified in the implementation plan.

Figure 8.1 presents a schematic of the planning process showing the various stages described above. It may be pointed out here that often the alternatives developed in the first iteration are not found suitable. In these cases, either new alternatives are formulated or some of the previously suggested ones are adequately modified. These are then evaluated as before. The procedure continues till an acceptable alternative has been identified. In some cases (although this has not been illustrated in the diagram), the evaluation benchmarks may be changed if it is found that it is impossible to find an alternative which meets all the requirements. Sometimes even the needs may be looked upon differently (like reducing the planning horizon in order to design a pavement for less number of axle repetitions) if the initial estimates of needs lead to unacceptable alternatives (like building a very expensive road).

8.3 DEFINITION OF GOALS AND OBJECTIVES

Any country, state, or locality has (or should have) a vision for its future. This vision is generally articulated through the various elected bodies and may change from time to time as the aspirations of the people change. Such vision often, if not always, includes the improvement of the quality of life of the people of the country, state, or locality.

It is this notion of improvement of quality of life of the people that leads to goals and objectives relating to better mobility and connectivity which one wants to achieve. Goals and objectives to be achieved through transportation are often stated as policies to be pursued by the different elected agencies (and are to some extent driven by political compulsions). These may be as general as improving the connectivity of all villages to their nearest market (or improving the air quality of large cities) to as specific as reducing the travel time of vehicles on a certain section of a highway. For example, in December 2000, the government of India announced a policy (a statement of goal) to connect all unconnected villages in India by 2003 (each with a population more than 1000 persons) through roads which could be used in all weathers.

As stated earlier, throughout this chapter the following example case study will be built upon and studied in order to better understand the transportation planning process.

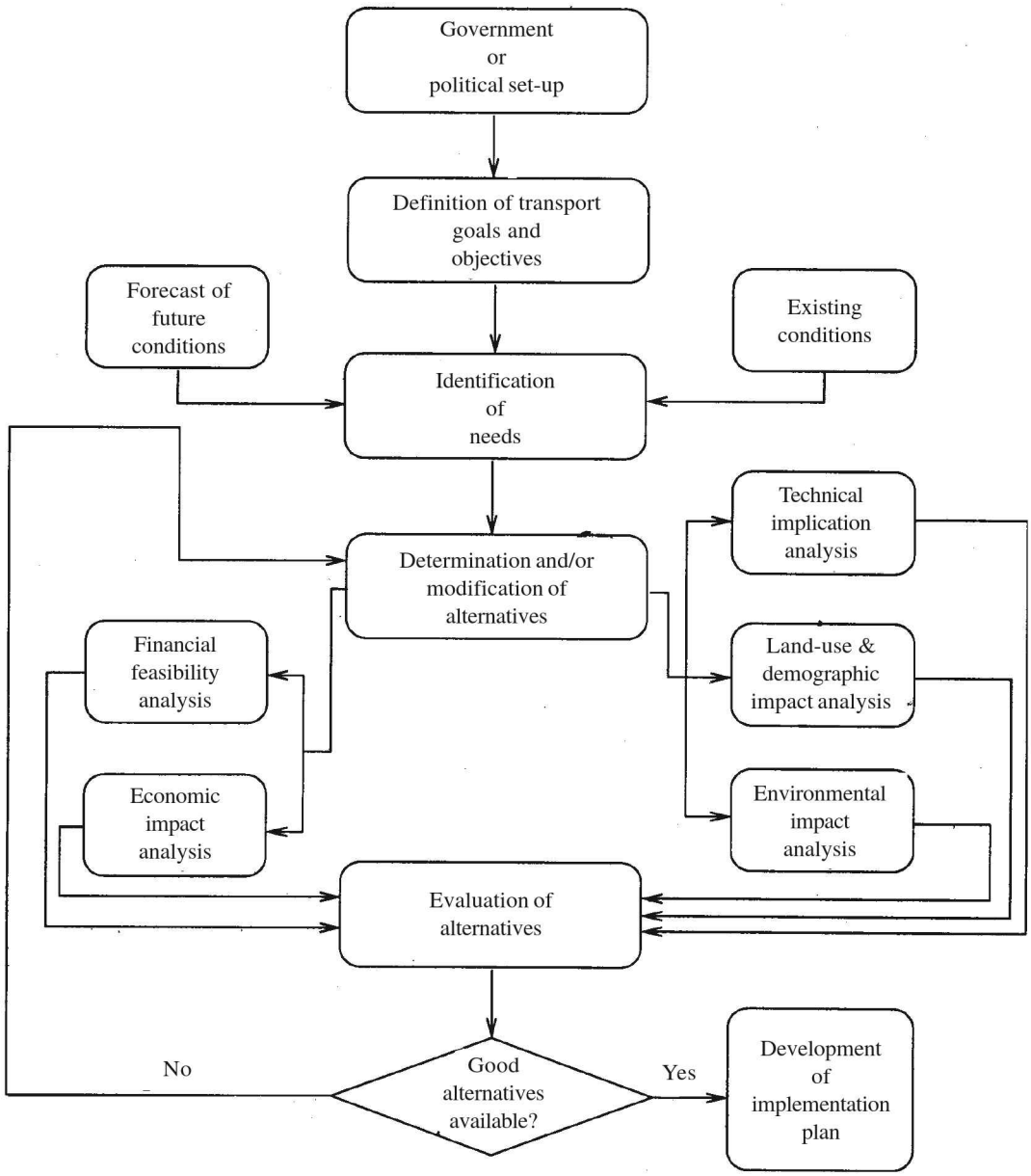


Figure 8.1 Schematic diagram of the transportation planning process.

Consider the case, where the government, based on the demands of the people of Kanpur and its adjoining areas, declares the objective of reducing the travel time of people between the cities of New Delhi and Kanpur (which is about 500 km south-east of New Delhi) so that people from Kanpur can travel to New Delhi and back on the same day. This example case is visited in every section to illustrate how the various stages of the planning process work.

8.4 IDENTIFICATION OF NEEDS

Once the goals have been stated, the next stage is to identify the needs of the transportation system that will satisfy the goals. Essential inputs to this identification process are an inventory of the existing state of the system and a forecast of the future conditions (especially demand). An implicit requirement of this process is the planning horizon—or the time period for which this planning is being done.

Consider again the example stated in Section 8.3. Based on the objective, we may define the needs at various levels. At the very basic level, we need to interpret the stated objective. First, obviously by a day we do not mean the entire 24 hours; rather it should be considered as a period of 16 hours starting from, say, 6:00 am. Second, it may not be worthwhile to travel to New Delhi and back if one does not get enough working hours in the capital city; hence here again we need to interpret the objective so as to assume that one should get at least 6 (if not 8) working hours in the city. These interpretations and an assumption that at least two hours (on an average) would be needed to travel to and from home/office to the transportation system terminus leaves one with 8 hours of travel time (for both ways). This then gives rise to the basic need: that of having a transport system in place which can cover 500 km in 4 hours.

We also need to look at the existing facilities available in Kanpur and the predicted demands for such a system in order to decide on certain specific needs. First, the existing facilities: (i) Although there are trains which can reach New Delhi in slightly less than 5 hours, none of them leave Kanpur in the morning; on the other hand, there are fast trains which leave New Delhi around 5:00 pm and stop at Kanpur, however, they are all long distance trains. (ii) Kanpur is not connected by air to New Delhi, although there is a small airport at Kanpur. (iii) The road connection from Kanpur to New Delhi at present takes about 10 hours. Next, we need to forecast the demand for such a transportation system over the next, say, 10 years. Such forecasts are made through travel demand analysis techniques which are the subject matter of Chapter 9. For the purposes of this example, we assume that on an average 600 passengers per day will presently use such a system and this will grow to about 1200 passengers per day in another 10 years.

Based on the existing facilities and the forecasts, we can say that there are no such existing facilities and that the need is to have a high-speed transport system capable of catering to about 1200 passengers per day when operating at full capacity.

8.5 GENERATION OF ALTERNATIVES

Based on the needs enumerated earlier, the transportation planners have to identify various alternatives which will match those needs. While determining the alternatives, a planner needs to look at (i) the various modes like roadways (either individual transport or public transport systems), railways, waterways, or airways, (ii) the technological aspects (such as high-speed trains, raised monorails, underground transit systems, driver information systems, and so forth), (iii) the traffic engineering aspects (such as changing or

improving the flow pattern in an area by making certain roads one-way, reducing delay on arterial streets by improving signalization or grade-separating intersections, disallowing certain movements at intersections, and so forth), and (iv) the regulatory aspects (such as reserving lanes for only high-occupancy vehicles, disallowing high polluting vehicles on the roads, imposing speed limits, and so on). Obviously, different aspects assume importance depending on the goals and needs.

For example, if the goal is defined as cleaner air, the planning instruments such as reserving lanes for high-occupancy vehicles, building better public transport systems (leading to a reduction in the number of vehicles on the road), coordination of signals, changing flow patterns (leading to reduced traffic congestion), and enforcing the use of only less polluting vehicles gain prominence. Whereas for the goals and needs of the proposed transport system connecting Kanpur and New Delhi, developing a high-speed public transport system assumes the utmost importance.

Specifically for the example at hand, the following alternatives could be suggested:

(i) Provide a train service leaving Kanpur at 6:30 am and reaching New Delhi at 10:30 am; provide a new train which will leave New Delhi at 5:30 pm and reach Kanpur in 4 hours (note that, given the high demand for such services, suggesting the use of one of the existing fast trains for the return trip is not tenable as they are all long distance trains and cannot provide adequate number of seats for Kanpur).

(ii) Develop the existing small airport in Kanpur to handle at least short-haul commercial flights.

(iii) Improve the existing road connecting Kanpur and New Delhi to provide average operating speeds of about 125 kmph.

8.6 EVALUATION OF ALTERNATIVES

Any alternative will have implications such as: (i) financially, (ii) economically, (iii) land-use wise and demographically, (iv) technologically, and (v) environmentally. Hence, an alternative has to be evaluated from each of the above five standpoints. A detailed discussion on these points is beyond the scope of this book, nonetheless, some idea is provided in the following text. Further, although the list is sequential, the interrelationship between one implication (or analysis) and the other is not; the interrelationship is indeed complex and can be gauged from a careful reading of the following.

Financial feasibility analysis. Most alternatives have some financial implications in this analysis. Therefore, the first objective is to determine how much a particular alternative will cost both in terms of capital investment and operating and maintenance costs and how the capital cost is structured, that is, whether the investment is phased over a period

of time or whether it has to made up-front. The second objective is to determine, whether it is possible to meet the costs of the alternative based on budgetary considerations (of governments, who are generally the primary investors in the transportation sector), loan availability, etc. If an alternative can be borne financially then it is financially feasible.

Consider the first alternative for the given example. In this case, the capital investment has to be made at the least on the following three accounts: (i) improving sections of rail tracks so that the high-speed train operation will not pose safety hazards, (ii) building new coaches which will be able to cope with the high-speed operation, and (iii) putting in place a more reliable automatic rail traffic control and collision avoidance system. Running costs in this case will primarily be due to: (i) higher maintenance cost because of lower tolerance on track quality (that is, even slight deterioration may mean replacement of affected track sections), (ii) more frequent maintenance checks, (iii) better maintenance of coaches, and the like. If the ministry of railways is able to (with or without the help of other agencies) fund this project then the alternative is financially feasible.

Economic impact analysis. The objective of this analysis is to assess the various economic effects of an alternative. The economic effects of an alternative include its impacts on (i) travel time of users, (ii) out-of-pocket cost to users, (iii) improved business opportunities, (iv) improved accessibility to various locations, and so forth. Some of the impacts may be beneficial while others may not be so.

Going back to the first alternative of the example, we can list the economic benefits that will be derived from the reduced travel time between Kanpur and New Delhi. The out-of-pocket cost, however, will be substantially greater than the prevalent rates since the ticket prices will definitely be higher. Accessibility to New Delhi, and its markets and facilities, from Kanpur will definitely improve and bring in a lot of other opportunities to Kanpur. The improved accessibility to New Delhi from Kanpur may, however, adversely affect the business of the Lucknow airport. Obviously, a complete economic impact analysis is a complex issue and requires a much better understanding of both principles of economics and transportation than is possible in a first-course on transportation.

Technical implications analysis. Each alternative assumes that its features are achievable technically. The technological requirements and their feasibility are the subject matters of this analysis. In the example being considered here, the first alternative assumes that it is technically feasible to have tracks and locomotives which can sustain an average speed of 125 kmph. Although it is technically feasible now, it would not have been so in the seventies (at least in India). Thus, this alternative would have been termed technically infeasible about thirty years ago, although it is considered technically feasible today.

Land-use and demographic impact analysis. Every transportation facility has an impact on the land-use and demographic characteristics of an area. The development of a road, for example, may change the land-use of an area from 'unused low land' to a 'bustling shopping complex', or a 'large hospital.' Changes in land-use may often lead to changes in the demographic statistics of an area. For example, building better connections between the city centre and the suburbs may see a shift of high-income group people from the city to the suburban locations.

Some of the land-use and demographic changes may be desirable while others may not be. We therefore, given an alternative, should analyze these changes so as to assess the impact of the alternative from the above standpoint. In the case of the example being looked at here, the first alternative (rail connection) will possibly lead to little or no land-use and demographic impact as the alternative does not really suggest building of a new facility (it is, however, assumed here that the proposed train service will not stop at any intermediate station). The second and third alternatives, however, may lead to some such changes. The revival of the airport may attract some business establishments close to the airport, while building a good road will make all places between Kanpur and New Delhi (along the road) more accessible to either of these cities. This increase in accessibility may see many offices shifting to the areas outlying these cities, as well as many residential complexes being built in these areas. These impacts are by no means a complete description of all the possible impacts that may happen due to the implementation of either of the last two alternatives.

Environmental impact analysis. Construction of any major transportation facility and its use has a direct impact on the environment. The impact on the environment is at many levels and only a simplistic view of this effect leads to the following three directions of impact: (i) the construction of the facility (say a road) may entail changing the natural features of the land which can lead to problems like loss of habitat for wildlife; (ii) the use of the facility (say by vehicles) may increase the air and noise pollution levels of the immediate surroundings; and (iii) a better facility may encourage frequent travel and hence greater use of the non-renewable fossil fuel (this is largely true at least in the present scenario).

The environmental impact of an alternative obviously varies with the alternative and some alternatives may in fact yield environmental benefits. For example, an alternative which encourages the use of the public transportation system will yield an environmental benefit through a reduction in vehicular traffic. In the present case study, the first alternative would hardly have any added (over and above what already exists) environmental impact. The second alternative (air connection) will have some impact in terms of increased fuel usage and a larger impact in terms of noise pollution. The third alternative (better road connection), on the other hand, will have the largest environmental impact and possibly on all the three accounts.

8.7 IMPLEMENTATION OF ALTERNATIVES

Once the implications of all the alternatives have been evaluated, these alternatives can be rated from the most advisable to the least. The most advisable alternative can then be chosen for implementation. The task of planning ends with the development of a detailed road map of how the alternative will be implemented. The implementation strategy should identify or put in place a method of identifying (i) the agencies and instruments that will be used to finance the project, (ii) the agencies or organizations that will implement the project on the ground, (iii) the agencies to be entrusted with maintaining and operating the facility, and (iv) the operating strategy.

As stated earlier, the end of the planning process signals the start of the implementation process. There still remains a considerable gap between this stage and the final stage (which is when the alternative has been fully implemented and can be used by the public at large). This is because, at the planning phase, we take a reasonably macroscopic view of an alternative while the implementation details are much more microscopic and may range from solving problems of getting electricity to the site to quality control of the work on a day-to-day basis.

In a nutshell, this chapter has presented a bird's-eye view of what may be referred to as the rational (or quantitatively biased) planning process. It has largely ignored (primarily because of the scope of this book) the qualitative and political nature of the planning process. The interested reader may refer to Banister [10] for a more detailed understanding of the planning process. Nonetheless, the chapter illustrates the challenges and importance of the planning process. The chapter also highlights the fact that transport planning is a multidisciplinary process and a good planning organization should have expertise ranging from finance, economics, environmental and ecological science, to engineering.

The next chapter describes one of the important stages in the planning process—the forecasting of travel demand.

EXERCISES

1. Assume that in your city (or town) it has been stated that the objective of the government is to reduce the congestion on the city (or town) streets. Go through the steps of the rational transportation planning process in order to develop a workable solution.
2. Often in the past, the transportation planning policies have been dictated by laws and court verdicts. Cite a few examples of this in India.
3. Sometimes, the transportation planning process gets affected to a large extent by political considerations. Cite a few examples of this in India.

4. Name the primary planning authorities (at various levels) in India.
5. Describe some of the important transport-related policies that have been stated and implemented in post-independence period of India. Write a report on one of these policies and how (or whether) its implementation has achieved the stated goals.
6. Write a report on the effects of construction of the ring roads in New Delhi on land-use and demographic patterns of the city. Write a similar report on the effects of constructing the eastern metropolitan bypass in Kolkata.

Transportation Demand Analysis

9.1 INTRODUCTION

This chapter introduces the topic of *transportation demand* and its determination. Transportation demand, simply stated, is the demand for trips that exists in any area. All of this demand, however, may or may not materialize into physical trips (vehicular or pedestrian)—and some of it generally remains latent and is referred to as *hidden demand*. The importance of analyzing the transportation demand in order to be able to predict the expected number of trips in a given network cannot be overstated. The demand for transportation forms the primary input to any decision related to creation and management of transportation and traffic facilities, such as roads, intersections, parking lots, transit system, and so on.

This chapter first describes the nature of transportation demand and how it can be analyzed. Then, it presents the *sequential demand analysis technique*—the most frequently used method of determining the transportation demand. Finally, it briefly describes some of the data collection mechanisms employed in travel demand analysis.

9.2 NATURE AND ANALYSIS OF TRANSPORTATION DEMAND

Transportation demand, unlike the demand for other commodities, such as wheat, coffee, housing, clothing, etc. is a *derived demand*. That is, a person demands to be transported not because he/she just wants to move (except for those rare cases when the person goes out for a joy ride!) but because he/she wants to achieve some other purpose such as reaching school, or office, or a movie theatre. In other words, the need for achieving some goal (such as reaching office or a shop) creates the need to travel. Hence, travel demand is primarily generated by the population's need to work, entertain (themselves), socialize, study, etc. Therefore, it is not surprising that two of the major aspects in travel demand analysis are *land-use* and *trip-purpose*.

Land-use refers to the pattern of land usage in an area. Land-use affects the transportation demand through generation and distribution of trips. The effect of land-

use on transportation demand is not necessarily a one-way effect but rather a part of cycle in which land-use changes transportation needs which in turn change land-use. Figure 9.1 shows a simple schematic of how land-use and transportation demand are related.

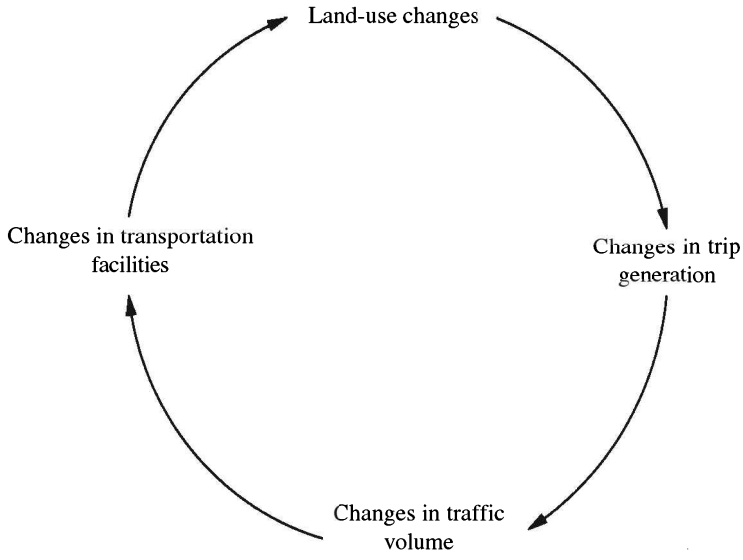


Figure 9.1 Relationship between land-use and transportation demand.

Trip-purpose refers to the purpose for which the trip is being undertaken. Travel demand behaviour changes with the trip-purpose. For example, a person hardly exercises any choice for work trips, i.e. does not necessarily decide every time whether to go to work or not. A person obviously does not decide where to go to work (generally it is fixed over a period of time for a large section of the population), even the choice of route and mode are not daily decisions. On the other hand, for recreational trips, an individual makes a large number of decisions, such as whether to go or not, where to go, and how to go. Consequently, the travel demand behaviour for work trips varies considerably from that of recreational trips. This example, can obviously be extended to other types of trips such as shopping trips, etc. Given the effect of trip-purpose on travel demand behaviour, the analysis of travel demand is done separately for different trip-purposes.

Although the above discussion throws light on some of the factors which affect travel demand, some more understanding of travel demand is necessary before we can analyze the demand and can, with some degree of confidence, predict the volume on various links of a network. Generally, a trip (which is the basic quantity in travel demand) materializes after the trip-maker makes certain decisions. These decisions can be broadly classified as follows:

The decision to travel. The trip-maker, given his/her requirements, makes a conscious decision to travel so that the requirements can be met.

The decision on the choice of destination. The trip-maker also makes a decision as to where he/she wants to go; for certain kinds of trip-purpose, such as work trips, this decision may not exist; yet for other kinds of trips, such as shopping trips, there may be certain alternative locations to choose from.

The decision on the choice of mode. The trip-maker also takes a decision as to what mode of transport to use for a given trip. This decision, however, is only available to those who have access to different modes and are not *captive* users of any particular mode.

The decision on the choice of route. The trip-maker on any given trip takes a definite decision on which route to take so as to reach the destination. Again, this decision is available to only those trip-makers who have access to modes which can use different routes as per the wishes of the trip-maker. Such modes would generally include personal automobiles or two-wheelers.

Although there is unanimity on the fact that the above decisions can aptly capture the entire trip-making behaviour of an individual and hence can be used to analyze travel demand pattern of an area, it is difficult to ascertain whether there exists any definite sequence in which these decisions are made. Generally it is assumed, primarily for the ease of analysis rather than anything else, that the decisions are made in a strict sequence as shown in Figure 9.2. Analysis techniques which assume that such a sequence exists are referred to as *sequential demand analysis techniques*.

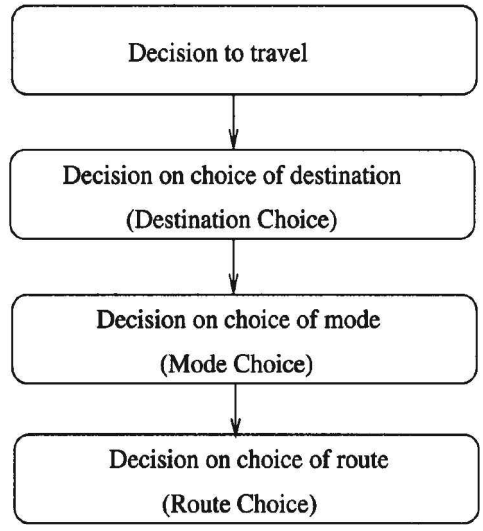


Figure 9.2 Schematic representation of the assumptions of sequential decision making.

Although even today transportation demand is analyzed sequentially, the assumption that the four major decisions (see Figure 9.2) of a trip-maker follow a strict sequence (i.e. are in a series) is possibly not the most appealing. Quite often, the *decision to travel* is changed because an appropriate destination does not exist; or an *initial choice of destination* is changed because it cannot be reached by the desirable mode of transport. It is possibly a more truer picture of reality if the decision-making framework is assumed to have feedback loops. One such possible structure is shown in Figure 9.3. In this structure, unlike in Figure 9.2, there are feedback loops indicating that decisions taken earlier can be changed based on a latter decision. For example, the decision to travel may be aborted at the *mode choice* stage if it is realized that none of the available modes suit the requirements.

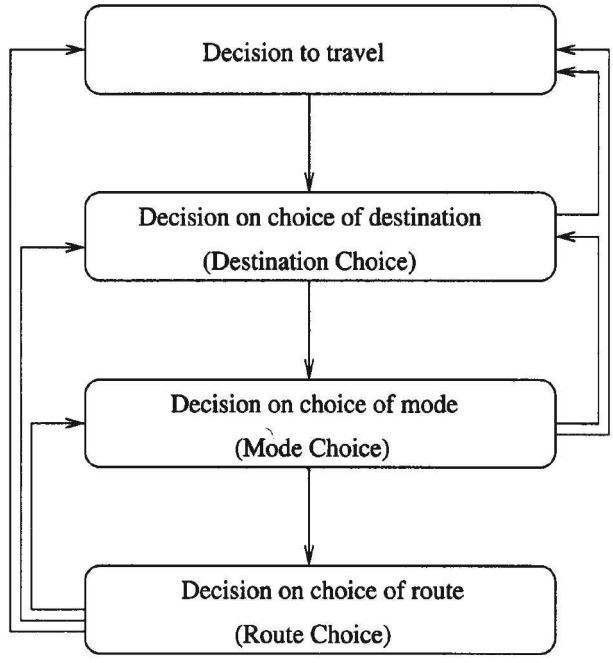


Figure 9.3 An example of the assumptions of non-sequential decision making.

9.3 SEQUENTIAL DEMAND ANALYSIS

As stated earlier, even though the assumption of the existence of a strict sequence in the decision-making process of a potential trip-maker may be debatable, generally *sequential demand analysis* is used to determine the travel demand. As will be seen throughout this section, even with this simplifying assumption of sequential decision making, the analysis of transportation demand remains sufficiently complex.

Figure 9.4 shows a schematic of the sequential demand analysis procedure. The figure attempts to not only illustrate the logic of the analysis procedure but also gives the names of the different classes of models used to mathematically describe each decision-making phase of the analysis procedure. Each of these classes of models is described in detail in the next four subsections.

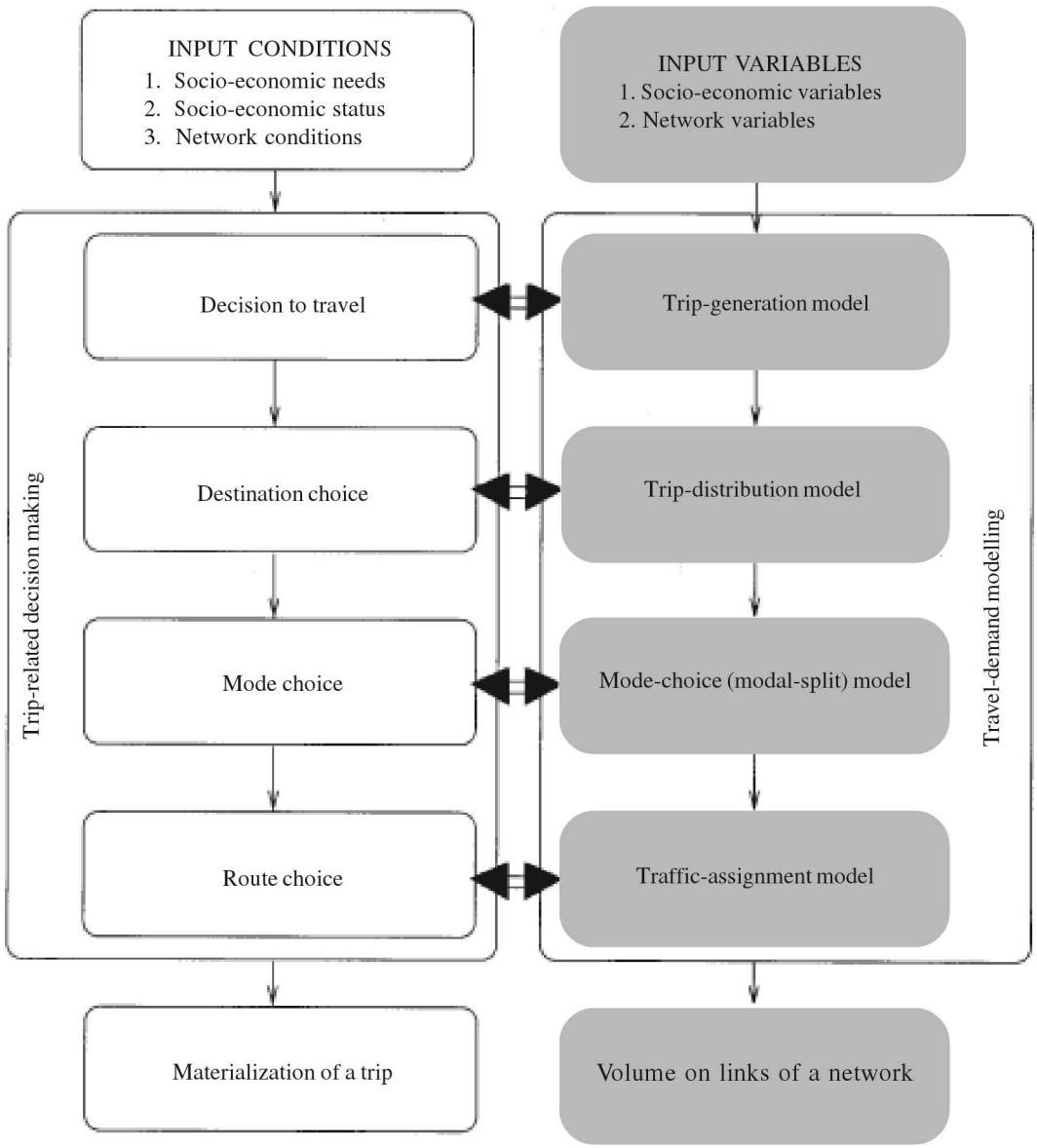


Figure 9.4 Sequential demand analysis procedure.

Before presenting the models, a general overview of the entire process is provided. In this analysis procedure, first the entire study area is divided into various zones. These zones are generally obtained from the land-use pattern of the area. Next, for each zone the total number of trips generated in that zone are estimated using the *trip-generation models*. The outputs of the trip-generation models are then used to determine the number of trips between all zone pairs using the *trip-distribution models*. Given the trip-distribution pattern (which is often referred to as the *origin–destination (O–D) matrix*), the relative shares of these trips for the different modes are estimated using the *mode-choice models*. Finally, the *traffic-assignment models* estimate the volume on each link of the network by determining the routes that will be used by the trips.

9.3.1 Trip-generation Models

The trip-generation models strive to predict the number of trips generated by a zone. These models try to mathematically describe the *decision-to-travel* phase of the sequential demand analysis procedure. It may be mentioned here that typically the term *trip-generation* is used to mean *trip production—generally the trips made from households*—and *trip attraction—the trips made to a particular urban location or activity*. However, it is felt that analysis of trip attractions should not be within the purview of trip-generation models which attempt to quantify a population's urge or propensity to travel. Rather, trip attractions are an outcome of the *destination-choice* phase of travel behaviour. Similar concerns about trip attractions being part of the trip-generation phase of urban demand analysis have been also voiced in Kanafani [132]. In keeping with this, the present section discusses trip-generation primarily in the context of trip productions. Trip attractions are assumed to be an outcome of the destination-choice phase and are discussed in the section on *trip-distribution models*.

There are basically two different model structures for trip-generation models—the *cross-classification models* and the *regression models*. However, both these model structures incorporate the same basic factors which affect the trip-generation of a zone; the models only differ in their characterization of these factors.

The factors (for any given trip-purpose) which affect the trip-generation of a zone are:

- The number of potential trip-makers in the zone; this data could be captured by variables like residential density, average household occupancy, age distribution of occupants, and so forth.
- The propensity of a potential trip-maker to make a trip; this is related to automobile ownership, accessibility to public transportation, and the like. For example, persons who own automobiles make more non-work trips than those who do not own automobiles.
- Accessibility of the zone to potential destinations for a given trip-purpose satisfaction; the variables like distance to potential destinations can capture this

factor. For example, persons who live close to various recreational facilities may make more number of recreational trips than those who live in areas which do not have nearby recreational facilities.

The cross-classification model

The cross-classification model, sometimes referred to as the category-analysis model, is based on the assumption that the number of trips generated by *similar* households or households belonging to the same *category* is the same. According to this model, if in Zone i there are n_k^i households in category k and if g_k is the average rate of trip-generation per household in category k then the relation for the trips generated (or produced) by Zone i, T_i, is given by

$$T_i = \sum_{\forall k} n_k^i g_k \tag{9.1}$$

The model predicts the trips produced by a zone by simply aggregating the total trips produced by all the households in that zone. However, two basic questions need to be answered here: (i) how do we define *similar* households, or alternatively how do we define categories of households, and (ii) how do we determine the rate of trip-generation for a given category of households. The answer to both these questions is: *through empirical observations and analysis*. What is done is that, first, data on demographic characteristics and trip-making behaviour of a large number of households are collected. This data is then analyzed to see what characteristics of the households are important in defining a homogeneous group—the households which produce approximately the same number of trips.

Based on the above analysis, tables are made which define each category of households by listing its properties in terms of different demographic variables. For example, a particular category of households may be defined as households with 3–4 members in the age group 6–60, with income in the range of Rs 30,000–40,000 per month, and one automobile. Finally, for each category of household the average number of trips generated is listed. The listing of the definition of categories and the associated trip-generation rates are generally referred to as *trip tables*.

This method of analysis although simple in its structure has few difficulties. The foremost is the problem with defining categories correctly—at best it is very difficult. There are other problems like handling additional data on trip-generation behaviour—the trip tables are not amenable to simple updating but generally have to be completely revamped every time new data is available.

Regression model

In this model, an additive functional form is assumed to exist between the factors which affect trip-generation and the number of trips generated. Generally, a linear function of the following form is used:

$$T_i = \alpha_1 z_{1,i} + \alpha_2 z_{2,i} + \cdots + \alpha_n z_{n,i} \tag{9.2}$$

where α_k are the parameters of the regression function and $z_{k,i}$ is the value of the kth variable (such as income, automobile ownership, number of members in a household, and the like) for the ith zone.

As can be seen using this model, determining the number of trips generated by a zone is a simple matter when all the parameters of the regression function are known. Obviously, the parameters are determined by using some parameter estimation techniques such as *Ordinary Least Squares* or *Maximum Likelihood Technique* on empirically obtained data on z_k variables and T_i. For a good description of the regression analysis and the parameter estimation techniques mentioned here, the reader may refer to any book on introductory statistical methods or basic econometrics; for example, Gujarati [95] and Wonnacott and Wonnacott [263].

Discussion. Generally, the models of trip-generation include variables which reflect the number of potential trip-makers and the propensity of potential trip-makers to make a trip. However, none of the present models incorporate variables which reflect the accessibility factor. This is possibly the single largest factor as to why trip-generation models cannot very well predict the number of trips generated.

9.3.2 Trip-distribution Models

The trip-distribution models strive to predict the number of trips that will be made between a pair of zones for a particular trip-purpose. These models try to mathematically describe the *destination-choice* phase of the sequential demand analysis procedure. There are various models of trip distribution. However, most of them incorporate the same basic factors which affect the number of trips between an origin zone and a destination zone. The models differ in their characterization of these factors and in the way these factors are assumed to affect the trip distribution.

The factors (for any given trip-purpose) which affect the number of trips between two zones are:

- The number of trips produced by the origin zone.
- The degree to which the in-situ attributes of the destination zone attract trip-makers. The attributes which gain importance vary with the trip-purpose. For example, if one is modelling the number of shopping trips attracted to a zone then the type of attributes of the zone which assume importance will be the total shopping floor area, the number of retail outlets, and the like. On the other hand, if one is modelling the number of work trips attracted to a zone then the type of attributes of the zone which assume importance will be the number of offices, the type of offices, and so forth.
- The factors that inhibit travel between a pair of zones. These factors could be, travel time, travel distance, travel cost, and so on.

Four different models are presented here. These are (i) Gravity model, (ii) Intervening opportunities model, (iii) Choice model, and (iv) Entropy model.

Gravity model

The gravity model uses the following basic form to determine the trips between an origin zone i and a destination zone j. That is,

$$t_{ij} = K_g^i \frac{f(a_j)}{h(d_{ij})} T_i \tag{9.3}$$

where

T_i is the total number of trips being produced by the origin zone i

a_j is a set of in-situ attributes of the destination zone j which attract trip makers

d_{ij} is a set of factors which inhibit travel between two zones

K_g^i is a calibration constant

$f(\cdot)$ and $h(\cdot)$ are positive monotonically increasing functions.

Note that $\Sigma_{\forall i}\, t_{ij}$ is basically the trip attractions of Zone j.

The expression $K_g^i \frac{f(a_j)}{h(d_{ij})}$ may be thought of as a factor which distributes the total trips produced by a zone among all the possible destination zones. In this sense, the sum of the expression over all destinations should be equal to unity. Thus,

$$\sum_j K_g^i \frac{f(a_j)}{h(d_{ij})} = 1 \tag{9.4}$$

The above equation implies that,

$$K_g^i = \left(\sum_j \frac{f(a_j)}{h(d_{ij})} \right)^{-1} \tag{9.5}$$

Substituting the expression for K_g^i in Eq. (9.3), the following relation, often referred to as the *origin–constrained gravity model*, is obtained:

$$t_{ij} = \frac{f(a_j)/h(d_{ij})}{\sum_j f(a_j)/h(d_{ij})} T_i \tag{9.6}$$

Sometimes, in the gravity model, the number of trips attracted to a zone (when such data is independently available), T_j, is used as a surrogate for $f(a_j)$. When such a

substitution is done, the gravity model is typically written as

$$t_{ij} = K_g^i K_g^j \frac{1}{h(d_{ij})} T_i T_j \tag{9.7}$$

Very often, in this case the following two constraints are imposed on the gravity model to obtain the two calibration constants.[1]

$$\sum_j t_{ij} = T_i, \qquad \forall j$$

$$\sum_i t_{ij} = T_j, \qquad \forall i$$

On imposition of the two constraints, the constants K_g^i and K_g^j are obtained as

$$K_g^i = \left(\sum_j \frac{K_g^j T_j}{h(d_{ij})} \right)^{-1} \tag{9.8}$$

$$K_g^j = \left(\sum_i \frac{K_g^i T_i}{h(d_{ij})} \right)^{-1} \tag{9.9}$$

Obviously, using the gravity model requires an iterative solution technique as K_g^i depends on K_g^j and vice versa.

EXAMPLE 9.1

Consider the following six-zone model of a town. Zones 1, 2, and 3 are fully residential areas and Zones 4, 5, and 6 are purely shopping areas. The shopping areas, the shopping trips attracted (per day), the shopping trips produced (per day), and the travel distances are as shown in Table 9.1. The cells which have a "–" imply that those data are irrelevant to the problem. Determine the trip distribution between the zones for the following different scenarios:

(a) Use the origin–constrained gravity model, assuming $f(a_j)$ to be a linear function of the shopping area (in square metres) with a slope of 0.01 and constant term of 10. Also assume $h(d_{ij})$ to be d_{ij}^2 where d_{ij} is the distance in km.

(b) Use the origin–destination constrained gravity model with the same relevant assumptions as those in (a).

---

[1]The resulting model is often called the *origin–destination constrained gravity model.*

Table 9.1 Data for Example 9.1 on gravity model

Zone	Shop area (m^2)	Trips produced	Trips attracted	Distance (km) to 1	2	3	4	5	6
1	—	1000	—	—	—	—	4	2	7
2	—	1000	—	—	—	—	3	1	6
3	—	2000	—	—	—	—	5	2	6
4	1000	—	800	4	3	5	—	—	—
5	2000	—	2000	2	1	2	—	—	—
6	3000	—	1200	7	6	6	—	—	—

Solution

(a) The trips of interest here are t_{ij}, where $i \in 1, 2, 3$ and $j \in 4, 5, 6$. Also note, as per the problem description $f(a_j) = 0.01 \times$ (shopping area in m^2) + 10, and $h(d_{ij}) = d_{ij}^2$. The trips are given by Eq. (9.6).

$$t_{14} = \frac{f(a_4)/d_{14}^2}{(f(a_4)/d_{14}^2) + (f(a_5)/d_{15}^2) + (f(a_6)/d_{16}^2)} T_1 = \frac{1.25}{9.57} 1000 = 131$$

$$t_{15} = \frac{f(a_5)/d_{15}^2}{(f(a_4)/d_{14}^2) + (f(a_5)/d_{15}^2) + (f(a_6)/d_{16}^2)} T_1 = \frac{7.5}{9.57} 1000 = 789$$

$$t_{16} = \frac{f(a_6)/d_{16}^2}{(f(a_4)/d_{14}^2) + (f(a_5)/d_{15}^2) + (f(a_6)/d_{16}^2)} T_1 = \frac{0.82}{9.57} 1000 = 85$$

$$t_{24} = \frac{f(a_4)/d_{24}^2}{(f(a_4)/d_{24}^2) + (f(a_5)/d_{25}^2) + (f(a_6)/d_{26}^2)} T_2 = \frac{2.22}{33.33} 1000 = 67$$

$$t_{25} = \frac{f(a_5)/d_{25}^2}{(f(a_4)/d_{24}^2) + (f(a_5)/d_{25}^2) + (f(a_6)/d_{26}^2)} T_2 = \frac{30.0}{33.33} 1000 = 900$$

$$t_{26} = \frac{f(a_6)/d_{26}^2}{(f(a_4)/d_{24}^2) + (f(a_5)/d_{25}^2) + (f(a_6)/d_{26}^2)} T_2 = \frac{1.11}{33.33} 1000 = 33$$

$$t_{34} = \frac{f(a_4)/d_{34}^2}{(f(a_4)/d_{34}^2) + (f(a_5)/d_{35}^2) + (f(a_6)/d_{36}^2)} T_3 = \frac{0.8}{9.41} 2000 = 170$$

$$t_{35} = \frac{f(a_5)/d_{35}^2}{(f(a_4)/d_{34}^2) + (f(a_5)/d_{35}^2) + (f(a_6)/d_{36}^2)} T_3 = \frac{7.5}{9.41} 2000 = 1594$$

$$t_{36} = \frac{f(a_6)/d_{36}^2}{(f(a_4)/d_{34}^2) + (f(a_5)/d_{35}^2) + (f(a_6)/d_{36}^2)} T_3 = \frac{1.11}{9.41} 2000 = 236$$

(b) In this case, the only difference from (a) is that the trips are given by Eq. (9.7) and the constants of Eq. (9.7) are given by Eqs. (9.8) and (9.9). The initial value of K_g^i is assumed to be the square root of the corresponding values obtained in (a). These values of K_g^i are used in Eq. (9.9) to obtain a set of K_g^j values which are in turn used in Eq. (9.8) to obtain a new set of K_g^i values. The process continues till all the K_g^i and K_g^j values converge. The final values of K_g^i and K_g^j obtained are as follows:

Final values of K_g^i

$K_g^1 = 0.359$, $K_g^2 = 0.136$, and $K_g^3 = 0.357$

Final values of K_g^j

$K_g^4 = 0.015$, $K_g^5 = 0.002$, and $K_g^6 = 0.032$

Using these values of K_g^i and K_g^j in Eq. (9.7), the following t_{ij} values are obtained:

$$t_{14} = 272, \; t_{15} = 444, \text{ and } t_{16} = 284$$

$$t_{24} = 182, \; t_{25} = 672, \text{ and } t_{26} = 146$$

$$t_{34} = 346, \; t_{35} = 884, \text{ and } t_{36} = 770$$

The reader must verify that for the above trips $\sum_j t_{ij} = T_i$ and $\sum_i t_{ij} = T_j$.

Discussion. It may be pointed out here that the expression $\dfrac{f(a_j)/h(d_{ij})}{\sum_j f(a_j)/h(d_{ij})}$ in Eq. (9.6) may be viewed as π_{ij}, the probability that destination j is chosen from origin i. Once such a view is taken, then the maximum likelihood technique[2] can be used to estimate the parameters of the gravity model. If observations on the trip distribution between different origin–destination pairs are made and denoted by n_{ij}, then the likelihood function can be written as

$$\mathcal{L} = \prod_{ij} (\pi_{ij})^{n_{ij}} \tag{9.10}$$

[2]For a good discussion on the maximum likelihood technique of estimation, the reader may refer to any good book on statistics or econometrics, for example, Gujarati [95], or Wonnacott and Wonnacott [263].

Other constraints like $\sum_i \pi_{ij} = 1$ could be accounted for by adding the constraint to the likelihood function and constructing a Lagrangian.

Intervening opportunities model

The postulate of travel behaviour on which this model is based, from Stouffer [221], is that the *probability of choice of a particular destination (from a given origin for a particular trip-purpose) is proportional to the opportunities for trip-purpose satisfaction at the destination and inversely proportional to all such opportunities that are closer to the origin.* The inverse proportionality to closer opportunities can be interpreted as proportionality to the probability that none of the closer destinations (opportunities) are chosen. Thus, in this model, the in-situ attractive properties of the destination are modelled as opportunities and the impedances are measured in terms of the number of opportunities which are closer.

In order to formulate the postulate as a mathematical model, the following notations are used:

L : Constant of proportionality (or calibration constant)

J : Total number of destinations

j : Index indicating the position of a destination (from the given origin), $j = 1$ for the nearest destination and $j = J$ for the farthest.

$V_i(j)$: Cumulative function of opportunities up to and including the jth destination from origin i.

$P_i(j)$: Probability that a destination is chosen from the ith origin by the time the jth destination is reached.

π_{ij} : Probability that the jth destination is chosen from the ith origin.

Based on the postulate, the following relation can be written

$$dP_i(j) = L\{1 - P_i(j)\}dV_i(j) \tag{9.11}$$

This relation states that a small addition in opportunities ($dV_i(j)$) after the jth destination causes a correspondingly small increase in the probability ($dP_i(j)$). This increase is proportional to the probability that no destination is chosen by the jth destination and the amount of additional opportunities.

Using $P_i(0) = 0$ (since it is assumed that $V_i(0) = 0$) and solving the differential equation, we get

$$P_i(j) = 1 - \exp[-LV_i(j)] \tag{9.12}$$

and since, by definition, $\pi_{ij} = P_i(j) - P_i(j-1)$, we have

$$\pi_{ij} = \exp[-LV_i(j-1)] - \exp[-LV_i(j)] \tag{9.13}$$

Note, however, nothing guarantees that $\sum_j \pi_{ij} = 1$. In fact, $1 - \sum_j \pi_{ij}$ could be interpreted as the probability that no destination is chosen. This, however, poses a problem in the sequential demand analysis framework because if no destination is chosen then it leads to a situation where a trip is produced but does not go anywhere. In order to correct this anomaly, the π_{ij} obtained in Eq. (9.12) is modified by adding a constraint which states that a destination must be chosen from the set of available destinations thereby precluding the possibility of not choosing any destination. Mathematically this can be done in one of two ways, both leading to the same result. We could use the Bayes theorem and determine the probability that destination j is chosen given that a destination is chosen, or we could simply normalize the π_{ij}s obtained in Eq. (9.13) by dividing the individual π_{ij}s with $\sum_j \pi_{ij}$. In either case the modified π_{ij}, written as π_{ij}^{m}, is given by

$$\pi_{ij}^{m} = \frac{\exp[-LV_i(j-1)] - \exp[-LV_i(j)]}{1 - \exp[-LV_i(J)]} \tag{9.14}$$

Note that the denominator can be looked upon as either (i) the probability a destination is chosen, P(J), or (ii) $\sum_j \pi_{ij}$. (The reader should verify these claims.)

Once the π_{ij}^{m} are obtained, the trips between i and j can be obtained as

$$t_{ij} = \pi_{ij}^{m} T_i \tag{9.15}$$

EXAMPLE 9.2

For the 1200 shopping trips from Zone A, three destinations exist. The destinations are Zones X, Y, and Z. The shopping areas available in each of the zones and their distances from Zone A are given in Table 9.2. Assuming a proportionality constant of 0.35 and assuming a thousand square metres of shopping area as one opportunity, determine the trip distribution from Zone A.

Table 9.2 Data for Example 9.2 on intervening opportunities model

Zone	*Shopping area* (in '000 sq. m)	*Distance from Zone* A (km)
X	2.0	7.0
Y	4.0	12.0
Z	2.0	4.0

Solution

First, the destinations should be arranged from the closest to the farthest; in this case, $j = 1$ for Zone Z, $j = 2$ for Zone X, and $j = 3$ for Zone Y, since there are only three destinations, $j = 3$.

Next, $V_i(j)$ need to be determined; $V_A(1) = 2$, $V_A(2) = 4$, and $V_A(3) = 8$.

Using Eq. (9.14) the π^m_{AX}, π^m_{AY}, and π^m_{AZ}, are calculated as follows:

$$\pi^m_{AZ} = \pi^m_{A1} = \frac{1-\exp(-2\times0.35)}{1-\exp(-8\times0.35)} = 0.536$$

$$\pi^m_{AX} = \pi^m_{A2} = \frac{\exp(-2\times0.35)-\exp(-4\times0.35)}{1-\exp(-8\times0.35)} = 0.266$$

$$\pi^m_{AY} = \pi^m_{A3} = \frac{\exp(-4\times0.35)-\exp(-8\times0.35)}{1-\exp(-8\times0.35)} = 0.198$$

Therefore, the trip distribution from Zone A is given by

$$t_{AX} = 0.266 \times 1200 = 319$$

$$t_{AY} = 0.198 \times 1200 = 238$$

$$t_{AZ} = 0.536 \times 1200 = 643$$

Discussion. Since the model gives the probability of choosing a destination from a given origin, we may employ the maximum likelihood technique to calibrate the constant L. The likelihood function is similar to that given in Eq. (9.10). However, it is seen that assumption of a constant L is not very good for most problems. A varying L adds complexity to the model and its calibration. For a more detailed discussion on intervening opportunity models, the reader may consult Kanafani [132], or Stouffer [221], or Ruiter [197].

Destination choice models

The postulate of travel behaviour on which this model is based is that *destination j will be chosen from origin i for a particular trip-purpose if the perceived utility derived from choosing j is greater than the perceived utility derived from choosing any other destination*. It is further assumed that, $u_i(j)$, the utility derived from destination j by an individual (or a group of similar individuals) in origin i has two components: (i) the deterministic component, $v_i(j)$ and (ii) the stochastic component, $e_i(j)$. The deterministic component is the approximate utility that can be obtained from the destination, given the destination's in-situ attributes and the impedances. On the other hand, the stochastic component can be thought of as an approximation of the random variability assumed to be present in the utility because of the fact that this is a quantity which is perceived by humans. Mathematically,

$$u_i(j) = v_i(j) + e_i(j) \tag{9.16}$$

Since $u_i(j)$ is a stochastic quantity, the answer to the question as to which destination provides a greater utility will be probabilistic. Keeping this in mind the basic postulate of travel behaviour is modified thus: *the probability that destination j will be chosen from origin i for a particular trip-purpose is equal to the probability that the perceived utility derived from j is greater than the perceived utility derived from each of the other destinations.* Mathematically,

$$\pi_{ij} = \text{Prob. } (u_i(j) > u_i(k), \ \forall k \neq j) \tag{9.17}$$

That is,

$$\pi_{ij} = \text{Prob. } \{v_i(j) + e_i(j) > v_i(k) + e_i(k), \ \forall k \neq j\}$$

$$= \text{Prob. } \{e_i(k) < v_i(j) - v_i(k) + e_i(j), \ \forall k \neq j\}$$

Once π_{ij} are obtained, the trip distribution between any i and j can be obtained by multiplying T_i by π_{ij}.

It is, however, clear from the above equations that the exact nature of π_{ij} will depend on the assumptions about the nature of the $e_i(k)$s. If it is assumed (i) that $e_i(k)$s are distributed identically for each k, (ii) that they are independent, and (iii) that they are distributed according to the Gumbel distribution, i.e.

$$\text{Prob. } [e_i(k) < z] = \exp(-\theta e^{-z})$$

then the resulting relation for π_{ij} can be obtained as follows:

$$\pi_{ij} = \int_{-\infty}^{\infty} \prod_{\forall k \neq j} \exp[-\theta \exp - \{v_i(j) - v_i(k) + x\}][\theta \exp(-x) \exp(-\theta \exp(-x))]\, dx \tag{9.18}$$

Now realizing that when $k = j$, then $\exp[-\theta \exp\{v_i(j) - v_i(k) + x\}] = \exp[-\theta \exp(-x)]$ we can rewrite Eq. (9.18) as

$$\pi_{ij} = \int_{-\infty}^{\infty} \exp[-\theta \exp(-x)] \sum_k \exp[-(v_i(j) - v_i(k)] \theta \exp(-x)\, dx \tag{9.19}$$

Now substituting $g = -\theta \exp(-x)$ and $w = \sum_k \exp[-v_i(j) - v_i(k)]$, we obtain

$$\pi_{ij} = \frac{1}{w} \tag{9.20}$$

or

$$\pi_{ij} = \frac{\exp[v_i(j)]}{\sum_k \exp[v_i(k)]} \tag{9.21}$$

The form for π_{ij} given in Eq. (9.21) is known as the **multinomial logit model**.

EXAMPLE 9.3

For the data given in Example 9.2 and assuming $v_A(j)$ = (0.5 × shopping area of Zone j in '000 sq. m) – (0.23 × distance to j in km), determine the trip distribution using the multinomial logit model.

Solution

For the given data, $v_A(X) = -0.61$, $v_A(Y) = -0.76$, and $v_A(Z) = 0.08$. Hence, the probabilities are:

$$\pi_{AX} = \frac{\exp(-0.61)}{\exp(-0.61) + \exp(-0.76) + \exp(0.08)} = 0.26$$

$$\pi_{AY} = \frac{\exp(-0.76)}{\exp(-0.61) + \exp(-0.76) + \exp(0.08)} = 0.22$$

$$\pi_{AZ} = \frac{\exp(0.08)}{\exp(-0.61) + \exp(-0.76) + \exp(0.08)} = 0.52$$

and the trip distribution is given by

$$t_{AX} = 0.26 \times 1200 = 312$$

$$t_{AY} = 0.22 \times 1200 = 264$$

$$t_{AZ} = 0.52 \times 1200 = 624$$

Discussion. Instead of assuming a Gumbel distribution for the random terms, if it is assumed that they are distributed normally then the resulting form for π_{ij} is called the **Probit model**. This model is, however, a lot more cumbersome than the Logit model. For a detailed and extensive discussion on the Probit model, the reader may refer to Kanafani [132].

Another point which may be mentioned here is about the calibration of destination choice models. The parameters which need to be calibrated are the parameters of the function $v_i(j)$. Since the choice models give the probability of choosing a destination from a given origin, we may use the maximum likelihood estimation technique to estimate the parameters. Note that the likelihood function will be the same as that given in Eq. (9.10).

Entropy model

The entropy model of trip distribution, unlike the other models discussed here, is not a behavioural model. That is, the entropy model does not strive to predict the trip distribution by modelling the human behavioural aspects related to choosing a

destination. This model, on the other hand, attempts to determine a distribution of trips which is most likely to occur assuming that each trip occurs independently of another.

For a total number of Q trips in a network, it is assumed that there are a total of Q independent decisions. Obviously, a given trip-distribution matrix can be obtained through different combinations of the decisions. Specifically, the number of ways a trip-distribution matrix $[t_{ij}]$, can be obtained is

$$\frac{Q!}{\prod_{ij} t_{ij}!} \tag{9.22}$$

In the entropy models it is assumed that the likelihood of a particular trip-distribution matrix occurring is proportional to the number of ways the matrix can be obtained. That is, the likelihood of a matrix occurring is proportional to the expression given in Eq. (9.22). The entropy models determine the estimates of t_{ij} by determining that set of t_{ij}s which maximizes the likelihood of a trip-distribution matrix occurring. Mathematically, the problem, therefore, is to solve the following

$$\max_{t_{ij}} \frac{Q!}{\prod_{ij} t_{ij}!} \tag{9.23}$$

Since Q is a constant, this is equivalent to solving

$$\max_{t_{ij}} -\prod_{ij} t_{ij}! \tag{9.24}$$

or

$$\max_{t_{ij}} -\sum_{ij} \ln t_{ij}! \tag{9.25}$$

Using the approximation $\ln x! \approx x \ln x - x$, the above is equivalent to solving

$$\max_{t_{ij}} -\left(\sum_{ij} t_{ij} \ln t_{ij} - \sum_{ij} t_{ij} \right) \tag{9.26}$$

Since $\sum_{ij} t_{ij} = Q$, this is equivalent to solving

$$\max_{t_{ij}} -\left(\sum_{ij} t_{ij} \ln t_{ij} \right) \tag{9.27}$$

The unconstrained optimization problem in Eq. (9.27) is the standard **maximum entropy model** of trip distribution. However, we can easily incorporate various constraints in this model. Some of the constraints which are often included in the

maximum entropy model are: $\sum_j t_{ij} = T_i$, $\sum_i t_{ij} = T_j$, $\frac{\sum_{ij} t_{ij} d_{ij}}{\sum_{ij} t_{ij}}$ = observed average trip length, and so forth.

Discussion. The constraints can be easily incorporated in the function to be maximized by constructing a Lagrangian. The Lagrangian can then be differentiated with respect to the t_{ij}s and the Lagrange's multipliers to obtain a set of expressions which can be equated to zero. These equations when solved will give the estimate of the trip distribution.

Also note that this method does not use any calibration constants and, therefore, the issue of calibrating the model does not arise.

9.3.3 Modal Split Model

The modal split models aim to determine the number of trips on different modes given the travel demand between the different pairs of nodes (zones). These models try to mathematically describe the *mode choice* phase of the sequential demand analysis procedure. Generally, *choice models* are used for modal split analysis. That is, it is assumed that the probability of choosing a particular mode is the probability that the perceived utility from that mode is greater than the perceived utility from each of the other available modes. Since choice models were discussed while presenting *destination choice models* in Section 9.3.2, they are not repeated here. This subsection only discusses the factors which are generally assumed to affect the perceived utility of modes. An example problem is also solved.

The factors which affect the choice of a mode (and hence the perceived utility from a mode) are:

- Socio-economic factors such as income, automobile ownership, age, and so on.
- Service-related factors like in-vehicle travel time, access to public transport (or transit systems), frequency of transit system operation, out-of-pocket cost, and the like.

EXAMPLE 9.4

For a particular zone pair, three modes of travel between the zones exist—private transport like automobiles (PT), bus (B), and the urban rapid transit system like the local trains (RT). It is given that all trip-makers have access to private transport and that the perceived utility of a mode m, i.e. $v(m)$ is given by

$$v(m) = -0.004t_m - 0.005c_m - 0.003w_m + 0.15d_m$$

where

t_m is the in-vehicle travel time in minutes for mode m

c_m is the out-of-pocket cost in rupees for mode m

w_m is the waiting time in minutes for mode m

d_m is a dummy variable which is 1 when the mode is private transport, 0 otherwise.

Assuming that the variable values are as shown in Table 9.3 and that 1000 trips are made from the origin zone to the destination zone, determine the number of trips made by the different modes. Use the Logit model.

Table 9.3 Variable values used in Example 9.4 on modal split model

Mode	*Variable values*			
	t_m (min)	c_m (Rs)	w_m (min)	d_m
PT	65	60	0	1
B	75	5	5	0
RT	25	8	20	0

Solution

First, calculate the perceived utility for each mode:

$$v(\text{PT}) = (-0.004 \times 65) - (0.005 \times 60) - (0.003 \times 0) + 0.15 = -0.41$$

$$v(\text{B}) = (-0.004 \times 75) - (0.005 \times 5) - (0.003 \times 5) = -0.34$$

$$v(\text{RT}) = (-0.004 \times 25) - (0.005 \times 8) - (0.003 \times 20) = -0.20$$

Next, we use the Logit model to determine the probability π_m that a particular mode m will be chosen.

$$\pi_{\text{PT}} = \frac{\exp(-0.41)}{\exp(-0.41) + \exp(-0.34) + \exp(-0.20)} = 0.302$$

$$\pi_{\text{B}} = \frac{\exp(-0.34)}{\exp(-0.41) + \exp(-0.34) + \exp(-0.20)} = 0.324$$

$$\pi_{\text{RT}} = \frac{\exp(-0.20)}{\exp(-0.41) + \exp(-0.34) + \exp(-0.20)} = 0.374$$

Hence, 302 trips will be made using private transport, 324 will use buses, and 374 will use the rapid transit system.

9.3.4 Traffic-assignment Models

Traffic-assignment models aim to determine the number of trips on different links (road sections) of the network given the travel demand between the different pairs of nodes (zones). These models try to mathematically describe the *route-choice* phase of the sequential demand analysis procedure. There are various models of traffic assignment. All of these models assume that travel time on the link is the only factor which the trip-makers consider while choosing a route. These models, however, differ in their assumptions concerning the variation in link travel times with the link volume (or link flow). In this subsection, three models are discussed, namely (i) all-or-nothing assignment model, (ii) incremental-assignment model, and (iii) user-equilibrium model. The notations used (other than those used in the preceding subsections) in describing these models are as follows:

x_a : Flow (or volume) on link a.

x_a^k : Flow (or volume) on link a as estimated in the kth iteration.

$\tau_a(x_a)$: Travel time on link a when flow on link a is x_a.

t_{ij} : The total demand (or the number of trips) between origin i and destination j.

$\delta_{a,k}^{i,j}$: Indicator variable which is 1 if link a is a part of the kth route between origin i and destination j; otherwise it is zero.

$\Phi_k^{i,j}$: Flow between i and j which uses the kth route.

All-or-nothing assignment model

In this model it is assumed that (i) the travel times on links do not vary with link flows, i.e. $\tau_a(x_a) = \tau_a$ and (ii) all trip-makers (users) have precise knowledge of the travel times on the links. Based on these assumptions about travel times and the postulate that a trip-maker will choose that path (or route) which minimizes his/her travel time, this assignment model assigns all the trips between a particular origin and destination pair to that route (or path) which offers the minimum travel time.

The exact nature of the assignment model is presented through the following algorithm.

Step 1. For every i–j pair (i.e. origin–destination pair) with $t_{ij} > 0$, determine the minimum travel time path (or route) using τ_a as the link travel times. The minimum path determination can be done using any of the various existing algorithms like Flyod's algorithm or Djkastra's algorithm. Detailed description of these algorithms can be found in Teodorovic [233] or any other book on theory of networks. Also, initialize all $x_a^0 = 0$.

Step 2. Set iteration counter $k = 1$. Select a particular i–j pair.

Step 3. Assign the entire t_{ij} to the minimum path between the i–j pair. If link a is a part of the minimum path set, $x_a^k = x_a^{k-1} + t_{ij}$, else set $x_a^k = x_a^{k-1} + 0$.

Step 4. If $k = N$ (where N is the total number of i–j pairs with $t_{ij} > 0$) then report x_a^k as x_a. Else, select another i–j pair; set $k = k + 1$ and go back to Step 3.

EXAMPLE 9.5

For the network shown in Figure 9.5 and the trip-distribution matrix given in Table 9.4 determine the link flows using the all-or-nothing assignment technique. Note that the numbers on the links of the network denote the travel times and the numbers in the circles denote the zone numbers.

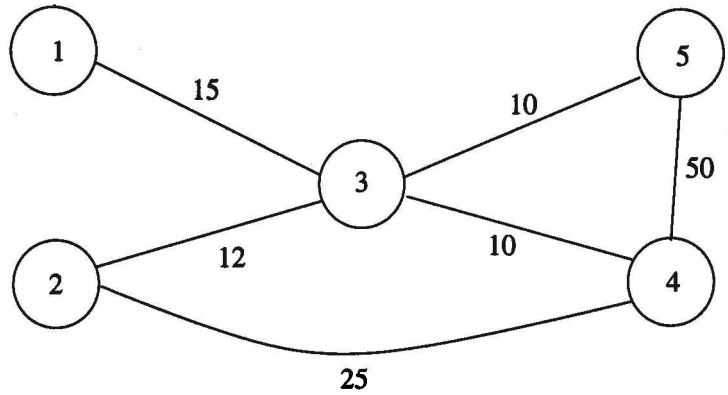


Figure 9.5 Network for Example 9.5 on all-or-nothing assignment technique.

Table 9.4 Trip-distribution matrix (O-D matrix) for Example 9.5 on all-or-nothing assignment model

Origin zone	*Destination zone*				
	1	2	3	4	5
1	0	0	200	100	150
2	0	0	300	300	50
3	200	300	0	100	100
4	100	300	100	0	0
5	150	50	100	0	0

Solution

Note there are 25 possible zone pairs out of which 9 have, $t_{ij} = 0$. Hence, $N = 16$.

Step 1

The minimum paths for the 16 zone pairs (obtained using Djkastra's algorithm) are as follows:

i–j pair	*Min. path*	*i–j pair*	*Min. path*
1–3	1 → 3	1–5	1 → 3 → 5
3–1	3 → 1	5–1	5 → 3 → 1
1–4	1 → 3 → 4	2–3	2 → 3
4–1	4 → 3 → 1	3–2	3 → 2
2–4	2 → 3 → 4	2–5	2 → 3 → 5
4–2	4 → 3 → 2	5–2	5 → 3 → 2
3–4	3 → 4	3–5	3 → 5
4–3	4 → 3	5–3	5 → 3

Step 2

$k = 1$. Consider the zone pair 1–3.

Step 3

$x^1_{1\to 3} = 0 + 200 = 200$; the rest of the x_a^k remain zero.

Step 4

Since $k \neq 16$, set $k = 2$ and select zone pair 1–5 as the next pair and go back to Step 3.

Step 3

$x^2_{1\to 3} = 200 + 150 = 350; x^2_{3\to 5} = 0 + 150 = 150$; the rest of the x_a^k remain zero.

Step 4

Since $k \neq 16$, set $k = 3$ and select zone pair 3–1 as the next pair and go back to Step 3.

In this manner, Steps 3 and 4 are repeated till all the zone pairs are chosen (i.e. $k = 16$). Finally, the following assignment is obtained.

$x_{1\to 3} = 450$	$x_{3\to 5} = 300$
$x_{2\to 3} = 650$	$x_{4\to 2} = 0$
$x_{2\to 4} = 0$	$x_{4\to 3} = 500$
$x_{3\to 1} = 450$	$x_{5\to 3} = 300$
$x_{3\to 2} = 650$	$x_{5\to 4} = 0$
$x_{3\to 4} = 550$	$x_{4\to 5} = 0$

Incremental-assignment model

The incremental-assignment model, in addition to the postulate that each trip-maker chooses a path so as to minimize his/her travel time, also assumes that the travel time

on a link varies with the flow on that link. Under such an assumption, the ideal way to assign traffic volume would be to assign a single trip to the road network assuming that the travel time on links during the assignment is constant. We could then update the travel times and repeat the process till all the trips are assigned. However, this procedure is not practical as any network would typically have a very large number of trips. The incremental-assignment models, therefore, try to approximate this ideal process by dividing the total number of trips into a few smaller parts and then assigning each part with a constant link travel time.

The exact nature of the assignment model is presented through the following algorithm.

Step 0. Divide the entire trip-distribution matrix (or origin-destination matrix) into n smaller part matrices. Note that, the sum of all the part matrices should be equal to the actual trip-distribution matrix.
Set counter $m = 1$.
Set $x_a^{m-1} = 0$ for all a.
(Also note that in the following, t_{ij}^m refers to the number of trips from i to j as per the mth part matrix.)

Step 1. Set $v_a = 0$ for all links.
Assuming $\tau_a(x_a^{m-1})$ as the link travel times, assign the trips of the mth part matrix using all-or-nothing assignment technique. Store the link volumes obtained from the all-or-nothing assignment technique as v_a.

Step 2. Update the link volumes using $x_a^m = x_a^{m-1} + v_a$.

Step 3. If $m = n$ then report x_a^m as x_a and stop. Else, set $m = m + 1$ and go to Step 1.

EXAMPLE 9.6

For the network shown in Figure 9.6 and the trip-distribution matrix given in Table 9.5, determine the link flows using the incremental-assignment technique. The link travel times $\tau_a(x_a)$ are given by: $\tau_a(x_a) = k_a[1 + 0.15(x_a/b_a)^4]$. The link number, the k_a value, and

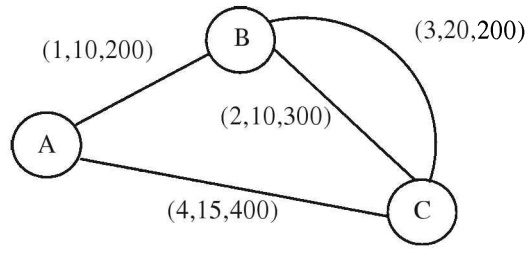


Figure 9.6 Network for Example 9.6 on incremental-assignment technique.

the b_a value, for a particular link are mentioned as (α, β, γ) on the links. Divide the trip-distribution matrix into four parts in the ratio 40:30:20:10.

Table 9.5 Trip-distribution matrix for Example 9.6 on incremental-assignment model

Origin zone	Destination zone		
	A	B	C
A	0	250	150
B	250	0	400
C	150	400	0

Solution

Step 0

The trip-distribution matrix is divided into the following four (i.e. $n = 4$) parts:

Part 1 matrix

Origin zone	Destination zone		
	A	B	C
A	0	100	60
B	100	0	160
C	60	160	0

Part 2 matrix

Origin zone	Destination zone		
	A	B	C
A	0	75	45
B	75	0	120
C	45	120	0

Part 3 matrix

Origin zone	Destination zone		
	A	B	C
A	0	50	30
B	50	0	80
C	30	80	0

Part 4 matrix

Origin zone	Destination zone		
	A	B	C
A	0	25	15
B	25	0	40
C	15	40	0

Set counter $m = 1$.
Set $x_1^0 = 0$, $x_2^0 = 0$, $x_3^0 = 0$, and $x_4^0 = 0$.

Step 1

Set $v_1 = 0$, $v_2 = 0$, $v_3 = 0$, and $v_4 = 0$.
Using Part 1 matrix, $\tau_1(0) = 10$ min, $\tau_2(0) = 10$ min, $\tau_3(0) = 20$ min, and $\tau_4(0) = 15$ min, and all-or-nothing assignment the following values for v_a are obtained:

$$v_1 = 200,\ v_2 = 320,\ v_3 = 0, \text{ and } v_4 = 120$$

Step 2

Using x_a^0 and v_a, the following quantities are obtained:

$$x_1^1 = 200,\ x_2^1 = 320,\ x_3^1 = 0, \text{ and } x_4^1 = 120$$

Step 3

Since m (=1) < n (= 4), set m = 2 and go to Step 1.

Step 1

Set $v_1 = 0$, $v_2 = 0$, $v_3 = 0$, and $v_4 = 0$.
Using Part 2 matrix, $\tau_1(200) = 11.5$ min, $\tau_2(320) = 11.9$ min, $\tau_3(0) = 20$ min, and $\tau_4(120) \approx 15$ min, and all-or-nothing assignment, the following values for v_a are obtained:

$$v_1 = 150,\ v_2 = 240,\ v_3 = 0,\ \text{and}\ v_4 = 90$$

Step 2

Using x_a^1 and v_a, the following quantities are obtained:

$$x_1^2 = 350,\ x_2^2 = 560,\ x_3^2 = 0,\ \text{and}\ x_4^2 = 210$$

Step 3

Since m (= 2) < n (= 4), set m = 3 and go to Step 1.

Step 1

Set $v_1 = 0$, $v_2 = 0$, $v_3 = 0$, and $v_4 = 0$.
Using Part 3 matrix, $\tau_1(350) = 24.1$ min, $\tau_2(560) = 28.2$ min, $\tau_3(0) = 20$ min, and $\tau_4(210) = 15.2$ min, and all-or-nothing assignment the following values for v_a are obtained:

$$v_1 = 100,\ v_2 = 0,\ v_3 = 160,\ \text{and}\ v_4 = 60$$

Step 2

Using x_a^2 and v_a, the following quantities are obtained:

$$x_1^3 = 450,\ x_2^3 = 560,\ x_3^3 = 160,\ \text{and}\ x_4^3 = 270$$

Step 3

Since m (= 3) < n (= 4) set m = 4 and go to Step 1.

Step 1

Set $v_1 = 0$, $v_2 = 0$, $v_3 = 0$, and $v_4 = 0$.
Using Part 4 matrix, $\tau_1(450) = 48.4$ min, $\tau_2(560) = 28.2$ min, $\tau_3(160) = 21.2$ min, and $\tau_4(270) = 15.5$ min, and all-or-nothing assignment the following values for v_a are obtained:

$$v_1 = 0,\ v_2 = 0,\ v_3 = 130,\ \text{and}\ v_4 = 80$$

Step 2

Using x_a^3 and v_a, the following quantities are obtained:

$$x_1^4 = 450,\ x_2^4 = 560,\ x_3^4 = 290,\ \text{and}\ x_4^4 = 350$$

Step 3

Since m (= 4) = n (= 4), report $x_1^4 = x_1 = 450$, $x_2^4 = x_2 = 560$, $x_3^4 = x_3 = 290$, and $x_4^4 = x_4 = 350$.

Discussion. Although the incremental-assignment technique overcomes the shortcoming of the all-or-nothing assignment technique by incrementally assigning the entire trip-distribution matrix and updating the link travel times with flow, it still suffers from a major drawback. Despite the fact that traffic assignment is an outcome of the route choice behaviour of humans, the incremental-assignment technique does not have any behavioural basis and therefore remains more of a computational technique than a mechanism of traffic assignment which mirrors the route choice behaviour of humans.

User-equilibrium model

The user-equilibrium model of traffic assignment is based on the fact that humans choose a route so as to minimize their travel time and on the assumption that such a behaviour on the individual level creates an *equilibrium* at the system (or network) level. Flows on links (whose travel times are assumed to vary with flow) are said to be in equilibrium when no trip-maker can improve his/her travel time by unilaterally shifting to another route. This notion of equilibrium flows is generally referred to as *Wardrop's principle*. Before presenting the model which can determine such equilibrium flows on a network, the idea of equilibrium flows or the concept of user-equilibrium needs to be explained further.

Consider the example network shown in Figure 9.7(a). Figure 9.7(b) gives the travel time function for each of the three single link routes between the origin O and destination D. The total demand from O to D is 110.

In the example, obviously if the demand is less than or equal to 100, everybody will travel using Route 2 since the travel time offered by Route 2 (between 30–40 min) will be less than any other route. However, when the demand is in excess of 100, if everybody travels on Route 2, then the travel time on Route 2 will become more than 40 min, implying that Route 1 will become more lucrative (offering about 40 min of travel time, since the volume on Route 1 is zero at this stage) and some can benefit by unilaterally shifting to Route 1. On doing so, if the volume carried by Route 2 falls below 100 then again Route 2 will become lucrative and some will shift to Route 2 from Route 1. Given this, and the fact that the total demand is 110, it can be said that if 100 use Route 2 and 10 use Route 1 then each one will face a travel time of about 40 min

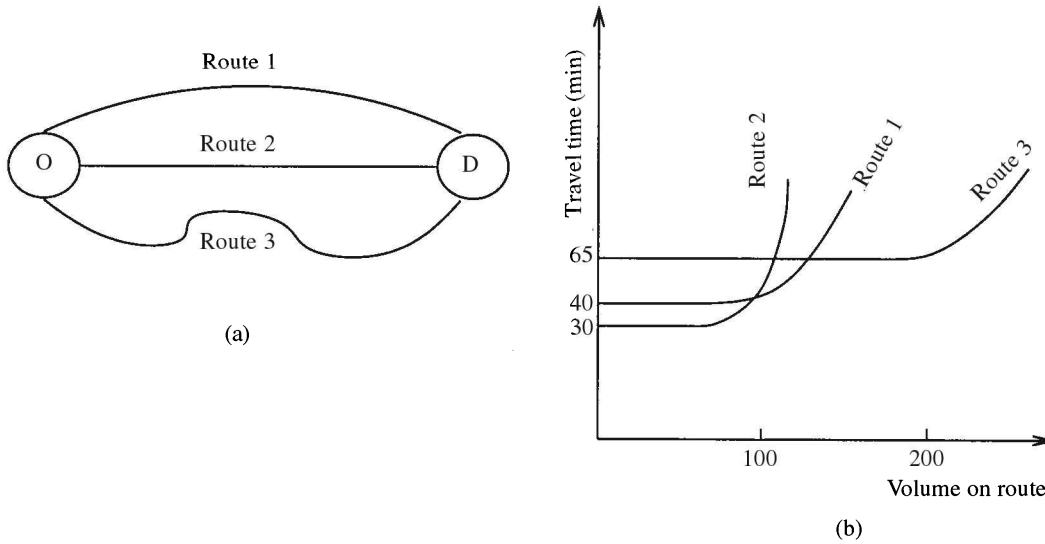


Figure 9.7 The use of user-equilibrium assignment technique.

and none can improve his/her travel time by unilaterally shifting from one route to another. This then will be the *equilibrium flow* to which the network has to eventually converge given that every traveller always tries to minimize his/her travel time. Note that under the equilibrium condition, Route 3 is not used as it entails more than 40 min of travel time.

The above example, in addition to clarifying the concept of equilibrium, also illustrates the fact that at equilibrium all routes which are used between a given origin and destination offer the same travel time which is less than all routes which are not used. Another interesting feature is that at equilibrium flow, the sum of the areas under the travel time versus flow plots is the least. This feature can be easily seen by visually inspecting the sum of the areas shown in Figure 9.8. Figure 9.8(a) corresponds to the equilibrium flow, and the sum of the areas is equal to *area* ACDE + *area* ABFG. Figure 9.8(b), corresponds to an instance of non-equilibrium flow conditions with Route 2 carrying all the flow (i.e. 110); in this case the relevant sum of the areas is simply equal to *area* HIJK. Figure 9.8(c) represents another instance of non-equilibrium flow where Route 2 carries a flow of 90 and the rest 20 are carried by Route 1; in this case the relevant sum of the areas is the *area* LMNO + *area* LPQR. Finally, Figure 9.8(d) represents another instance of non-equilibrium flow conditions with Routes 1, 2, and 3 carrying flows of 30, 70, and 10, respectively; the relevant sum of the areas in this case is *area* TUVW + *area* TXYZ + *area* $T\alpha\beta\gamma$. Note that the sum of the areas is the least in (a)—the case representing the equilibrium flow. The interested reader may try to verify this feature by constructing other such examples or by proving them mathematically.

The user-equilibrium model formulation is based on this area-related observation about equilibrium flows and the fact that the travel time on a route is simply a linear sum

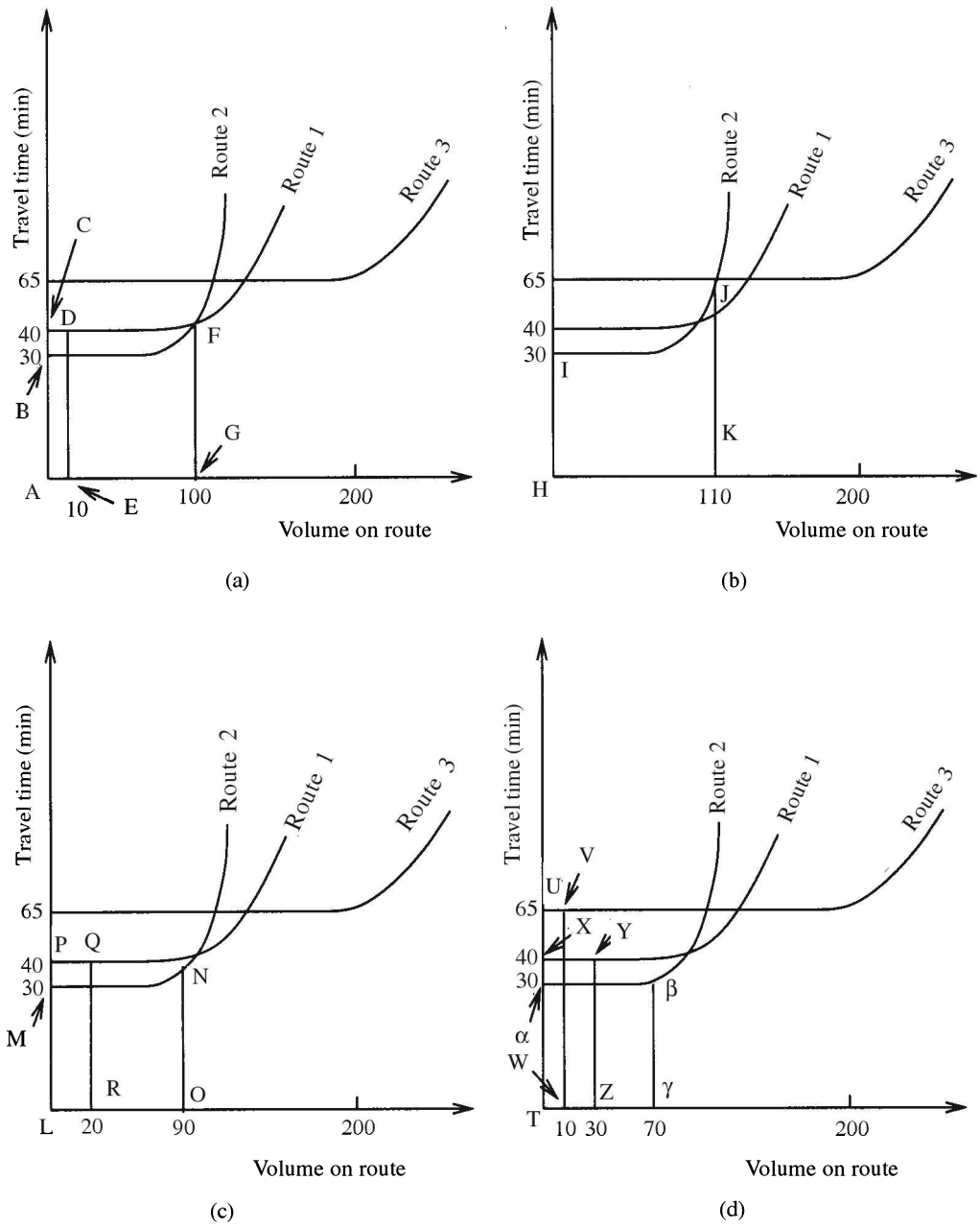


Figure 9.8 Areas under travel-time versus flow plots under various flow distributions.

of the travel times on the constituent links. The formulation, which is a non linear programming problem, is given in Eq. (9.28).

$$\underset{x_a}{\text{minimize}} \sum_a \int_0^{x_a} \tau_a(w)\,dw$$

subject to

$$\begin{aligned} \sum_k \Phi_k^{i,j} &= t_{ij} && \forall i, j \\ \sum_i \sum_j \sum_k \Phi_k^{i,j} \delta_{a,k}^{i,j} &= x_a && \forall a \\ \Phi_k^{i,j} &\geq 0 && \forall i, j, k \end{aligned} \tag{9.28}$$

where w is a variable denoting flow. Nonlinear programming problems of this kind (with linear constraints and a nonlinear objective function) can be effectively solved using convex combination algorithms. The interested reader may refer to Taha [225] for more details on convex combination algorithms. Here, only the principle is given so that the latter discussions on solving user-equilibrium problems are more meaningful to the reader.

Principles of convex combination methods for solving optimization problems with non-linear objective function and linear constraints

The convex combination algorithm determines the optimum point by proceeding in the following iterative manner.

- First, a feasible solution (i.e. a solution which satisfies all the constraints), say f_1, is determined.
- Next, another feasible solution, say f_2, is determined.
- Since all the constraints are linear (thereby giving rise to a convex solution space), it can be shown that any point f_3 on a straight line joining f_1 and f_2 (note $f_3 = \alpha f_1 + (1 - \alpha) f_2$; where $0 \leq \alpha \leq 1$) is also feasible. The convex combination algorithm then determines the value of α which gives the best value for the objective function. Note that the determination of the best value of α, say α^*, requires an unconstrained optimization of the objective function written in terms of α.
- Once the value of α^* is determined, the point f_3 is precisely identified by using the value of α^* in place of α in the earlier equation for f_3.
- Next f_3 is set equal to f_1; also another feasible point is determined and f_2 is replaced with the new feasible point. The process goes back two steps and continues repeatedly till the objective function value changes negligibly from one iteration to the next. The optimal solution is f_1 obtained in the last iteration.

Solving the mathematical programming formulation of the user-equilibrium model

As stated earlier, the convex combination algorithm is used to solve the mathematical programming formulation of the user-equilibrium model. Here the procedure is presented step-wise. For a better understanding of the procedure, parallels are drawn with the steps presented in the earlier subsection.

Step 1. Set iteration counter $n = 1$. Determine a feasible solution by performing *all-or-nothing* assignment with $\tau_a(0)$ as the travel time on link a. *Note that the feasibility conditions of the programming formulation will be met by performing the all-or-nothing assignment.* Let this solution vector be s_1^n.

Step 2. Update the travel time of the links using the link volumes obtained in the s_1^n.

Step 3. With the updated travel time information, perform an *all-or-nothing* assignment to obtain another feasible solution. Let this solution vector be s_2^n.

Step 4. Determine $\alpha^{*,n}$ by finding that value of α which minimizes the following quantity

$$\sum_a \int_0^{[s_1^n + \alpha(s_2^n - s_1^n)]_a} \tau_a(w)\,dw$$

where $[s_1^n + \alpha(s_2^n - s_1^n)]_a$ indicates the flow value for link 'a' as per the solution vector $[s_1^n + \alpha(s_2^n - s_1^n)]$.

Step 5. Obtain the new feasible point, $s_1^{n+1} = s_1^n + \alpha^{*,n}(s_2^n - s_1^n)$.

Step 6. Check whether the value of the objective function given in Eq. (9.28) or in Step 4 here has changed significantly. If it has not, then stop and report s_1^{n+1} as the solution. Otherwise, set $n = n + 1$ and continue from Step 2 onwards.

Example 9.7 illustrates how the solution procedure suggested here works.

EXAMPLE 9.7

For the network shown in Figure 9.9, and 1000 trips per day from Node A to Node B, determine the link flows using the user-equilibrium assignment technique. The link travel times $\tau_a(x_a)$ are given by: $\tau_a(x_a) = k_a[1 + 0.15(x_a/b_a)^4]$. The link number, the k_a value, and the b_a value for a particular link are mentioned as (α, β, γ) on the links. Note that in this case each route has one link.

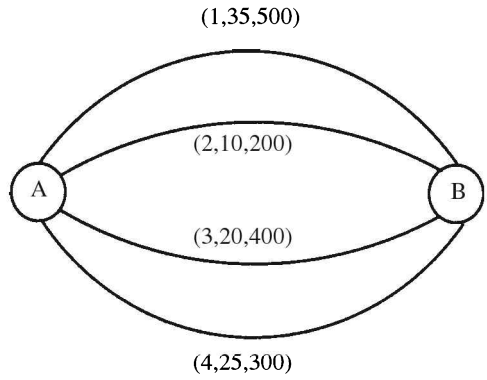


Figure 9.9 Road network for Example 9.7 on user-equilibrium method of traffic assignment.

Solution

Table 9.6 provides the output of every step at each iteration. As per the table the flow on link a, x_a is: $x_1 = 0$, $x_2 = 359$, $x_3 = 470$, and $x_4 = 171$. Note that all links which are used have approximately the same travel time and this is less than the travel time offered by the unused link (in this case Link 1).

Table 9.6 Solution of Example 9.7 on user-equilibrium model of traffic assignment

n	Step		Links				Objective function value
			1	2	3	4	
1	1	$\tau_a(0)$	35	10	20	25	
		s_1^n	0	1000	0	0	1975.0
	2	$\tau_a(s_1^n)$	35	947	20	25	
	3	s_2^n	0	0	1000	0	
	4	$\alpha^{*,n}$	0.596				
	5	s_1^{n+1}	0	404	596	0	197.0
	6		$n = 2$; go to Step 2				
2	2	$\tau_a(s_1^n)$	35	35	35	25	
	3	s_2^n	0	0	0	1000	
	4	$\alpha^{*,n}$	0.161				
	5	s_1^{n+1}	0	339	500	161	189.98
	6		$n = 3$; go to Step 2				

(Contd.)

Table 9.6 (contd.)

3	2	$\tau_a(s_1^n)$	35	22.3	27.3	35.3	
	3	s_2^n	0	1000	0	0	
	4	$\alpha^{*,n}$	0.035				
	5	s_1^{n+1}	0	362	483	155	189.44
	6		n = 4; go to Step 2				
4	2	$\tau_a(s_1^n)$	35	26.1	26.3	25.3	
	3	s_2^n	0	0	0	1000	
	4	$\alpha^{*,n}$	0.020				
	5	s_1^{n+1}	0	354	473	172	189.33
	6		n = 5; go to Step 2				
5	2	$\tau_a(s_1^n)$	35	24.8	25.8	25.4	
	3	s_2^n	0	1000	0	0	
	4	$\alpha^{*,n}$	0.007				
	5	s_1^{n+1}	0	359	470	171	189.33
	6		Stop; Report s_1^{n+1} as solution				

9.4 COLLECTION OF TRANSPORTATION DEMAND DATA

While discussing the models for transportation demand analysis it became clear that empirical data on travel behaviour and demographics are required for calibrating some of the models. Specifically, the following types of data are necessary:

- Demographic data; these include
  - Income at individual, household, or regional level
  - Private transport vehicle ownership at individual, household, or regional level
  - Age distribution
  - Profession
- Travel behaviour
  - Number of trips produced per individual or household
  - Destination choice behaviour in terms of origin–destination matrix

The purpose of this section is to provide some idea as to how such data are obtained. Generally, the demographic data is obtained from census studies and are modified (if the census is old) using the growth patterns of the area. Transportation professionals,

typically, are not involved in the collection of such data. However, data on trip-making behaviour are collected by transportation professionals and therefore their collection mechanisms are briefly described here. Most of the techniques used fall under two broad classes: (i) the direct method and (ii) the indirect method.

In the direct method, the data on travel behaviour are collected by directly observing the trips or by interviewing the prospective trip-makers. The method of interviewing is generally referred to as *home interview method* or *road-side interview method*. The names are self-explanatory. In this method, data collectors interview people on their trip-making behaviour (like how many trips they make, the origins and destinations of their trips) and on certain demographic aspects. The indirect methods, on the other hand, rely on (i) traffic counts, observations of licence plates, etc. and (ii) the mathematical techniques to derive information on travel behaviour. Among the methods of this class are cordon line count-based methods, link volume count-based methods, ordinal information-based methods, and so forth.

The reader may refer to Richardson et al. [196] and Moser and Kalton [169] for details on direct methods of data collection. The reader can refer to Reddy and Chakroborty [194, 195] for a survey of indirect methods and relevant references.

EXERCISES

1. Given the following data set, develop a trip table with four categories of households. List the problems you face while doing so.

Household number	*Monthly income* (in '000 Rs)	*Occupants between* 6 and 60 years	*Trips produced per day*
1	10	2	4
2	15	5	4
3	35	5	14
4	30	6	14
5	40	3	8
6	15	5	6
7	20	3	3
8	32	3	8
9	40	3	9
10	38	2	6
11	40	6	10
12	12	6	6
13	35	2	8
14	40	2	7
15	15	2	3

2. Using the trip table obtained in Exercise 1, determine the total number of trips produced from a zone with 2000 households in each category.

3. Determine which of the following three trip distribution tables (giving the trips made from Zone A to Zones B, C, and D) is likely to be obtained if we used the maximum entropy model of trip distribution.

	B	C	D
A	3	4	5

	B	C	D
A	4	4	4

	B	C	D
A	4	6	2

4. For the 1000 shopping trips from Zone A, four destinations exist. The destinations are Zones W, X, Y, and Z. The shopping areas available in each of the zones and their distances from Zone A are given in the following table. Assuming a proportionality constant of 0.35 and taking a 1000 sq. m of shopping area as one opportunity, determine the trip distribution from Zone A.

Zone	*Shopping area* (in '000 sq.m)	*Distance from Zone* A (km)
W	2.0	5.0
X	1.0	8.0
Y	5.0	11.0
Z	3.0	5.0

5. A Logit model is being developed for mode-choice behaviour. The choice is between two modes—automobile and bus. The utilities derived from using an automobile and a bus are u_A and u_B respectively. The following models for v_A and v_B are proposed (note that, the notation used is $u_i = v_i + e_i$, where e_i is a stochastic term): $v_A = a_A + bt_A$ and $v_B = bt_B$, where t_k is the travel time by mode k and a_A and b are constants.

   Based on the above description, answer the following:

   (a) Show that $\pi_A/\pi_B = e^{v_A}/e^{v_B}$; where π_i is the probability of choosing mode i.

   (b) What is the purpose of including a constant term in v_A?

   (c) What sign should one expect for b and why?

   (d) If it is observed that people are more inclined to use their automobiles (that is, they are biased towards the automobile mode) what sign should one expect for a_A?

   (e) Suppose a new mode, say rail, is introduced and $v_R = bt_R$, what changes can be observed in π_A, π_B, and π_A/π_B?

6. For the incremental assignment Example 9.6 assign the O–D matrix (a) in three increments by dividing the matrix into three parts in the ratio 40:30:30, and (b) in five increments by dividing the matrix into five parts in the ratio 40:30:10:10:10. Compare the results and state what you observe.

7. Use the user-equilibrium method to assign the trips given in Exercise 6.

8. Show that for Example 9.7 on user-equilibrium mehtod, the objective function value at $n = 1$ is 1975. Also, show that the value of $\alpha^{*,n}$ for $n = 1$ is 0.5964.

PART IV

PAVEMENT ENGINEERING

Pavement Materials and Characterization

10.1 INTRODUCTION

This chapter discusses the engineering characterization of various materials used in pavement construction, namely soil, aggregate, bitumen, bituminous mix, cement, cement concrete, stabilized soil, and other cemented materials. Some pavement materials such as cement, cement concrete, and stabilized soil are only briefly described here. The reader interested in in-depth treatment of these materials may, however, refer to books [174, 175, 99] listed in references. Also, only those properties of a material which are important from pavement engineering point of view have been discussed in this chapter.

10.2 SOIL

Soil is the ground support on which roads are built. Soil which originated from weathering of rocks, is the loose mass of mineral available in abundance over the crust of the earth.

The subgrade, which is the bottom-most layer of a pavement, is made up of compacted soil. Road embankments are built with soil. Soil is sometimes used as one of the ingredients in the base/sub-base layer of a pavement. Characterization of soil, that is, knowledge of behaviour of soil is an essential part of pavement engineering.

Comprehensive coverage of engineering properties of soil is available in geotechnical engineering books [141]. Only some important aspects which need special mention with reference to pavement engineering are discussed in this section.

10.2.1 Characterization

A number of parameters are used for characterization of soil. Some of them such as resilient modulus, Poisson's ratio, and permeability of soil are discussed here.

Resilient modulus

The elastic modulus is an important parameter used in the analysis and design of a pavement structure. For soil and granular material, the equivalent term is *resilient modulus* (M_R). Figure 10.1 shows a triaxial state of stress, to which a sample (of soil or granular material) placed in a triaxial cell is subjected. To simulate the dynamic loading condition as observed in the field

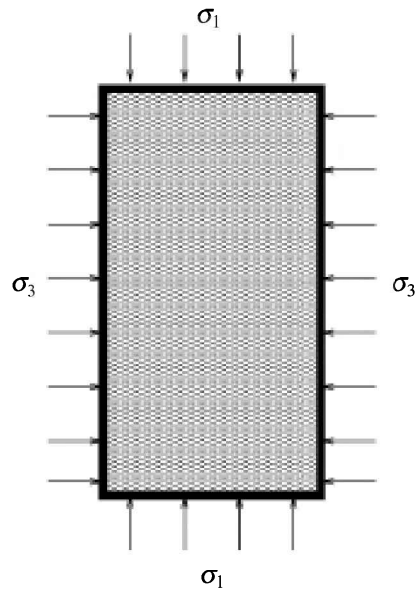


Figure 10.1 A triaxial state of stress.

due to traffic, the stress $\sigma_1 - \sigma_3$ (known as deviatoric stress or pressure) is made pulsating. This dynamic nature of triaxial testing is intended to match the loading and the unloading durations in the same way as they occur in the in-service road. Deformation of the sample occurs when the load is applied to it, and recovery takes place when the load is removed. Dynamic triaxial testing on soil or granular material shows that a fraction of the total strain is unrecoverable, called *permanent deformation*, even when the load is removed. Figure 10.2 shows the concept of recoverable and non-recoverable strains.

The permanent deformation is specially prominent when the sample is subjected to a large number of repetitions. On an in-service road as well, permanent deformation occurs due to repetitive traffic loading called *rutting*. Rutting is discussed in detail in Section 12.3.6. The M_R of soil (or granular material) is defined as the deviatoric stress divided by recoverable strain. Thus,

$$M_R = \frac{\sigma_1 - \sigma_3}{\varepsilon_{re}} \tag{10.1}$$

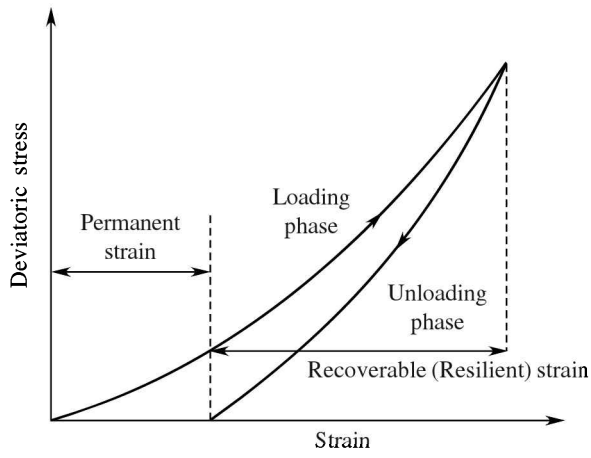


Figure 10.2 Recoverable and permanent strain in dynamic triaxial loading.

where ε_{re} is the recoverable axial strain. Based on tests on various materials, different empirical formulae for M_R value have been suggested.[1]

If the facilities for repeated triaxial test are not available, the M_R value of the subgrade soil may be estimated from its CBR value with the help of the following equations [146]:

$$M_R = 10 \times \text{CBR} \qquad \text{for} \qquad \text{CBR} \leq 5\% \tag{10.5}$$

and

$$M_R = 17.6 \times \text{CBR}^{0.64} \qquad \text{for} \qquad \text{CBR} > 5\% \tag{10.6}$$

The reader may note that Eqs. (10.5) and (10.6) are not dimensionally balanced, hence M_R must have a specific unit. The M_R value obtained from these empirical formulae is in MPa unit. Equations (10.5) and (10.6) are also recommended by the Shell pavement design manual [206]. Equation (10.6) can be used for CBR values even up to 100% for evaluation of M_R [206]. Indian pavement design guidelines

[1]For silty clay soil, M_R values of the subgrade are given by the following bilinear equations [207]:

$$M_R = k_2 + k_3[k_1 - (\sigma_1 - \sigma_3)] \qquad \text{when} \qquad k_1 > (\sigma_1 - \sigma_3) \tag{10.2}$$

$$M_R = k_2 + k_4[(\sigma_1 - \sigma_3) - k_1] \qquad \text{when} \qquad k_1 < (\sigma_1 - \sigma_3) \tag{10.3}$$

where k_1, k_2, k_3, and k_4 are the material constants.

For clayey soil, the M_R value is given by

$$M_R = k_1 \times \left(\frac{\sigma_1 - \sigma_3}{\sigma_3}\right)^{-k_2} \tag{10.4}$$

where k_1 and k_2 are the material constants. Dynamic triaxial test on subgrade soil collected from NH-6, showed [179] that the values of k_1 ranged from 35.7 to 57.3 and those of k_2 varied from 0.110 to 0.353.

IRC:37–2001 [89] recommend both Eqs. (10.5) and (10.6) for computation of the M_R value of soil. Pavement design guidelines published by the Asphalt Institute [237] and AASHTO [2] also use the same form of equation as that of (10.5) except that there are slight variations in the coefficients.

Poisson's ratio

Poisson's ratio μ is defined as the ratio of lateral strain ε_l to the axial strain ε_a, caused by a load parallel to the axis along which ε_a is measured. It is found that for most of the pavement structures, the influence of μ value is normally small [162]. This allows the use of typical constant values for analysis rather than direct testing. The μ values of clayey subgrade vary from 0.4 to 0.5 and a value of 0.5 is adopted for the wet condition [22]. The μ values of saturated clays and sand can be taken as 0.5 and 0.35, respectively [162]. Further discussion on Poisson's ratio of various pavement materials is contained in Section 11.3.2.

Permeability

Permeability of soil is the ease with which water can flow through it. An idea of the permeability of soil helps a designer to take into account the sub-surface drainage considerations of a pavement structure. Drainage design aspects have been separately discussed in Section 12.6. The basic law on permeability, known as Darcy's (D'Arcy) law, is given by

$$Q = k \times i \times A \tag{10.7}$$

where

Q is the quantity of flow or discharge
k is the permeability of the media (soil in this case)
i is the hydraulic gradient
A is the cross-sectional area perpendicular to the direction of flow.

The coefficient of permeability is determined either in the laboratory by (i) the constant head or (ii) the falling head test, or in the field by pumping tests. The factors affecting permeability of soil particles are particle size, shape, relative distribution (gradation), degree of saturation, degree of compaction, etc. For example, sand has a high coefficient of permeability whereas the permeability of clay is low. Permeability of soil can be increased by adding flocculants (e.g. lime, gypsum) and can be decreased by deflocculants (e.g. cement slurry) [143].

10.2.2 Some Tests on Soil

This section discusses the shear test, the CBR test, the plate load test, and the triaxial test for characterization of soil. The reader may refer to other tests for example given in [141], for various specific purposes of soil characterization.

Shear test

Some of the shear strength tests are (i) direct shear test, (ii) vane shear test, and (iii) triaxial shear test. In direct shear test the soil sample is put within a rectangular mould, sandwiched between two porous layers. The mould has some holes for drainage purpose. Normal pressure is applied to the mould and the upper-half of the shear test box is moved gradually with the application of horizontal shear force till the sample fails. This failure load is measured for various values of normal loads. The shear strength equation given by Coulomb's law is

$$S = c + \sigma \tan \phi \tag{10.8}$$

where

S is the shear strength
c is cohesion
ϕ is the angle of internal friction
σ is the normal stress on the shear plane.

The friction, which is developed because of interlocking of particles, is contributed by larger soil particles. Thus, the angularity of particles and the degree of compaction affects the value of internal friction for a particular soil. For saturated clays, the angle of internal friction can be assumed to be zero. Clay particles have little friction; they contribute to shear strength in the form of cohesion which is the mutual attraction between the soil particles.

Direct shear test has the following demerits:

(i) Shear stress distribution is not uniform within the sample.
(ii) The area of the sliding surface decreases as the test progresses.
(iii) The horizontal plane of failure is an imposed one.

CBR test

CBR test stands for California Bearing Ratio Test. This test was originally developed by the California Division of Highways prior to the World War II [65]. The CBR test procedure is very popular in many countries for determining the subgrade strength of the soil, because of its simplicity and low cost for conducting the test. The CBR test can also be performed for marginal aggregates. It is an *ad hoc* penetration test whose results are still used to design pavements based on some experience-based curves plotted between CBR, thickness of the pavement, and the number of traffic repetitions. The Bureau of Indian Standards (BIS) publication on laboratory determination of CBR writes, *the test is arbitrary and the results give an empirical strength number which may not be directly related to fundamental properties governing the strength of soils such as cohesion, angle of internal friction, etc.* [109].

For CBR test on the remoulded sample, soil is compacted in the CBR mould (inner diameter 150 mm) with optimum moisture content (determined from standard or modified proctor test). The material should pass through the 20 mm sieve. The larger size

materials, if present, are replaced by an equal amount of material passing through the 20 mm sieve and retained by the 4.75 mm sieve [109]. Compaction of the soil sample is done either by static or by dynamic methods as follows:

(i) In the static method, a calculated quantity of soil mixed with requisite moisture is put into the mould in such a way that after the desired level of compaction it occupies exactly the volume up to the top level of the collar. After initial tamping with a steel rod, a filter paper is put on the soil sample and the displacer disc is placed above it. Soil is compacted with a compression machine until the top of the displacer disc flushes with the top of the collar of the mould.

(ii) In the dynamic method, soil mixed with required moisture content is compacted into the mould in three layers using the standard soil rammer. For subgrade soil intended for pavement construction, heavy compaction (as per modified proctor density) is used for heavily trafficked roads such as national highways, expessways, or major district roads. In other cases, standard compaction is adopted [89].

The mould is kept immersed in water for four days. While immersed, a weight, equivalent to the expected surcharge on subgrade by the pavement, is loaded on the sample. Swelling of the sample is measured, if required, during soaking by a dial gauge fixed over the sample. The mould, after four days of soaking is taken out and water is allowed to drain off. The sample, along with the surcharge, is then subjected to loading in the CBR equipment.

In the CBR test, a plunger of diameter 50 mm penetrates the mould of diameter 150 mm, at the rate of 1.25 mm/min where the soil sample is placed (see Figure 10.3). The load values corresponding to the penetration values of 0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0, 7.5, 10, and 12.5 mm are measured and plotted on a graph. The curve obtained may be of two possible types:

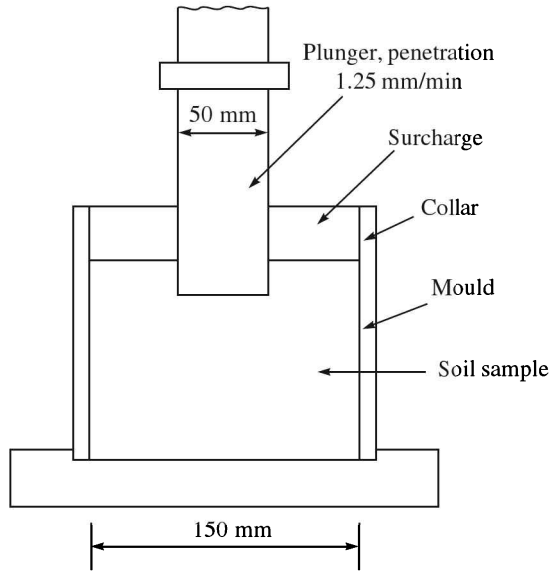


Figure 10.3 Schematic diagram of the mould and the plunger in CBR test.

(i) The smooth curve B (Figure 10.4) in which the penetration increases as the pressure is increased and later the rate of increase of pressure gradually decreases. This case requires no further correction to the plot.

(ii) The load does not increase or increases only slightly with the initial increase in penetration, though later, of course, the load starts increasing. This situation may arise when the loading is slightly inclined or the sample has surface irregularities. Some seating load is thus required after which the soil sample starts taking the load. The nature of the curve obtained is shown in curve A in Figure 10.4. In this case, a correction needs to be applied before proceeding to calculate the CBR value. A tangent is drawn to the point on the curve where the change of direction of curvature takes place. The point where the tangent touches the x-axis is taken as the new origin and axes are shifted to that point [109].

The CBR value is defined as:

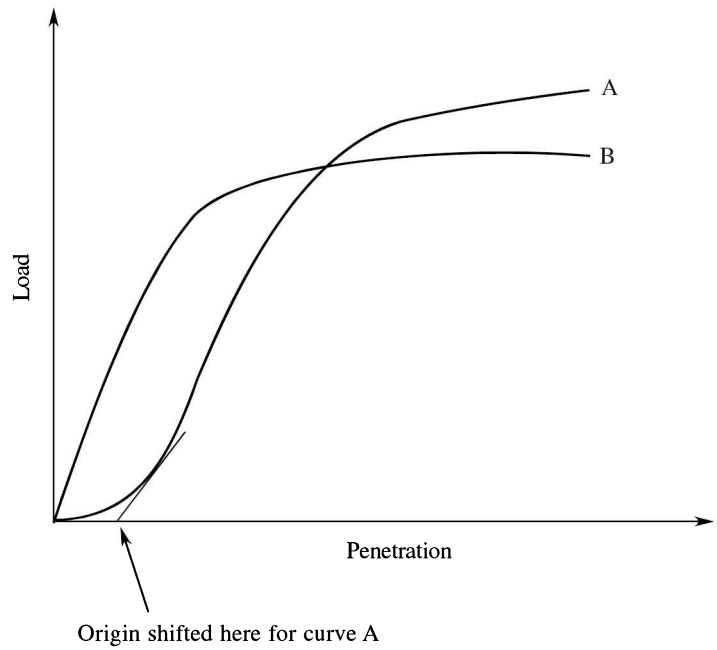


Figure 10.4 Two possible plots of CBR test result.

$$\text{CBR} = \frac{\text{pressure sustained by the specimen at 2.5 mm or 5 mm penetration}}{\text{pressure sustained by the standard aggregates}} \tag{10.9}$$

The standard aggregates are the aggregates from California on which the CBR test was actually evolved and the values of the pressure sustained by them are 70 kg/cm^2 and 105 kg/

cm^2 for 2.5 mm and 5.0 mm penetration respectively [109].

Generally, the 2.5 mm CBR value is found to be greater than the 5 mm CBR value. If it is not so, the test is repeated. If the repetition test also yields the 5 mm CBR value to be greater than the 2.5 mm CBR value, then the 5 mm CBR value is chosen as the CBR value of the sample.

As per the bituminous pavement design guidelines published by the Indian Roads Congress [89], the CBR test should be carried out only on the remoulded sample. In-situ CBR tests are generally not encouraged.

EXAMPLE 10.1

The following are the load values obtained corresponding to the recommended values of penetration from a CBR test on a subgrade soil sample. Calculate the CBR value.

Pen. (mm)	0	0.5	1.0	1.5	2.0	2.5	3.0	4.0	5.0	7.5	10.0	12.5
Load (kg)	0	2.6	5.3	9.7	14.5	19.5	24.2	32.1	38.7	51.5	63.8	74.8

Solution

The load–penetration curve for the CBR test results is shown in Figure 10.5. The figure shows a change in the nature of the curve from concave upwards to convex upwards. Thus, correction to the curve is needed. The tangent is drawn to the point where the direction of curvature changes (i.e. the point at which the slope is maximum). The point at which this tangent intersects the x-axis is the new origin, based on which the CBR values are calculated.

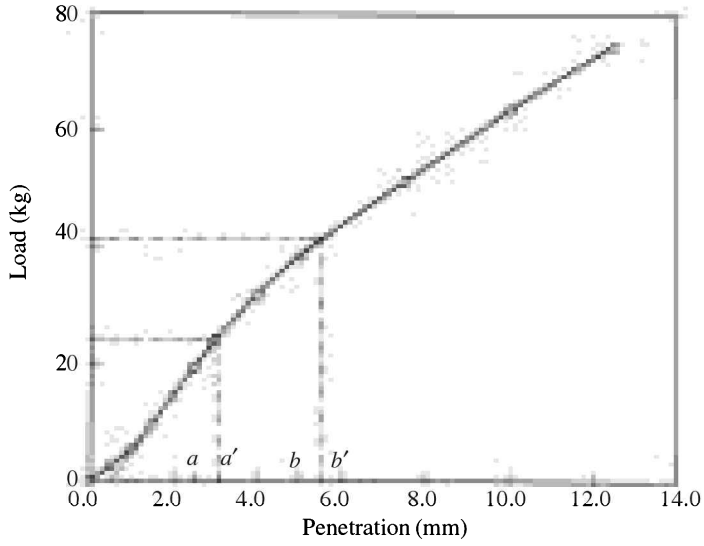


Figure 10.5 Load versus penetration plot for the CBR test.

As a result of this correction, the earlier points '*a*' and '*b*' on the *x*-axis, correspoinding to 2.5 mm and 5.0 mm penetration, shift to *a′* and *b′* respectively as shown in Figure 10.5. The corresponding load values are read as 24.4 kg and 41.6 kg.

Thus, the CBR value corresponding to 2.5 mm (as per new coordinate) penetration is

$$\frac{24.4/[\pi(2.5)^2]}{70} \times 100 = 1.77\%$$

And, the CBR value corresponding to 5.0 mm (as per new coordinate) penetration is

$$\frac{41.6/[\pi(2.5)^2]}{105} \times 100 = 2.02\%$$

(The pressure values are obtained by dividing the load values by the cross-sectional area of the plunger.)

Here, the CBR value corresponding to 2.5 mm penetration is found to be lower than the CBR value for 5.0 mm penetration. Therefore, the CBR test needs to be repeated. If the same values are obtained from the repeat CBR test, the CBR value of the subgrade soil is taken as 2.02%.

Plate load test

The plate load test provides the value of the modulus of subgrade reaction of the subgrade soil and the bearing capacity of soil [110]. It is also used for estimation of elastic modulus of subgrade which is discussed in Section 11.3.1.

The modulus of subgrade reaction k is defined as the pressure sustained per unit deformation of subgrade at the specified deformation or pressure level, with the specified plate size used in the plate load test. Physically, it is similar to the spring constant of soil. Figure 10.6 presents a schematic diagram of a plate load test arrangement.

The reaction frame is loaded with sand bags and the hydraulic jack is adjusted to apply load gradually to the subgrade soil. The plates are grooved to ensure rough interface between soil and the plate [110]. Initially, a seating load of 0.07 kg/cm^2 is applied and then released [110]. Then, a load of 1 kg/cm^2 is applied without any impact. A settlement reading is taken when there is no appreciable settlement or when the rate of settlement is less than 0.02 mm/min. Again, an additional load is applied and the corresponding settlement is noted. Generally, there are a number of dial gauges put in different directions to find the average value of settlement. For the plate load test, loading may be continued up to a settlement of 25 mm (under normal circumstances) [110]. The modulus of subgrade reaction is defined as

$$k = \frac{p}{1.25} \tag{10.10}$$

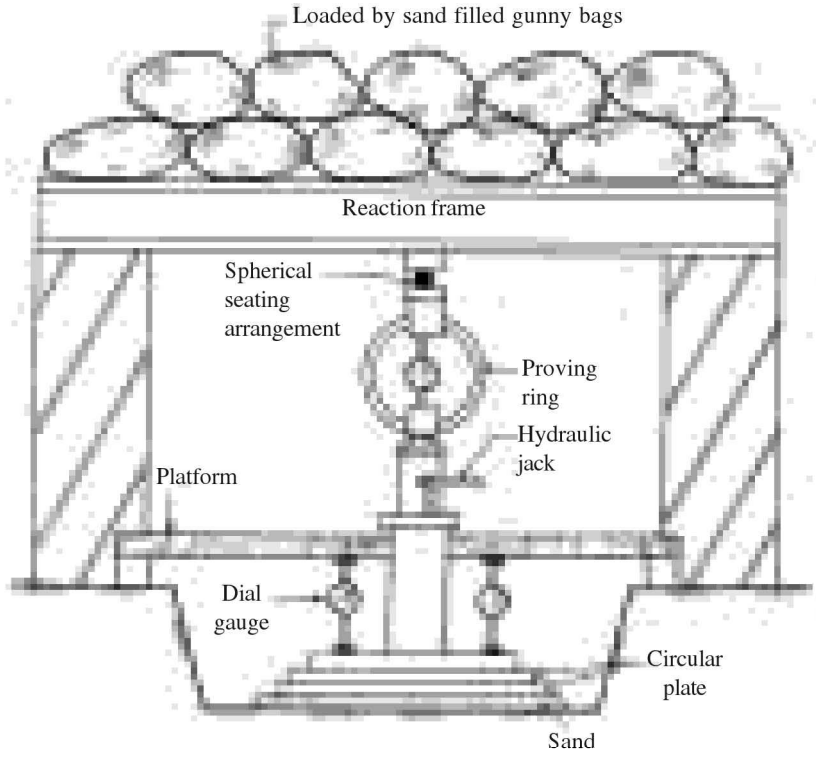


Figure 10.6 A plate load test set-up.

where p is the pressure applied to the plate load test corresponding to the displacement of 1.25 mm.

The unit of k in Eq. (10.10) is MPa/mm. A standard plate size of 750 mm diameter is generally recommended for the plate load test. If a different size of the plate is used, the k value obtained from the plate load test can be approximately converted to the new k value with help of the formula

$$k_1 \times d_1 = k_2 \times d_2 \tag{10.11}$$

where k_1 and k_2 are the moduli of subgrade reaction for the plate load test with diameter d_1 and d_2 respectively.

The IRC guidelines for the design of rigid pavements, IRC:58–1988 [91] recommends the use of the relation, $k_{750} = 0.5 \times k_{300}$, where k_{750} and k_{300} are the moduli of subgrade reaction for 750 mm and 300 mm diameter plates respectively.

The plate bearing test can be individually performed on base, sub-base, or subgrade. One test per kilometre is recommended for the assessment of the k value [91]. Performing a plate load test is costlier than the CBR test. The CBR value of the soil is

used to predict the k value, only if the plate load test data is not available. Generally, the CBR value is used for designing the bituminous pavement and the k value for concrete pavement. It is, however, a matter of convention only. IRC:58–1988 [91] provides the following relationship (Table 10.1) between the CBR and k values.

Table 10.1 k values corresponding to the CBR values for homogeneous soil subgrade [91]

CBR value (%)	2	3	4	5	10	20	50	100
k value (kg/cm^3)	2.08	2.77	3.46	4.16	5.54	6.92	13.85	22.16

For estimation of strength of layered subgrade where the strength of individual layers are different, the plate load test is expected to give a better estimation of subgrade strength (compared to the CBR test), as in this case all the layers involve themselves in sharing the load.

EXAMPLE 10.2

The modulus of subgrade reaction of soil tested with a 350 mm diameter rigid plate is 5.25 MPa/mm. Estimate the modulus of subgrade reaction when a plate of diameter 750 mm is used instead. Also, estimate the pressure that would cause a plate settlement of 1.25 mm when the 750 mm diameter plate is used.

Solution

Using Eq. (10.11),

$$5.25 \times 350 = k_{750} \times 750$$

Therefore,

$$k_{750} = 2.45 \text{ MPa/mm}$$

Thus, by the definition of modulus of subgrade reaction, 2.45 MPa of pressure on the 750 mm diameter plate will cause a settlement of 1.25 mm.

Triaxial test

The triaxial test is conducted on a cylindrical specimen where the length of the sample is generally twice its diameter. A vertical pressure σ_1 is applied from the top and a uniformly distributed pressure σ_3 in the form of fluid pressure, is applied along the curved surface of the specimen (see Figure 10.1). The term $\sigma_1 - \sigma_3$ is the deviatoric pressure, and σ_3 is called the *confining pressure*. In triaxial testing, various values of deviatoric stresses corresponding to σ_3 values, for the failure of the sample due to shear,

are noted, and accordingly the *Mohr rupture envelope* is obtained. This failure envelope is then used to calculate the c and ϕ values of the sample. Depending on various drainage and consolidation conditions, different sets of results may be obtained from the triaxial test. The subgrade conditions in a pavement generally correspond to the consolidated, undrained situation.

In a dynamic triaxial test, the deviatoric stress, $\sigma_1 - \sigma_3$, is made pulsating so as to simulate the traffic loading conditions. This test is used to determine the resilient modulus of soil or of the granular material sample, as explained earlier. If the confining pressure is not applied, the test is called the *Unconfined Compressive Strength* (UCS) test. The UCS tests are generally used for the strength estimation of the cemented materials.

10.3 STONE AGGREGATES

Stone aggregates are one of the major components of road structure. They are used in the bituminous or in the concrete layer, or in other bound form. Unbound aggregates are used for base or sub-base course. Aggregates bear load due to particle interlocking and sustain the wear and tear due to vehicular movement.

10.3.1 Source

Aggregates are of three types by origin:

(i) Igneous (acidic igneous rocks, e.g. granite; basic igneous rock, e.g. basalt)

(ii) Sedimentary (chalk, limestone, etc.)

(iii) Metamorphic rock (schist, gneiss, etc.)

The constituents of rock may be feldspar, quartz, iron oxide, calcite, dolomite, mica, gypsum, and so on. Aggregates are processed in stone quarries. The rock at the quarry face is broken into smaller parts either by earth moving equipment or by explosion. It is further broken into smaller pieces of suitable gradation using gyratory or impact crushers. The aggregates are screened through different sieves and batched for various construction requirements. *Scalping* is a process by which undesirable materials are removed from the aggregate batches.

Table 10.2 presents a summary information on the availability of various naturally occurring rocks suitable for highway construction in different regions of India [51]. Various pavement layers mentioned in the table are discussed in detail in Chapter 11 (Section 11.2).

Table 10.2 Availability of various types of aggregates in India [51]

Type of Rock	*Geological group*	*Properties*	*Suitability*	*Location/Availability*
Basalt	Igneous	Hard, durable and resistant to abrasion; fine grained	Good for base and surface courses	Maharashtra, Bihar, Gujarat W.Bengal, and M.P.
Granite	Igneous	Hard, durable, resistant to abrasion, coarse grained and quite brittle	Very good for bituminous courses and WBM	J.&K., Tamil Nadu, Punjab, Rajasthan, U.P., M.P., Assam, Karnataka, W.Bengal, Maharashtra, Bihar, Orissa, Kerala, Gujarat
Lime-stone	Sedimentary	Hard but liable to get polished due to traffic, high water absorption, excellent adhesion to bitumen, fine grained	Good for base courses	Maharashtra, W.Bengal, Rajasthan, A.P., Andaman Islands, Bihar, H.P., M.P., and U.P.
Quartzite	Sedimentary and Metamorphic	Hard, durable, but brittle; poor adhesion to bitumen	Good for base courses	W.Bengal, A.P., H.P., Tamil Nadu, U.P., Karnataka, Gujarat, Punjab, and Rajasthan
Sand	Sedimentary	Moderately hard and durable; fine to medium grained	Good for road bases	A.P., M.P., W.Bengal, Punjab, Bihar, Rajasthan, H.P., Andaman Islands, J.&K., U.P., and Tamil Nadu

10.3.2 Characterization

Stone aggregates, also called granular materials, when not bound by any cementing material, show strength only under confinement. The relationship between M_R and the confining pressure σ_3 for the granular materials may be put forth [22] as

$$M_R = k_1 \sigma_3^{k_2} \tag{10.12}$$

where k_1 and k_2 are the material constants.

Typical values of k_1 and k_2 for partially saturated granular material with nonplastic filler for samples collected from eastern zone of NH-6 are 3.18 and 0.73 respectively [179]. As an alternative formulation, Monismith et al. [168] suggested the following equation

$$M_R = k_1 (\theta)^{k_2} \tag{10.13}$$

where

θ is the first stress invariant (= $\sigma_1 + 2\sigma_3$ in this case)

k_1 and k_2 are the material constants.

Stress invariants are certain combinations of the stress components which remain unchanged irrespective of the choice of the coordinate system [48, 241]. Equation (10.13), known as the *k-θ model*, has been widely used for the characterization of granular

materials and is supported by data obtained from dynamic triaxial tests with constant confining pressure.[2]

Equations (10.12), (10.13), and (10.14), are nonlinearly stress dependent. Its implication could be understood in the following way. A granular layer, within a pavement, is thought of where the M_R values of individual elements are assumed to be equal to some fixed value initially. The analysis of the pavement structure would show that stresses at various elements are different, depending on their locations with respect to the load point. Now, by using any of the Eqs. (10.12), (10.13), and (10.14), a new set of M_R values of the individual elements are obtained, which will be found to be different from each other. The analysis is repeated and a different set of stress values other than the earlier analysis, is obtained. This forms an iterative procedure and, therefore, finite element method or some other technique may be employed in such a case till the solution converges. This also justifies why equivalent thickness conversion used in pavement design leads to further approximation[3]. Equivalency between the layers may be established by equating stress or displacement at a critical point, however, because of this nonlinearity, the values (stress or displacement) of points other than the point considered would not match.

Because of this inherent complication and computationally exhaustive nature, simpler formulations are made available from practical application considerations. A widely accepted expression for computation of the M_R value of granular material is given as [38]

$$\frac{M_{R_{subbase}}}{M_{R_{subgrade}}} = 0.2 \times (h)^{0.45} \tag{10.15}$$

where $M_{R_{subbase}}$ and $M_{R_{subgrade}}$ are the M_R values of granular sub-base and subgrade respectively and h is the thickness of the granular layer (mm).

---

[2]However, in a realistic situation, both the vertical and the lateral stresses are pulsed as the wheel load passes. Secondly, the granular materials are often in a partially saturated state and, therefore, the effective stress in the pavement layer is different from what is applied [18]. Thirdly, M_R should increase with an increase in the value of repeated deviatoric stress, but for deviatoric stress less than 70 kPa, an opposite behaviour was observed by some scientists (May and Witczak [155], Shackel [205] and Uzan [250] etc.). Thus, an additional term was introduced in Eq. (10.13) to account for the effect of shear on the M_R value of the granular materials. The modified equation is

$$M_R = k_1 \theta^{k_2} \tau_{oct}^{k_3} \tag{10.14}$$

where

τ_{oct} is the octahedral shear stress

k_1, k_2, and k_3 are the material constants.

There are various other formulations proposed for finding the M_R of granular materials.

[3]In equivalent thickness conversion, some factors are assigned, which are used to convert the thickness of a given pavement layer into the equivalent thickness of some other layer.

Equation (10.15) is used by Shell [38] for characterization of granular layer and the ratio of moduli of granular and subgrade is found to range between 2 and 4. Equation (10.15) is also used in the IRC guidelines for the design of bituminous pavements [89].

10.3.3 Tests on Aggregates

The aggregates are tested for engineering properties to assess their suitability as road construction materials. Various tests on aggregates have been formulated keeping in view the following parameters:

(1) *Strength:* The aggregates should be sufficiently strong, so that they can bear the traffic load without getting crushed. The name of the equipment used to measure the strength of aggregates is known as the Crushing Strength Testing Equipment.

(2) *Hardness:* The aggregates undergo continuous wear and tear under the wheels of vehicles. They also get rubbed with each other due to application of the traffic load. These two phenomenon are known as *abrasion* and *attrition* respectively. Los Angeles test, Deval's abrasion test, Polished stone test are the tests under this category.

(3) *Toughness:* Toughness of a material is its ability to sustain impact loading. The aggregates on an in-service road are also subjected to impact loading due to the vehicle movement. Toughness of aggregates is measured by the *impact test.*

(4) *Durability:* Gradual deterioration of aggregates takes place as they are continuously exposed to the environment. This aspect, called durability of aggregates, is tested by *soundness tests.*

(5) *Shape:* Aggregates have varied shapes. Depending on the specific construction purpose, angular or rounded aggregates are recommended for highway construction. Too flaky or elongated aggregates are generally not encouraged for construction purposes. Flakiness index, elongation index, and angularity number are some of the indices used to determine the shape characteristics of aggregates.

(6) *Water absorption:* The water absorption test determines the tendency of aggregates to absorb water.

(7) *Adhesion with bitumen:* When aggregates are mixed with bitumen, a thin film of bitumen is formed over the aggregates, which helps in holding the whole mass together. Aggregates show a varied degree of affinity towards bitumen and water. If the aggregates have a relatively higher affinity towards water, bitumen is sometimes stripped off from the aggregates leading to failure of the bituminous mix. *Stripping test* is done to check the adherence between aggregates and bitumen.

The basic philosophies of the various possible tests on aggregates have been discussed above. We will now discuss some of the characteristic tests carried out on aggregates.

Crushing strength test

In the crushing strength test, a total load of 40 tonne is applied at the rate of 4 tonne per minute on stone aggregates (passing 12.5 mm and retained on 10 mm sieve, surface dry condition) kept in a mould. The aggregates in the mould are placed in three layers by tamping each layer 25 times. They are then sieved through a 2.36 mm sieve. The percentage weight value passing through the 2.36 mm sieve with reference to the total weight of aggregates gives the measure of crushing strength of the aggregates [111]. The mean of the two such results rounded off to the nearest whole number gives the aggregate crushing value.

The 10% *fines test* is an iterative test used to find the load required to achieve a crushing strength of 10%. As per the provisions of the Code [111], the 10% fines test is recommended for the materials whose crushing strength value is 30 or above as anomalous results may be obtained for the crushing strength value at and beyond 30.

Abrasion test

In the abrasion test, the aggregates are subjected to abrasion in a shielded (dust-tight) container of specific dimensions with a specific number of steel balls and individual weights kept inside. The container rotates at a given frequency for a given duration. The abrasion test results tend to show a correlation with the impact test value for a given type of aggregates [161].

Los Angeles test. In Los Angeles test, a cylindrical container (500 mm width and 700 mm diameter) rotates at 20–33 rpm along its axis for 500–1000 times, depending on the aggregate gradation. A measured quantity (5 or 10 kg depending on the gradation) of aggregates is put inside the cylinder. Specified number of steel balls (depending upon the gradation of aggregates), each of 48 mm diameter and of varying weights (390–445 g), are put as abrasive charge into the rotating cylinder. When the test is over, the aggregates are sieved through a sieve size of 1.7 mm, washed and dried in an oven at 105–110°C. The difference in the weights (before and after the test), expressed as a percentage of the original weight, is the Los Angeles abrasion value [111].

Deval's abrasion test. In Deval's abrasion test, there are two cylinders inclined at 30° with respect to the horizontal plane. Six cast iron balls of diameter 48 mm each (of total weight 2500 g) are used as the abrasive charge. After 10,000 revolutions at a speed of 30–33 rpm, the aggregates are taken out and sieved through a sieve size of 1.70 mm, washed and dried. The loss in weight with reference to the original weight of the test sample is expressed as the Deval's abrasion value [111]. Figure 10.7 shows a Deval's abrasion test equipment installed in the laboratory.

Polished stone test. Aggregates' surfaces tend to get polished due to traffic. The degree of polish imparted depends on the traffic conditions (such as volume, weight, speed,

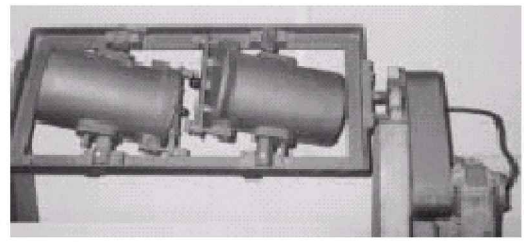


Figure 10.7 Deval's aggregate abrasion test equipment.

acceleration, braking), on the road geometry (such as alignment, grading), and on the nature of the aggregates.

Aggregates are tested in the laboratory to determine their susceptibility to polishing through an *accelerated polish test* (Figure 10.8). The test procedure was first developed in the UK and is known as the *British Polish Stone Test* [143]. In this test, the aggregates of acceptable size are pasted on a curved tile by means of mortar such that the aggregates protrude out of the mortar level. Many such samples (fourteen in number) placed on the curved tiles are mounted around the periphery of a steel wheel. Wheels with pneumatic tyres are allowed to rotate (320–325 rpm) over this steel wheel maintaining a given normal force (40 kg) between them. Emery powder and water at specified rate are sprinkled over the surface of tyre and the specimen wherever they come in contact. Thus, the aggregates so impregnated on the curved tiles are subjected to accelerated polishing for a given duration (3 hours) after which they are taken out from the steel rim [143, 111] and tested for skid resistance in the British Pendulum Tester, the details of which have been discussed in Section 14.2.2. After applying a correction factor for the curvature of the sample, the resulting parameter obtained is the *Polished Stone Value* (PSV).

Figure 10.8 An aggregate polish test equipment.

Impact test

Aggregate impact value gives the relative strength of aggregates against impact loading [111]. In the impact test, an impact load of 13.5 to 14.0 kg of weight is allowed to fall 15 times from a height of 38 cm on to the aggregates (sieved through 12.5 mm sieve but retained by 10 mm sieve) placed in a mould (placed in the three layers by tamping each layer 25 times). The percentage of those passing through the sieve size of 2.36 mm with respect to the total weight of aggregates is calculated and expressed as the impact test value of the aggregates. The average of two such values rounded off to the nearest whole number is referred to as the aggregate impact value [111]. Figure 10.9 shows a photograph of an aggregate impact test set-up.

Soundness test or Sulphate resistance test

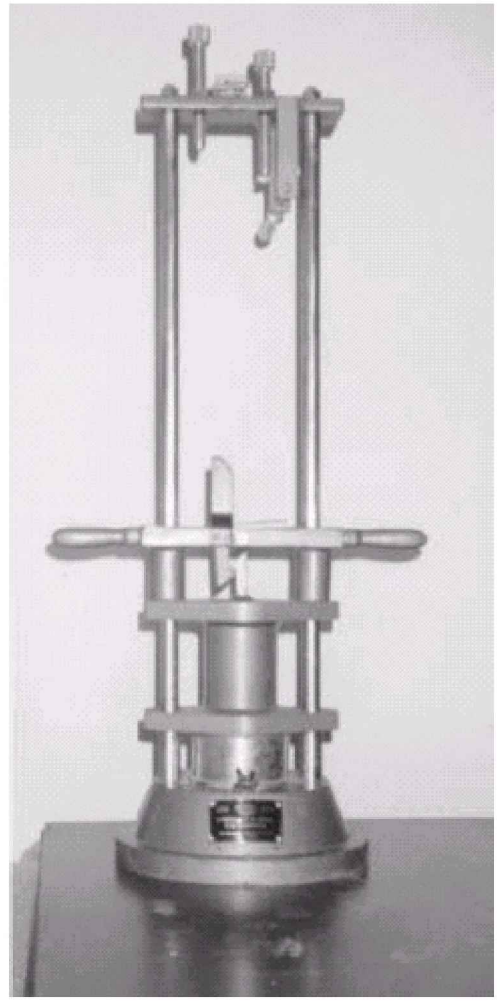


Figure 10.9 The aggregate impact test set-up.

This test measures the resistance of aggregates to disintegration caused due to weathering. In this test, the aggregates are subjected to alternate cooling and heating cycles in the presence of sulphates as abrasive agents. The sulphate solution penetrates the fine cracks in the aggregates. Due to repetitive drying and wetting, sulphate salts get deposited in these cracks and grow in size, causing disintegration of aggregates along the weak shear planes [143].

In this test, the aggregate samples are weighed and immersed in a solution of sodium sulphate or magnesium sulphate. After 16–18 hours of immersion in sulphate solutions, the aggregates are allowed to drain for 15 minutes and then oven dried at 105–110°C [112], thus, completing one cycle. The average loss in weight after 10 such cycles should not exceed 12% (for test with sodium sulphate) or 18% (for test with magnesium sulphate) for aggregates to be of recommendable quality [215].

Shape test

Flakiness Index. The Flakiness Index of aggregates is the percentage of that mass of the aggregates whose least dimension is less than 0.6 time their mean dimension. The Flakiness test is applicable to the aggregates of size larger than 6.3 mm [113].

The aggregate samples are first sieved through specified sieve sizes. 200 aggregates of any fraction (say, those passing through the sieve size x and retained by sieve size y) are then chosen. The selected aggregates are next individually made to pass through the gauge (slots close to rectangular shape), specified for that size of fraction, chosen from the metal gauges intended for flakiness testing. The slot width is 0.6 times the mean dimension of the aggregates (i.e. $0.6 \times (x + y)/2$). The aggregates passing through the particular slot are identified as the flaky aggregates for that size fraction. Similar exercises are carried out for the aggregates of other size fractions.

The mass of aggregates that passes through the corresponding gauge for each size of fraction is calculated as a percentage of mass of the total number of aggregates of the respective sieve fraction (say, p_i). And the mass of the total number of aggregates in an individual size fraction is expressed as the percentage of the total mass of aggregates taken for the study (say, q_i). Then, the Flakiness Index (FI) is obtained as the sum of weighted percentage of the aggregates passing through the appropriate gauge, i.e.

$$\text{FI} = \sum_{i=1}^{m} p_i q_i$$

where m is the number of size fractions [113]. This is in fact equal to the total weight of flaky aggregates divided by the total weight of the aggregate sample taken for the study.

Elongation Index. The elongation test, too, is not applicable to aggregates of size less than 6.3 mm [113]. The elongation index of aggregates is the percentage by weight of the particles whose longest dimension (i.e. length) is greater than 1.8 times their mean dimension. The test procedure is the same as that for the flakiness index, with the exception that in this case a different type of a metal gauge (length gauge) is used as the aggregates are tested for their longest dimension. The elongation index, similarly, is the total weight of the aggregates retained in the respective length gauges, expressed as a percentage of the weight of the total aggregates gauged [113].

Angularity number. The angularity of aggregates is the converse of roundness [113]. In this case, oven-dried aggregates of appropriate size (passing through sieve size x and retained by sieve size y), as recommended by IS code [113] are chosen. The aggregates are filled in three layers in a metal cylinder of specified dimensions closed at one end. A tamping rod is used to apply 100 blows to each layer while filling, and the top surface is levelled. Tamping should be done carefully to avoid breaking of aggregates as otherwise they are adjudged unsuitable for angularity testing [113]. The angularity number (AN) is determined as

$$\text{AN} = 67 - \frac{W_a}{W_w \times G_a} \times 100 \tag{10.16}$$

where

W_a is the weight of the aggregate mass

W_w is the weight of water occupying the mould

G_a is the specific gravity of the aggregates.

The AN value is rounded off to the nearest whole number [113]. The number, 67, is a reference number which signifies the maximum attainable density level (100 minus voids) with spherical aggregates of the same size as that of the aggregates under test. The more angular the shape of the aggregates, the more will be voids, resulting in an increase in the angularity number. The aggregates suitable for bituminous pavement construction should generally be angular in shape for better interlocking. In most of the type of pavement construction with bituminous mixes, it is generally recommended [215] that the aggregates should have at least two fractured surfaces. The angularity of aggregates, however, makes the mix less workable. For this reason, less angular aggregates are preferred for concrete construction.

The Ministry of Road Transport and Highways (MORT&H) Specifications for Roads and Bridge Works [215], uses another parameter, namely the 'combined flakiness and elongation index', as a suitability measure of shape of the aggregates. This parameter is determined by first finding the flakiness index of the aggregates of the representative sample. The nonflaky aggregate particles are separated out and the elongation index of only these aggregates is calculated. The values of flakiness and

elongation indices, thus found, are added up to calculate the 'combined flakiness and elongation index'.

Specific gravity and water absorption of aggregates

Concepts of two different specific gravities of aggregates, namely (i) the bulk specific gravity and (ii) the apparent specific gravity are generally used for practical purposes. Their definitions are explained with the help of Figure 10.10, which shows that an aggregate assumes two weights, A and B, at oven dry and saturated (with water) surface dry conditions respectively. If the weight of the aggregates in water is C (not shown in Figure 10.10) then the bulk specific gravity is $A/(B - C)$ and the apparent specific gravity is $A/(A - C)$. It may be noted that in the saturated surface dry condition, water occupies the permeable pores in the aggregate. This does not necessarily mean that all the aggregate pores have been occupied by water [Figure 10.10(b)].

(a) Oven dry aggregate [weight = A]

(b) Saturated surface dry conditions [weight = B]

Figure 10.10 The concept of bulk and apparent specific gravity of aggregates.

The specific gravities of the aggregates can now be defined as:

The bulk specific gravity of an aggregate at a given temperature is the mass of the aggregate divided by the mass of equal volume of distilled water at that temperature.

The apparent specific gravity of an aggregate at a given temperature is the mass of the impermeable portion of the aggregate divided by the mass of equal volume of distilled water at that temperature.

Based on the above concept, Indian specifications for determination of specific gravities [114] of aggregates have enumerated various recommendations for aggregates of different sizes. For some size of aggregates, they can be directly weighed, when immersed in water, with the help of a weighing balance (with sample suspension facility). The smaller size of aggregates can be weighed using specific gravity pychnometer. Oven drying as per Indian specifications [114] should be done at 100–110°C for 24 hours. The percentage water absorption is found out by calculating $100 \times (B - A)/A$.

Stripping value test

As per the recommendation [115] for the stripping value test, 200 g of aggregates, passing 20 mm and retained 12.5 mm, is heated up to 150°C. Bitumen is also mixed in a separate container at 160°C. Aggregate and bitumen (5% by weight of dry aggregates) are thoroughly mixed so that aggregates get properly coated with bitumen. Bitumen-coated aggregates are put in a 500 ml beaker in normal air temperature and allowed to cool down for two hours. Distilled water is added to the beaker to immerse the aggregates completely and the whole set-up is kept for 24 hours in a water bath maintained at 40°C. After 24 hours, the individual aggregates are visually tested to observe the percentage of surface area of the aggregates from where the bitumen has stripped off. This exercise is repeated for all the aggregates. The average of three such experimental results rounded off to the nearest whole number [115] is taken as the stripping value of the aggregates and bitumen considered.There are various other tests to study the moisture sensitivity of bituminous mixes.

Closing remarks

Various physical property tests of aggregates used for highway construction have been discussed in this subsection. The list of tests discussed is not exhaustive as there are a large number of other tests recommended for various other specific applications such as Micro Deval's abrasion test, British abrasion test, Freeze-thaw soundness test, Washington degradation test, Texas ball mill test, Sand equivalent test etc. Some of these tests are application specific and developed by local agencies in various countries.

The code of practices recommend various acceptability limits of the aggregates. These acceptability limits may vary depending upon the type of construction, for example, strong materials may be needed at the top layers where the stresses are likely to be high, whereas weaker materials may be used at the bottom layers of a pavement. Table 10.3 gives an example of the acceptability limits of some of the properties of aggregates suitable for highway construction, recommended by the Ministry of Road Transport and Highways (MORT&H) Specifications for Road and Bridge Works, Government of India [215].

Table 10.3 An example of acceptability limits of physical properties of aggregates for Bituminous Macadam construction [215].

Test	*Test method*	*Requirement*
Los Angeles abrasion value	IS:2386 (Part IV)	40% maximum
Aggregate impact value	IS:2386 (Part IV)	30% maximum
Combined flakiness and elongation index	IS:2386 (Part I)	30% Maximum
Aggregate stripping value	IS:6241	25% Maximum
Soundness (5 cycles)		
loss with sodium sulphate test	IS:2386 (Part V)	12% Maximum
loss with magnesium sulphate test	IS:2386 (Part V)	18% Maximum
Water absorption	IS:2386 (Part III)	2% Maximum

10.3.4 Aggregate Gradation

A stockpile of aggregates, obtained from stone crusher, contains aggregates of different dimensions in varied proportions. Aggregates are sieved through a standard mesh system and the cumulative per cent passing value is plotted against the sieve sizes (generally, in logarithmic scale) and a curve is obtained. This curve is known as the *aggregate gradation* (or *particle size distribution*) *curve*. Figure 10.11 shows some forms of the

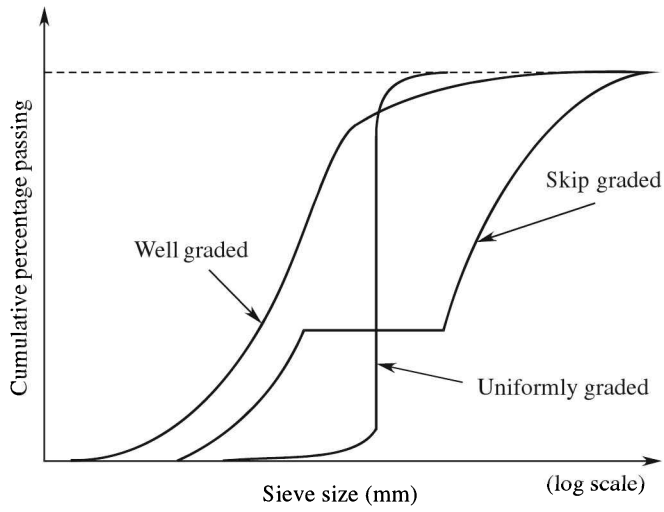


Figure 10.11 Various possible forms of the aggregate gradation curve.

aggregate gradation curve. The curve is called *uniformly graded* if aggregates are mostly of single size; it is called *skip graded* if the aggregates corresponding to specific sieve size(s) are missing; and it is called *well graded* if the aggregate sizes are distributed in different proportions in such a way that the cumulative curve takes a nice 'S' shape.

Voids in aggregates

The total volume of broken pieces of aggregates always shows a larger value than the volume of the original boulder from where the pieces are derived, upon crushing. This is due to voids created within the broken pieces. The term *void ratio* is defined as the ratio of the volume of voids to that of the solid pieces and the *packing ratio* is defined as the volume of solid pieces to the total volume.

Efforts have been made to theoretically model the void ratio when the aggregate gradation curve is known. If the aggregates are ideally considered to be spherical and of equal size, various arrangements, as shown in Figure 10.12, are possible. Recently [107], it has been proved that the 'closed pack face centred cubic lattice' has the

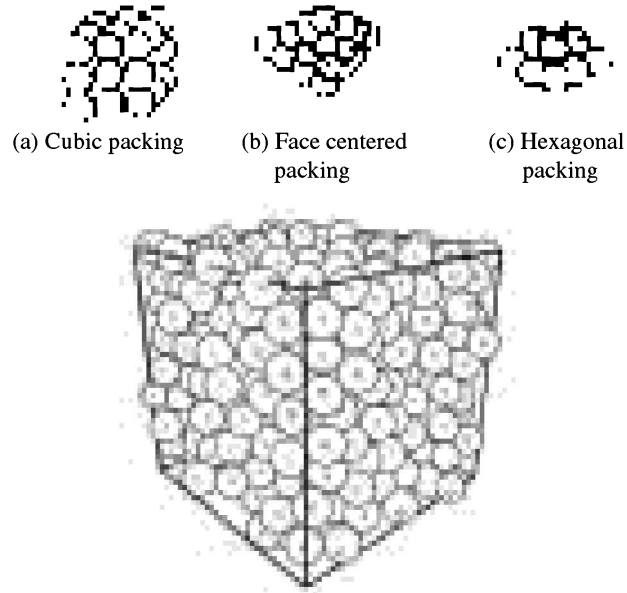


Figure 10.12 Possible arrangements of spherical aggregates of equal diameter.

minimum void[4] and a packing ratio of $\pi/\sqrt{18}$. Prediction of the resultant void ratio for aggregates of random shape and size distribution is extremely difficult due to mathe-matical complexities [222, 203, 4]. Therefore, experimental or computer simulation results are used for all practical purposes.

When aggregates of two different sizes are blended, generally a resultant void ratio less than the void ratio of each of the two individual components is obtained. Only for a particular proportion of the two, the void ratio attains a minimum value. This minimum void ratio also depends on the size ratio between the aggregate particles, which may be defined as the ratio of the small to large aggregate sizes. If the size ratio is low, the void ratio varies sharply with the variation in proportion of aggregates. Figure 10.13 presents these variations of void ratio when binary aggregates (of various size ratios) are mixed in different proportions.

The objective of good aggregate gradation is to achieve the desired density and overall strength of the aggregate mix through frictional forces at the contact points, by proper proportioning of their size distribution. Figure 10.14 shows the structure of a mix of binary

[4]This is an interesting and commonly faced problem by the reserchers from various disciplines. Kepler, way back in 1611, postulated that cubic or hexagonal arrangement is possibly the most dense arrangement with spherical equal size particles. This is known as *Kepler's conjecture*, which could not be proved until recent past. Experiments were done with ball bearings or similar objects, which were shaken, sunk in oil, kneaded in a rubber balloon, and the maximum packing ratio obtained was 0.637 or less [243]. Only in recent times, rigorous mathematical proof [107] has established the theoretical value of packing ratio to be $\pi/\sqrt{18}$.

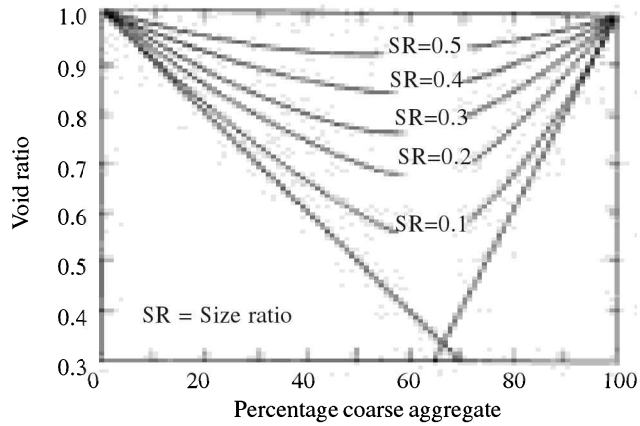


Figure 10.13 Variation of void ratio with size ratio and relative proportions in a binary mix [43].

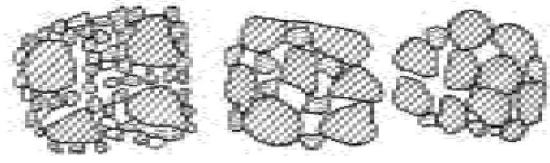


Figure 10.14 Possible mix structures for binary aggregates [199, 268].

aggregates where three cases may arise depending on their proportion and size ratio. The 'balanced stage' is the most desirable situation.

Thus, one can develop relative proportions of a given choice of aggregates of different dimensions in order to achieve the maximum possible density by mixing them one by one, or by any other trial and error method. Based on experimental results, Fuller [64], in 1907, first postulated an equation of gradation curve which resulted in maximum density. Since then, a large number of experimental studies have been undertaken which show that the Fuller's maximum gradation curve is mostly applicable to the aggregates of various shapes, sizes, and proportions. The latest is the Superpave [224, 223] technology to have used the Fuller's equation as the base curve for its recommendations on the choice of aggregate gradation. Various considerations in Superpave are briefly discussed in Annexure II of this chapter. The Fuller's equation is given by

$$p_d = \left(\frac{d}{D}\right)^n \tag{10.17}$$

where

p_d is the 'percentage passing' corresponding to particle size d
D is the maximum particle size considered
n is a constant varying from 0.45 to 0.50.

Closing remarks

A very dense packing of aggregates may not always be desirable for bituminous road. A certain amount of void space is mandatory for bituminous mix so that there is enough room for bitumen to be able to impart greater durability to the mix. Air voids are essential to take care of the additional compaction caused by traffic and to prevent bleeding of bitumen on the road surface.

Various types of aggregate gradations are recommended for use in India [215], which have been evolved from experience gained about their performances. These specifications are known by different names, such as Bituminous Concrete (BC), Dense Bituminous Macadam (DBM), Bituminous Macadam (BM), and so on. These are discussed in detail in Sections 13.8.5 and 13.8.6. Table 10.4 shows an example of how the recommended gradations are different for different mixes. For example, cumulative percentage passing of BC grading-I corresponding to sieve size 13.2 mm should lie between 79 to 100, whereas for DBM grading-I the range is 55 to 75. Figure 13.14 also shows the same thing where the gradations of BC and Semi-dense Bituminous Concrete (SDBC) are compared graphically.

Table 10.4 Comparison of gradations of BC, DBM and BM [215]

Sieve size	*Cumulative percentage passing*		
	BC (Grading I)	DBM (Grading I)	BM (Grading I)
45.0 mm	—	100	100
37.5 mm	—	95–100	90–100
26.5 mm	—	63–93	75–100
22.4 mm	—	—	60–95
19.0 mm	100	—	—
13.2 mm	79–100	55–75	35-61
11.2 mm	—	—	—
9.5 mm	70–88	—	—
5.6 mm	—	—	—
4.75 mm	53–71	38–54	13–22
2.8 mm	—	—	—
2.36 mm	42–58	28–42	4–19
1.18 mm	34–48		
600 μm	26–38		
300 μm	18–28	7–21	2–10
150 μm	12–20		
90 μm	—	—	—
75 μm	4–10	2–8	0–8

10.3.5 Batch Mixing Problem

An aggregate batch is generally designated by its maximum size. An aggregate batch, of a given maximum size, obtained from a particular quarry, shows a specific gradation. Another aggregate batch with a different maximum size will have a different gradation. Also, it could be that an aggregate batch of the same size but from a different quarry would have a different gradation. Perhaps, none of these gradations will conform to the gradation range recommended for a particular type of construction. Thus, it becomes necessary to mix various aggregate batches in certain proportions to achieve a gradation which lies within the range of upper and lower limits of the specified gradation. This is known as the *batch mixing problem* whose objective is to find suitable proportions between the aggregate batches, so that the resultant gradation shows the best possible match with the specified gradation.

Let us consider a hypothetical case where there are n number of batches available and a_{ij} denotes the cumulative percentage passing for the ith sieve size of the aggregates in the jth batch. It is assumed that there are m such sieve sizes. If the batches are mixed in proportions of p_j, the cumulative percent passing of the resultant mix for the first sieve size becomes $a_{11}p_1 + a_{12}p_2 + a_{13}p_3 + \cdots + a_{1n}p_n$. This value should lie between the upper and the lower specified gradations, say s_1^u and s_1^l respectively. Thus, the whole batch mixing problem could be presented in matrix form (variables are replaced by capital letters) as:

$$[A]\{P\} \leq \{S^u\} \tag{10.18}$$

$$[A]\{P\} \geq \{S^l\} \tag{10.19}$$

Some methods employed to solve this set of equations are as follows:

- Trial and error methods (graphical, analytical, or computational) are used where various trial proportions are mixed and checked for the resultant gradations. The trial and error methods work fine when the number of batches are not more than three, but the process becomes difficult to handle when the number of batches increases.
- The above equations may be transformed into equivalent equality equations by assuming that the resultant gradation is expected to be close to the mid-point of the specified gradation. That is, the above set of equations take the form $[A]\{p\} \approx \{(S^u + S^l)/2\}$. This is a system of redundant equations, where the number of variables (i.e. batch proportions) are less than the number of equations (i.e. number of sieve sizes). This system of equations can be solved by the least square technique which sometimes gives excellent results. But the least square technique has got no control over one or some of the proportions being negative (though $\Sigma_{j=1}^{n} p_j = 100$ is satisfied).
- The linear programming approach possibly shows a better performance (if at all a solution exists) than the other two. One more inequality constraint $\Sigma_{j=1}^{n} p_j \leq 100$, needs to be added to the above set of equations. An objective

function of maximization of $\Sigma_{j=1}^{n} p_j$ is chosen, and this should be equal to 100 if a solution exists. By this technique, aggregate batch mixing taking into account some other properties like fineness modulus, or cost considerations, can also be incorporated as mixing conditions [57, 58]. If the solution is not found for the aggregate batches considered, one or some of the batche(s) can be omitted, and calculation can be repeated.

10.4 BITUMINOUS MATERIAL

10.4.1 Source

Bitumen is derived from fractional distillation of petroleum. Naturally occurring bitumen is commonly known as asphalt. Rock asphalts are porous limestone and sandstone rocks into which natural asphalt is impregnated. Rock asphalts are found in Italy, Sicily, and California and these materials are excellent from the road construction point of view. There are asphalt lakes, too, found in Trinidad and Iraq and are called *lake asphalt* [143]. Tar is produced from destructive distillation of coke, and pitch is produced from fractional distillation of tar [143]. Tar is not encouraged to be used for bituminous pavement construction nowadays, because of its low durability and high temperature susceptibility.

The terms bitumen and asphalt are generally used synonymously. In the USA [237], Australia [182] and the UK [245], the term 'asphalt' is used. In India, earlier, MOST (Ministry of Surface Transport) specifications [216] used the term asphalt which has now [217, 215] been replaced with the term bitumen. This book uses the two terms interchangeably.

The components derived from fractional distillation of petroleum, at various temperature levels, are (i) gas, (ii) naptha, (iii) kerosene, (iv) diesel and lubricating oil, (v) bitumen and furnace oil, and (vi) residue. This bitumen is known as *penetration grade bitumen* because the specification or the grade of this bitumen, by which it is designated, is obtained from the penetration test, described later in Section 10.4.5. There could be two other forms of bitumen *emulsion*, where bitumen is in suspension form as small globules in water, and *cutback*, where bitumen is dissolved in suitable solvents (see Section 10.4.4). In bituminous construction, the choice between penetration grade bitumen, bitumen emulsion, or cutback bitumen is made depending on factors like, weather conditions, availability, economy, and available construction time. These three types of bitumens are in general referred to as *bituminous binders*.

10.4.2 Composition

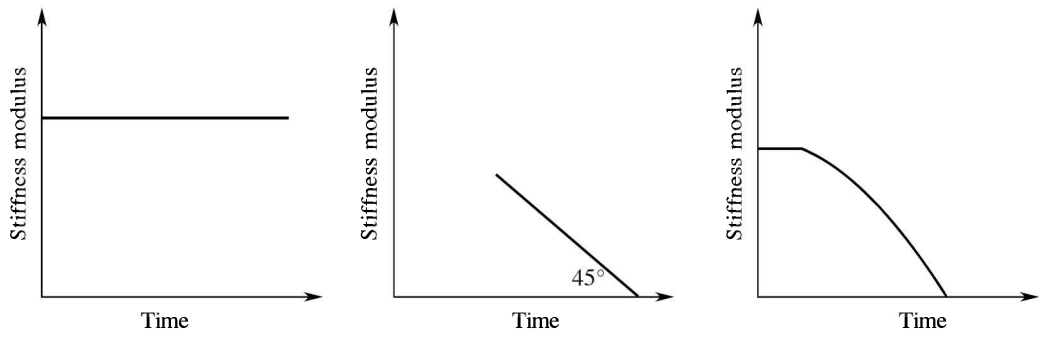
Bitumen is basically a hydrocarbon. Less than 10% by weight is due to atoms of sulphur, nitrogen, and oxygen which are attached to the hydrocarbon molecules. Carbon content in bitumen is 80–87% by weight. Three basic components of bitumen are

(i) asphaltene, (ii) maltene, and (iii) carbene [143]. Asphaltene is hard, relatively inert and aromatic. Maltene is a solvent by nature and imparts viscoelasticity to bitumen. It is resin like intermediate molecular hydrocarbon. Carbene is the fraction which is insoluble in carbon tetrachloride (CCl_4) [143]. The chemical bonds in bitumen are weak and break when heat is applied. When bitumen is cooled, it comes back to original structure, but not necessarily the same structure as before [224]. The phenomenon of changing chemical structure of bitumen with heating has made the understanding of behaviour of bitumen very complicated.

10.4.3 Characterization

Bitumen as a material has drawn attention of the engineers since a long time because it is (i) waterproof, (ii) durable, (iii) resistant to strong acids, and (iv) possesses good cementing properties. At normal temperature, bitumen is semi-solid, that is, it takes time to flow. At higher temperatures, bitumen behaves like a viscous liquid, whereas at a very low temperature bitumen is as brittle as glass. Bitumen is believed to behave 'viscoelastically' at the standard operating temperature of highways. Theoretically, it is difficult to model the exact behaviour of bitumen.

Detailed discussion on rheological characteristics of bitumen is beyond the scope of this book. A brief note on simple rheological models is given in Annexure I at the end of this chapter. Figure 10.15 represents the time versus stiffness modulus curve for (i) an elastic solid, (ii) a newtonian fluid, and (iii) a viscoelastic material, like bitumen.



(a) Elastic solid (b) Newtonian fluid (c) Viscoelastic material

Figure 10.15 Time versus stiffness modulus curve for an elastic solid, a newtonian fluid and a viscoelastic material.

10.4.4 Other Forms of Bitumen

Oxidized bitumen

The residual material obtained from distillation of petroleum is blown with hot air under

specific physical conditions—the derived material is the oxidized bitumen which is stiff and does not loose its stiffness due to further weather action. This air blown bitumen is, in general, considered unsuitable for road construction, but used as roof waterproofing material or as a joint filler in the concrete pavements.

Cutback bitumen

Cutback bitumen is produced by diluting penetration grade bitumen with suitable petroleum distillate (e.g. coal tar, naptha, creosote oil, or a mixture of these). After the application of cutback bitumen, the solvent present in cutback evaporates (or gets photo-oxidized) leaving behind bitumen to bind with aggregates. The cutback bitumen is categorized as (i) rapid curing (where naptha and gasolene are used as solvents), (ii) medium curing (kerosene as solvent), and (iii) slow curing (light oil as solvent) types. Cutback bitumen is designated in terms of its kinematic viscosity in centistokes [116]. Viscosity, flash point, percentage water content, specific gravity etc. are some of the tests carried out on cutback bitumen.

Bitumen emulsion

An emulsion is a two-phase system consisting of two immiscible liquids, one being dispersed as finite globules in the other. In bitumen emulsion, bitumen globules are suspended as emulsion in water with the help of emulsifiers, which are used to stabilize the bitumen emulsion. Emulsifiers break into ions and charge the bitumen particles. Charged bitumen particles repel each other and the suspension remains stable; this stability remains so long as water does not evaporate, freeze, or emulsifier does not break. Most of the bitumen emulsions are either of anionic or of cationic type in water media. Cationic emulsifiers are those which make bitumen particles positively charged, e.g. ammonium salt. Anionic emulsifiers make bitumen globules negatively charged, e.g. sodium stearate. Cationic emulsions treated with electro-negative aggregates such as gravel and siliceous types of aggregates are found to give good performance even under adverse moisture conditions [158]. Aggregates like limestone are electropositive and the anionic emulsion is the obvious choice in that case.

Depending on the setting speed of bitumen emulsion, they are subdivided into three groups, namely rapid setting (RS), medium setting (MS), and slow setting (SC) [117]. Some of the tests carried out on bitumen emulsion [117] are:

(i) Viscosity
(ii) Percentage water content
(iii) Specific gravity
(iv) Settlement time
(v) Sieve test
(vi) Miscibility with water

(vii) Particle charge
(viii) Storage stability
(ix) Stability to mixing with cement

Commercially, bitumen emulsion is manufactured by either of the two methods, namely (i) colloid mill and (ii) high-speed mixers. In the colloid mill method, there is a high-speed stator-rotor arrangement through which hot bitumen and water are passed using different pipes. The hydraulic shear force breaks bitumen into small globules of size of the order of 1–2 microns and later emulsifiers are added to stabilize the mix. In the high-speed mixer method, water is kept in a mixer just below the boiling point and hot bitumen is fed into the container slowly. The high-speed stirrer breaks bitumen into small globules to form bitumen emulsion. However, in this case dispersion may not be so uniform as in the case of the colloid mill.

Bituminous emulsion can be used even under adverse moisture conditions and has a wide range of applications, such as surfacing for low volume roads, curing purposes, bases for high volume roads, surface dressing, tack coat, premix carpets, soil stabilization, slurry seals, localized patch and pot-hole repair, etc. (consult Chapter 13 for discussion on the terms used here). Also, a brief discussion on emulsified bituminous mixes (EBMs) is contained in Section 13.10.1.

Closing remarks

The use of solvent in cutback bitumen, water in emulsion bitumen, and heating for penetration grade bitumen serves the common objective of making bitumen less viscous for mixing and laying purposes. The effect of heating in penetration grade bitumen is, however, shortlived, whereas in cutback and emulsion bitumen, the state of liquifaction achieved lasts longer. The cutback and emulsion forms of bitumen are specially recommended where there is (i) heavy rain which may cause hindrance to conventional hot mix construction, or (ii) where the environment is cold and humid making heating of bitumen difficult.

10.4.5 Tests on Bituminous Binder

This section discusses various standard tests for laboratory characterization of bitumen used as a binder. Broadly, these tests can be classified into:

(i) Viscosity
(ii) Ductility
(iii) Specific gravity
(iv) Durability
(v) Purity
(vi) Safety

Viscosity test

Viscosity is a property of fluids which opposes their flow. The higher the viscosity, the slower is the movement of the fluid. The viscosity can be expressed by the standard formula

$$F = \eta A \frac{dv}{dy} \tag{10.20}$$

where

F is the viscous force
η is the viscosity coefficient
dv/dy is the velocity gradient along the transverse direction of flow.

Viscosity of bitumen is an important parameter in bituminous construction. Different ranges of viscosities, depending on the type of binder, are recommended at various stages of construction, such as transportation, laying, mixing, and compaction. The viscosity of bitumen can be measured by (i) direct or by (ii) indirect methods.

In the direct method, the equipment used can estimate the viscosity of bitumen in terms of its absolute value (e.g. Stokes, Poise). Dynamic shear rheometer, Brookfield viscometer, Cannon-Manning vacuum viscometer, Asphalt Institute viscometer [224] etc. are some of the examples of such equipment. In a shear rheometer, a shear force is applied to a thin film of bitumen layer and viscosity is measured directly from the basic viscosity equation [Eq. (10.20)]. In Brookfield viscometer, bitumen film is placed between two concentric cylinders, one of which rotates at a known angular speed. By measuring the torque applied to the rotating cylinder, the viscosity of bitumen (in terms of its absolute value) can be found out. The Cannon-Manning vacuum viscometer and Asphalt Institute vacuum viscometer are based on the principle that the time required to flow through a given tube can be related to viscosity of the liquid [128].

In the indirect viscosity tests, the consistency of bitumen, rather than its viscosity, is measured which indirectly gives an idea of the viscosity range of the binder. Some indirect tests are:

(i) Standard tar viscometer
(ii) Penetration test
(iii) Softening point test
(iv) Float test

The indirect viscosity measurement equipment is popular because of its simplicity and low cost.

Standard tar viscometer test. Tar, emulsion, or cutback bitumen can be tested by the standard tar viscometer test [118]. In this test, bitumen/tar is put in a cup having a specified size orifice (10 mm standard, 4 mm for cutback bitumen). The cup is kept in a water bath to maintain a constant operating temperature. A stopwatch is started when

bitumen starts passing through the orifice and stopped after a definite quantity of bitumen has passed through the orfice and collected in a container kept at the bottom. The standard tar viscometer result is expressed in terms of time (in seconds) required for the flow of a specified volume (50 mm as per IS 1206) of bitumen at the specified test temperature.

The penetration test. "Penetration of a bituminous material is the distance in tenths of a millimetre, that a standard needle will penetrate vertically into a sample of bitumen under standard conditions of temperature, load, and time of loading'' [119]. Figure 10.16 shows a schematic diagram of the principle of penetration test. A highly polished needle of 1 mm diameter with total moving weight of 100 g, is allowed to penetrate through a bitumen sample put in a specified container as per the specified procedure, for a period of 5 seconds, and the penetration is measured [119]. The average of three such determinations is taken as the penetration value.

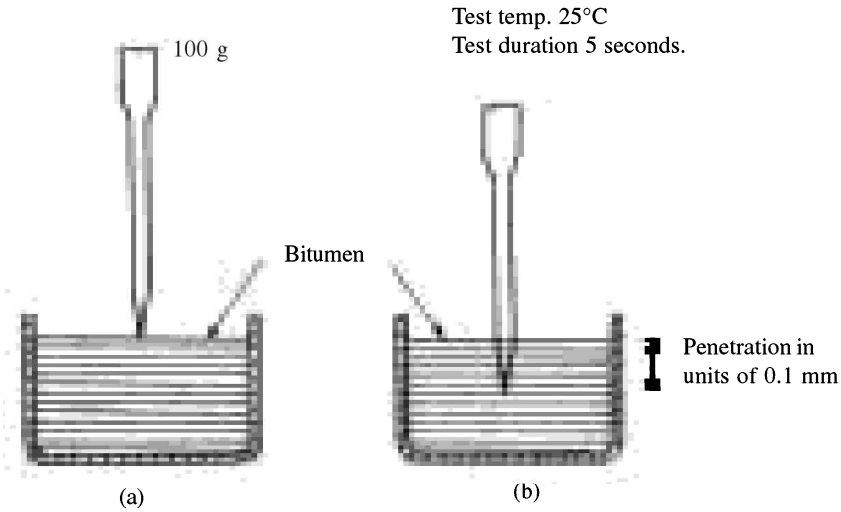


Figure 10.16 Schematic diagram of the penetration test set-up.

The standard temperature for the penetration test is 25°C. The average penetration value determines the grade of the bitumen. For example, for the 80/100 grade of bitumen, the penetration value is expected to lie between 80 to 100. The type of bitumen on which the penetration test can be performed is, therefore, called the *penetration grade* of bitumen. Different penetration grades of bitumen are used for different climatic conditions and design recommendations [88, 89]. In India, bitumen grades of 30/40, 60/70, and 80/100 are commonly used for highway construction. As is obvious, the 80/100 bitumen is suitable for cold climate locations whereas the 30/40 grade of bitumen is more suitable in hot climate regions. Figure 10.17 shows a photograph of the penetration test equipment.

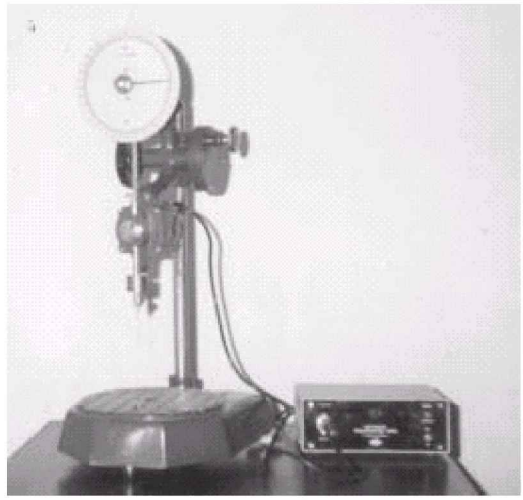


Figure 10.17 A laboratory penetration test set-up.

The following contains a discussion on empirical relationships between penetration of bitumen and various other parameters. Penetration and viscosity are related by the following empirical formula [102].

$$\eta = \frac{1.58 \times 10^{10}}{(\text{Pen}_T)^{2.16}} \tag{10.21}$$

where

Pen_T is the penetration value at any specified temperature T

η is viscosity of bitumen (in Poise).

Equation (10.21) is valid for penetration values greater than 60 [102]. Experiments have shown that penetration and time of penetration are proportional to each other. The penetration versus temperature follows the following relationship [102]:

$$\log_{10}(\text{Pen}_T) = (A \times T) + B \tag{10.22}$$

where T is the temperature, and A and B are two constants. The term A represents the slope of the 'penetration versus temperature' curve and is known as *temperature susceptibility*. Two different samples of bitumen (say, derived from two different refineries) having the same penetration grade may show different temperature susceptibilities, even though they may have the same viscosity at a particular temperature. In Figure 10.18, the viscosities of bitumen 'P' and bitumen 'Q' are the same at temperature T_1, but different at other temperatures, say T_2. This is of major importance in the selection of a performance-based binder, the criterion which is being increasingly emphasized today (see Annexure II of this chapter).

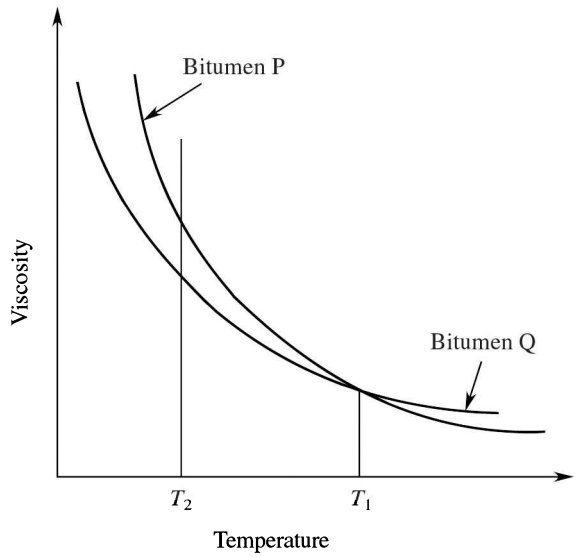


Figure 10.18 Schematic diagram of variation of viscosity with temperature.

Softening point. Softening point is a temperature at which a bituminous binder attains a particular degree of softening under the given physical conditions of the test [120]. Softening point of bitumen is determined by the *Ring & Ball test* as shown in Figure 10.19. Either water or glycerine, depending on the expected softening point of bitumen, is chosen as the medium

Figure 10.19 Softening point test.

(in the beaker). Glycerene is recommended when the expected softening point (also called the Ring & Ball temperature) is greater than 80°C. The liquid is heated at a constant rate of 5°C/minute by an external heating arrangement. Two steel balls are kept over the bitumen sample in brass rings, as shown in Figure 10.19. The average of the temperatures at which the balls touch the bottom metal plate, is the softening point of the bitumen sample. *It is found that for most of the types of bitumen, the penetration value at the softening point is* 800 [102].

EXAMPLE 10.3

The softening point value and penetration value of a sample of bitumen are 49°C and 85 respectively. What is its expected viscosity at 30°C?

Solution

Available data:

$$\text{Pen}_{25^\circ\text{C}} = 85$$

and

$$\text{Pen}_{49^\circ\text{C}} = 800 \text{ (as the penetration value at the softening point is 800.)}$$

Using Eq. (10.22) and solving for A and B, the values obtained are

$$A = 0.0405/^\circ\text{C} \quad \text{and} \quad B = 0.9151$$

Therefore, again using Eq. (10.22),

$$\text{Pen}_{30^\circ\text{C}} = 134.92$$

Now, from Eq. (10.21)

$$\eta_{30^\circ\text{C}} = 3.95 \times 10^5 \text{ Poise}$$

In Figure 10.20(a)–(d), the various relationships between temperature, penetration, and viscosity are presented schematically. A, B, and C are three different grades of bitumen. In Figure 10.20(a), it is seen that the slopes of the three lines are different as their temperature susceptibilities are different. They are straight lines as indicated by Eq. (10.22). Penetration at Ring & Ball (R&B) temperature being equal to 800, has the same value for all types of bitumen. Thus, in Figure 10.20(b), a line parallel to the x-axis is obtained. In Figure 10.20(c), the penetration value at 25°C of a particular bitumen is plotted against the R&B point temperature of that particular bitumen. A softer grade of bitumen will have a higher penetration value, and its softening point will be lower. Thus if the R&B temperatures are plotted against the logarithm of penetration value at 25°C, the points are observed to lie along a straight line for different samples of bitumen. Figure 10.20(d), where the logarithm of penetration is plotted against $T - T_{\text{R\&B}}$, represents the same trend as that in Figure 10.20(a). This figure shows that at a temperature equal to ring and ball temperature, the penetration of bitumens A, B, and C having

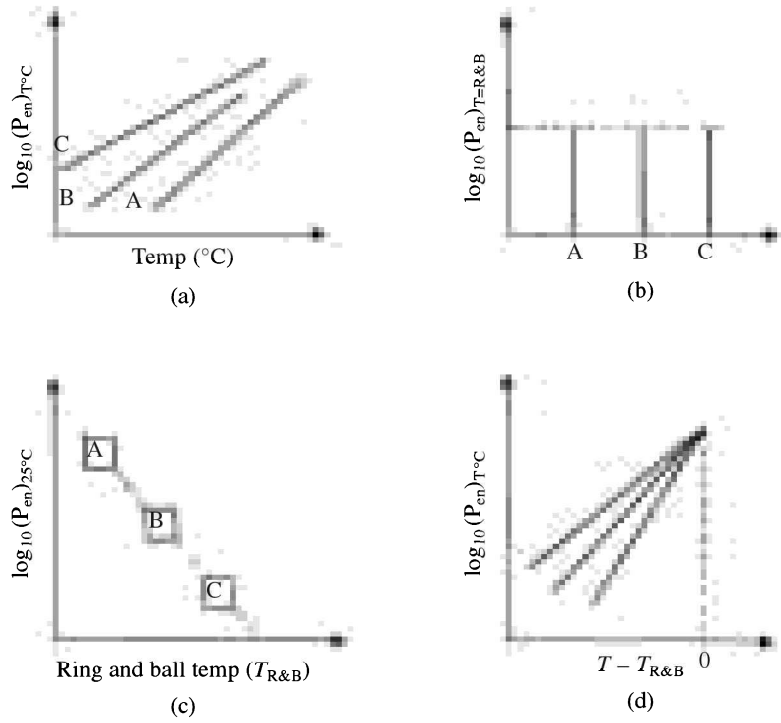


Figure 10.20 Schematic diagrams of variation between penetration, temperature and R&B temperature.

different temperature susceptibilities converges to a single point, which is 800.

Float test. Float test is generally recommended for that type of bitumen whose consistency is found unsuitable for penetration or softening point test. A brass collar filled with bitumen at a cooled temperature is fitted to an aluminium float. Time is noted when it is floated in a water bath maintained at a constant temperature (60°–65°C). The time is again noted when water starts entering through the bitumen plug which has attained molten state due to relatively higher temperature of water. This time is referred as the float test value [121]. Figure 10.21 illustrates the float test.

Ductility test

Ductility of a given grade of bitumen is obtained by measuring the distance in centimetres up to which the bitumen sample elongates before breaking [122]. The bitumen sample is put in a briquette of specific dimensions and is pulled apart at a specified rate (50 mm/min) maintaining a specified temperature of water bath (27°C), till failure. The average of three normal test readings is reported as the *ductility value* of the given bitumen sample [122]. Ductility value should be above 50 cm for satisfactory

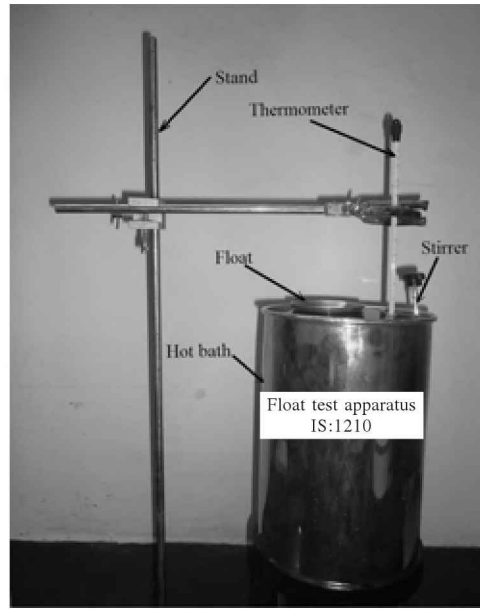


Figure 10.21 The Float test [121].

performance of bitumen. Bituminous mix made up of more ductile bitumen generally has a better longevity. The ductility value of bitumen reduces with its age.

Direct tensile tester is used to determine the ductility of bitumen at low temperatures [224], where bitumen generally exhibits brittle behaviour (the test temperature may vary from 0 to –36°C). In this test, the bitumen sample is pulled at a constant elongation rate. The load at which the stress reaches its maximum value is called failure load. Stress is calculated with respect to the original cross-sectional area of the sample. Failure strain is the change in length (up to failure) divided by the effective gauge length. As per Superpave recommendation (vide Annexure II of this chapter), 1% failure strain is acceptable for a sample of bitumen.

Specific gravity test

The specific gravity of bitumen/tar can be found out by the pychnometer method, as is done for small aggregates. For a sufficiently solid bitumen sample which can be handled in the form of pieces, its specific gravity can be determined by hanging the sample in a balance and weighing it in air and water [123]. Figure 10.22 shows the arrangement to determine the specific gravity of bitumen by the pychnometer method.

If

a is the mass of the dry pychnometer,
b is the mass of the pychnometer filled with water,
c is the mass of the pychnometer partially filled with bitumen,

Status of the specific gravity bottle	Empty	Completely filled with water	Partially filled with bitumen	Partially filled with bitumen, rest water	Completely filled with bitumen/tar
Weight	*a*	*b*	*c*	*d*	*e*

Figure 10.22 Determination of specific gravity of bitumen.

d is the mass of the pychnometer partially filled with bitumen and rest with water, and
e is the mass of the pychnometer completely filled with bitumen

then the specific gravity of bitumen is given by the two formulae:

$$\text{Specific Gravity} = \frac{c-a}{(b-a)-(d-c)} \tag{10.23}$$

$$\text{Specific Gravity} = \frac{e-a}{b-a} \tag{10.24}$$

The consistency of bitumen (i.e. whether pouring in and filling the complete pychnometer is convenient or not) determines whether the pychnometer is to be filled completely or partially with bitumen. The specific gravity of bitumen varies tentatively from 0.95 to 1.05.

Durability test

Durability is the ability of a substance to maintain its property when subjected to air, temperature, water, and other environmental factors. Bitumen, due to weathering, becomes hard and loses its desirable properties. Hardening of bitumen occurs due to three main reasons, namely oxidation, volatalization, and cold temperature. Bitumen gets oxidized gradually when exposed to the environment and therefore hardens. The oxidation of bitumen is a slow process. Oxidative hardening is accelerated if the temperature is high and the bitumen is thin, as happens during hot mixing itself. Volatalization is the evaporation of relatively volatile matter when temperature is high. Cold temperature hardening is a temporary process and bitumen regains its property after it is brought back to normal temperature. Durability tests focus on bitumen aging which causes permanent deterioration of its desirable properties. Thin film oven test, rolling thin film oven test, and the pressurized aging vessel test are the tests used to

measure the durability of bitumen. These tests measure both the aging of bitumen due to oxidation and the consequent loss of physical properties and the loss of volatile matter. An appreciable loss in volatile matter during test indicates (which is measured in terms of the loss of weight) that aging may occur during mixing and construction operations. Interestingly, bitumen due to its oxidation, gains weight after the test, the oxidized bitumen being heavier than the normal bitumen.Thus, due to aging weight of bitumen sample may increase or may decrease.

In the thin film oven test (TFO), 50 ml of the bitumen sample is put in a cylindrical flat-bottom pan of internal diameter 140 mm and depth 9.5 mm. The thickness of the bitumen sample is approximately 3 mm. The pan is placed on a rotating shelf in an oven where it is kept at a temperature of 163°C for 5 hours. The pan rotates with an approximate speed of 5 to 6 revolutions per minute. After the test, the loss in weight is measured [124] and the sample is tested for penetration and softening point [128].

In rolling thin film oven (RTFO) test [224], a specially designed bottle is used which is continuously exposed to a fresh film of bitumen. The bottle kept on a rack, rotating at a rate of 15 revolutions per minute, is periodically exposed to heated air jet, with air flow rate as 400 ml/min. The RTFO can accommodate more samples than TFO and requires less time to achieve the same level of aging.

In the pressurized aging vessel test, bitumen is exposed to high pressure and temperature simultaneously, to accelerate the aging procedure. Bitumen samples are kept in a sample rack inside the pressurized vessel and then heated. After the temperature equilibrium is achieved, pressure (2070 kPa) is gradually applied. The samples are kept confined for 20 hours, at specified pressure and temperature (it could be 90°, 100°, or 110°C, depending upon the requirement). When the test time is over, pressure is gradually reduced (gradual release of pressure is necessary to avoid foaming), sample rack is taken out and kept in a hot chamber (163°C) to remove air bubbles, and then referred to further testing. Pressurized air vessel is being used for research on bitumen aging, and different practices follow different procedures and specifications. The specifications mentioned here are as per the Superpave recommendations [224].

Purity test

Three purity tests on bitumen, namely solubility test, spot test, and water content test are discussed in this subsection.

Solubility test. Bitumen is mostly soluble in benzene, toluene, and trichloroethylene. Some impurities present in bitumen such as free carbon, salt, and others which are not soluble in these solvents, can be filtered out. The insoluble part of bitumen is measured and expressed as a percentage by mass of the bitumen sample taken [125].

Spot test. About 2 g of bitumen is dissolved in 10 ml of naptha and the two spots are put on a filter paper, one at 1 hour and then the other at 24 hours after the solution is prepared. If the strains of the spots are uniform in colour, then the purity of bitumen is

confirmed (AASHTO T 102-74). The spot test measures the presence of cracked bitumen. If cracked bitumen is present, the spot towards its centre appears deep in colour than that along the periphery.

Water content test. The standard method of water content test is known as *Dean and Stark method* and is conducted in Dean and Stark assembly [126]. Bitumen, kept in a 500 ml heat resistant glass container is heated just above the boiling point of water. The evaporated water is condensed and collected. The quantity of water present in bitumen is expressed by mass as percentage of the sample.

Safety

Flash point and fire point tests deal with the safety considerations while heating bitumen in the field for bituminous construction. These tests are carried out in Cleveland or in Pensky-Martin [127] flash point tester. In these tests, a brass cup filled with bitumen is heated (and stirred) at a prescribed rate. Periodically, a small flame is put over the sample surface. The temperature at which sufficient vapours are produced so that it catches fire in the form of an instantaneous flash is called the *flash point*. The *fire point* is the lowest temperature at which the application of test flame causes the material to ignite and burn for at least 5 seconds [127]. Flash point is lower than the fire point.

Closing remarks

Some selected major tests on bituminous binder have been discussed in this subsection. A large number of other tests such as distillation test, ash content test, naphthalene and phenol content test, volatile matter test, equiviscous temperature test, etc. are recommended depending on the specific application.

10.5 BITUMINOUS MIXES

The preceding sections discussed aggregates and bitumen as individual ingredients which when mixed in definite proportions constitute a bituminous mixture. A bituminous mixture (or mix) is prepared by mixing a chosen gradation of aggregates with a requisite quantity of bitumen of a given grade. Various aggregate gradations have been recommended in Indian specifications. These gradations are chosen depending mainly upon the design recommendations, type of road, earlier experiences, funds, local material availability, and so on. The Optimum Bitumen Content (OBC) of a mix is obtained from the stability tests where the strength and other properties of bituminous mix are determined by varying the quantity of bitumen. Some of the stability tests are Marshall test, Habbard stabilometer, Hveem stabilometer, etc. Marshall test is mostly used (in India as well) for the determination of OBC because of its simplicity and low equipment cost. Bituminous mixes are covered below under three subsections—mix volumetrics, mix design, and dynamic modulus and fatigue testing of bituminous mixes.

10.5.1 Mix Volumetrics

Mix volumetrics describes the relative volume proportions among the various constituents of the bituminous mix.

Figure 10.23 shows a phase diagram of a single aggregate. Some amount of bitumen goes inside the aggregate pores, called the absorbed bitumen, and this bitumen remains ineffective as far as the mix behaviour is concerned. Still, there can be some other (or part of) air voids within the aggregate which remain inaccessible by bitumen. These are unoccupied air voids within the aggregate.

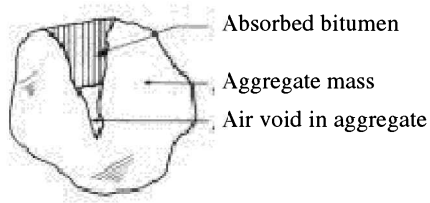


Figure 10.23 Schematic diagram of an individual aggregate showing air void and absorbed bitumen.

Based on the volume considered in evaluating the specific gravity of an aggregate, three definitions of specific gravity are proposed. The definitions are as follows:

Bulk specific gravity of the aggregates (G_{sb})

$$G_{sb} = \frac{M_{agg}}{\text{volume of (aggregate mass + air void in aggregate + absorbed bitumen)}} \quad (10.25)$$

where M_{agg} is the mass of the aggregate.

Effective specific gravity of aggregates (G_{se})

$$G_{se} = \frac{M_{agg}}{\text{volume of (aggregate mass + air void in aggregate)}} \quad (10.26)$$

Apparent specific gravity of aggregates (G_a)

$$G_a = \frac{M_{agg}}{\text{volume of aggregate mass}} \quad (10.27)$$

The reader may refer to Figure 10.10 and the definitions of bulk gravity and apparent specific gravity discussed there. It may be noted that these two basic definitions remain unchanged. Only due to the presence of bitumen as another volumetric component, the definition of effective specific gravity has been introduced here.

Figure 10.24 presents a schematic diagram of various volume components of the bituminous mix as a whole. The figure shows the total aggregate volume as V_{agg}, the amount of air void present as VA. The total volume of voids in the aggregate mix (when

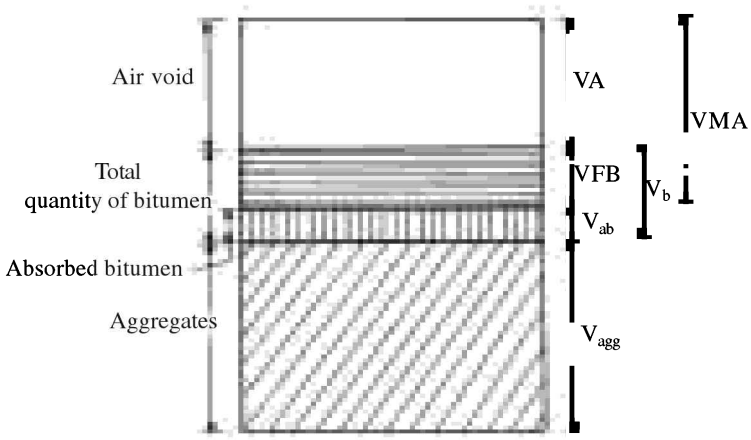


Figure 10.24 Schematic phase diagram of mix volumetrics.

there is no bitumen) is called Voids in Mineral Aggregates (VMA), out of which a part is occupied by usable bitumen, which is called Voids Filled with Bitumen (VFB). The value of VFB is expressed as a fraction of VMA.

Figure 10.25 shows a schematic cross-section of a bituminous mix sample. As shown in the figure, there are two types of air voids which may remain in the mix, namely the intra-aggregate air void and the inter-aggregate air void. It may be noted that while considering the parameter Air Voids (VA), only the inter-aggregate air voids are taken into account.

The reader may note that Figure 10.25 is just a schematic representation. A picture

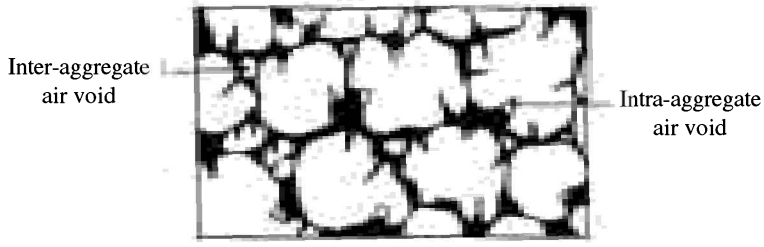


Figure 10.25 Air voids in a bituminous mix.

obtained from image analysis of bituminous mix is shown in Figure 10.26 for better understanding. With this background, the various terms related to the volumetrics of the mix can now be defined.

Theoretical maximum specific gravity of the mix (G_{mm}) is defined as

$$G_{mm} = \frac{M_{mix}}{\text{volume of the (mix-air voids)}} \tag{10.28}$$

where M_{mix} is the mass of the bituminous mix.

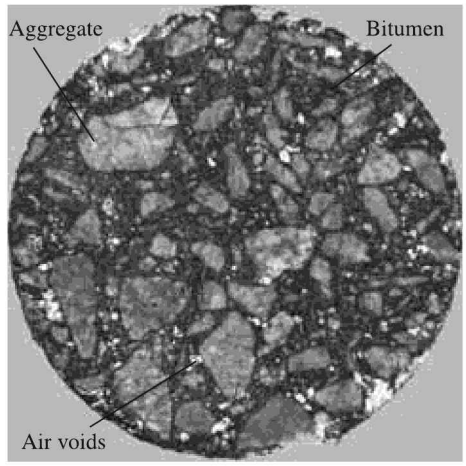


Figure 10.26 Cross-section of bituminous mix obtained from image analysis.

G_{mm} is measured by vacuum test, where aggregates are separated from the mix and put in a water jar. Vacuum is applied gradually, so that water can occupy all the intergranular voids. The volume of the mix without air voids is then measured, and the G_{mm} value is calculated (ASTM D 2041-95).

Bulk specific gravity of the mix (G_{mb}) is defined as

$$G_{mb} = \frac{M_{mix}}{\text{bulk volume of the mix}} \tag{10.29}$$

Bulk specific gravity is obtained by measuring the total mass of the mix and its volume. Volume is determined by measuring dimensions if the sample has a regular shape (like, Marshall sample), or for better accuracy it can be measured by the volume of water it displaces. However, while the sample is immersed in water, some water may be absorbed by the pores of the mix. Therefore, the mix is covered with a thin film of paraffin and the volume of the sample is measured by knowing the volume of paraffin used to coat its surface.

The effective specific gravity of the aggregates in a mix (G_{se}) is calculated as

$$G_{se} = (M_{mix} - M_b) \div \left(\frac{M_{mix}}{G_{mm}} - \frac{M_b}{G_b} \right) \tag{10.30}$$

where

M_b is the mass of bitumen used in the mix

G_b is the specific gravity of bitumen.

Voids in Mineral Aggregates (VMA) can be calculated as

$$VMA = \left[\left(\frac{M_{mix}}{G_{mb}} - \frac{M_{mix} P_s}{G_{sb}} \right) \div \frac{M_{mix}}{G_{mb}} \right] \times 100 \tag{10.31}$$

where P_s is the fraction of aggregates present, by total mass of the mix (that is, $M_{agg} = P_s \times M_{mix}$). Finally,

$$\text{VMA} = \left(1 - \frac{G_{mb}}{G_{sb}} \times P_s\right) \times 100 \tag{10.32}$$

Air Voids (VA) can be obtained as

$$\text{VA} = \left[\left(\frac{M_{mix}}{G_{mb}} - \frac{M_{mix}}{G_{mm}}\right) \div \frac{M_{mix}}{G_{mb}}\right] \times 100 \tag{10.33}$$

or

$$\text{VA} = \left(1 - \frac{G_{mb}}{G_{mm}}\right) \times 100 \tag{10.34}$$

Once the VMA and VA are known, Voids Filled with Bitumen, VFB, can be calculated easily using the following equation:

$$\text{VFB} = \frac{\text{VMA} - \text{VA}}{\text{VMA}} \times 100 \tag{10.35}$$

The percentage absorbed bitumen P_{ab} is given by

$$P_{ab} = \left[\left(\frac{M_{agg}}{G_{sb}} - \frac{M_{agg}}{G_{se}}\right) \div \frac{M_{agg}}{G_{sb}}\right] \times 100 \tag{10.36}$$

or

$$P_{ab} = \left(1 - \frac{G_{sb}}{G_{se}}\right) \times 100 \tag{10.37}$$

This percentage absorbed bitumen P_{ab} in Eq. (10.37) has been expressed in terms of the total mass of the aggregates. It can also be expressed as the percentage absorbed bitumen of the total weight of the mix.

Significance of volumetric parameters

Bitumen holds the aggregates in position and the load is taken by the aggregate mass through the contact points. If all the voids are filled by bitumen, the load is transmitted by hydrostatic pressure through bitumen, and the strength of the mix, therefore, reduces. That is why the stability of the mix starts reducing when bitumen content is increased beyond a certain value. Also during the summer season, bitumen melts and occupies the void space between the aggregates and if the void space is not available, it causes bleeding. Thus, some amount of void is necessary in a bituminous mix, even after the final stage of compaction. Therefore, minimization of air voids cannot be the governing

criterion for bituminous mix design (as also discussed in Section 10.3.4). However, excess void makes the mix weak from the point of view of its elastic modulus and fatigue life. The chances of oxidative hardening of bitumen are more for a mix with more voids.

The mix specification puts various limits to the volumetric parameters to be accepted for different types of mixes such as BC, DBM, SDBC, etc. Table 10.5 gives the minimum VMA requirement set in the MORT&H specification [215].

Table 10.5 Minimum percentage void in mineral aggregates (as per Table 500-12 of MORT&H Specification [215]).

Nominal maximum particle size (mm)	*Minimum VMA* (%) *related to design air voids* (%)		
	3.0	4.0	5.0
12.5	13.0	14.0	15.0
19.0	12.0	13.0	14.0
25.0	11.0	12.0	13.0
37.5	10.0	11.0	12.0

Note: Nominal particle size is one size larger than the first sieve to retain more than 10%.

EXAMPLE 10.4

The bulk specific gravities of coarse aggregate, fine aggregate, and fines are found to be 2.5, 2.60, and 2.65 respectively. The specific gravity of bitumen is given as 1.10. For preparation of the Marshall sample, their relative proportions used are 55, 30, 10, and 5% by weight respectively. The mass of the sample measured in air is 1245.2 g, the mass of paraffin coated sample in air is 1295.8 g, and the mass of the paraffin coated sample in water is measured as 703.7 g. Calculate:

(i) The bulk specific gravity of the mix
(ii) VMA
(iii) VA
(iv) VFB
(v) The percentage absorbed bitumen

The specific gravity of paraffin is 0.9 and the theoretical maximum specific gravity of the mix, obtained from the vacuum test is 2.441.

Solution

The volume of the paraffin coated sample can be obtained as the difference of the weight loss while immersed in water and the volume of the paraffin used. Thus,

Bulk volume of the sample = $(1295.8 - 703.7) - \dfrac{(1295.8 - 1245.2)}{0.9} = 535.877$ cc

(i) From the definition of G_{mb} [Eq. (10.29)]

$$G_{mb} = \frac{1245.2}{535.877} = 2.324$$

G_{sb} can be calculated from the individual specific gravities of the aggregates. Thus,

$$G_{sb} = 95 \div \left(\frac{55}{2.50} + \frac{30}{2.60} + \frac{10}{2.65} \right) = 2.546$$

(ii) From Eq. (10.32)

$$\text{VMA} = \left(1 - \frac{2.324}{2.546} \times 0.95 \right) 100 = 13.28\%$$

(iii) From Eq. (10.34)

$$\text{VA} = \left(1 - \frac{2.324}{2.441} \right) 100 = 4.79\%$$

(iv) From Eq. (10.35)

$$\text{VFB} = \frac{13.28 - 4.79}{13.28} \times 100 = 63.93\%$$

(v) From Eq. (10.30)

$$G_{se} = (100 - 5) \div \left(\frac{100}{2.441} - \frac{5}{1.10} \right) = 2.608$$

From Eq. (10.37)

$$P_{ab} = \left(1 - \frac{2.546}{2.608} \right) 100 = 2.377\%$$

of the mass of the aggregates. Therefore, the percentage absorbed bitumen is 2.5% of the total weight of the mix.

10.5.2 Mix Design

Selection of Optimum Bitumen Content (OBC) is a delicate balancing act in which there are a number of variables. A balance is to be maintained such that all the specification limits recommended in the code of practice are simultaneously satisfied. There are two aspects in a mix design specification which are described here.

(a) Specifications limit the volumetric parameters, like, upper and lower limits of VMA, VA, and VFB. Figure 10.27 presents a conceptual probability diagram,

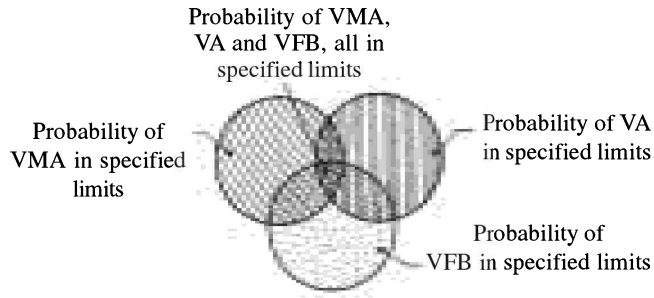


Figure 10.27 Probability diagram for mix volumetrics.

where the individual probabilities of VMA, VA, VFB being within their respective specified limit are plotted [176, 69]. One of the objectives of the mix design is to find a suitable bitumen content for which all the volumetric specifications are simultaneously satisfied.

(b) The stability parameters are another consideration in mix design which are obtained from the stability tests. Flow and stability are the two parameters measured by the Marshall test while the stabilometer and cohesiometer values are measured by the Hveem method. Suitable bitumen content is chosen from the test results where these parameters are within the specified limits.

In this subsection, mix design by the Marshall method and the Hveem method has been briefly discussed.

Marshall method

Marshall test was developed by Bruce Marshall of the Mississippi State Highway Department in 1940, and is still being used as a popular test method for design of bituminous mixes. The U.S. Army Corps of Engineers formalized the procedure as ASTM D 1559 and AASHTO T 245 after incorporating some modifications to the originally proposed procedure [223]. In the Marshall test, load is applied to a cylindrical specimen (refer Figure 10.28) of bituminous mix and the sample is monitored till its failure.

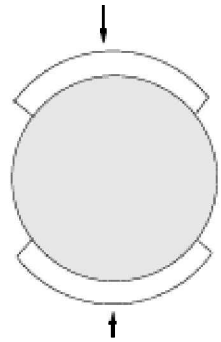


Figure 10.28 Schematic representation of the Marshall test.

For preparation of the Marshall specimen, about 1200 g aggregates are weighed and heated up to 154–160°C. Bitumen is separately heated up to 175–190°C and the measured quantity of bitumen is poured in the container where the aggregates are being heated. Aggregates and bitumen are thoroughly mixed in such way that the upper surfaces of the aggregates appear to be uniformly coloured with bitumen film. The mix is poured into the Marshall mould (64 mm height and 100 mm diameter) and compacted with 75 blows on each face. The mould is taken out and kept under normal laboratory temperature for 12 hours. It is immersed in a water bath kept at constant temperature of 60°C for 30 minutes and after that it is taken out for testing in the Marshall testing machine. Load is applied vertically at the rate of 50 mm per minute on the sample at 60°C and the maximum load at which the sample fails is recorded as the Marshall stability value. Figure 10.29 shows a photograph of a Marshall test equipment installed in the laboratory.

For a given aggregate gradation, the stability value initially increases with the increase

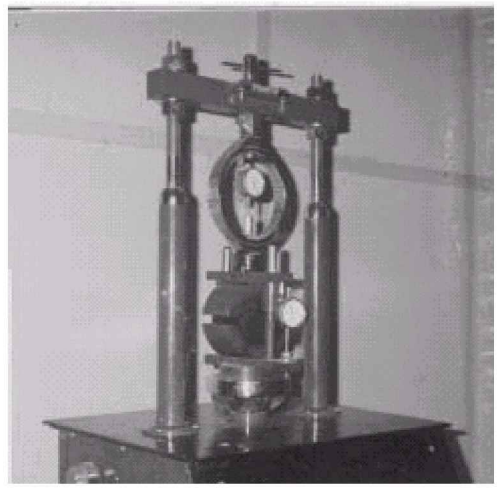


Figure 10.29 A photograph of a Marshall test equipment.

in bitumen content as the aggregate–bitumen bond gradually gets stronger. But with further increase in the bitumen content, the applied load is transmitted as hydrostatic pressure, keeping the friction across the contact points of aggregates immobilized. This makes the mix weak against plastic deformation and the stability falls. Figure 10.30 is a schematic representation of this trend.

Flow, in the Marshall test, is the deformation undergone by the specimen at the maximum load where failure occurred. The flow value increases with the increase in bitumen content. The increase is slow initially, but later the rate increases with the increase in bitumen content, as shown in Figure 10.31.

The VMA value, for a given aggregate gradation should theoretically remain constant.

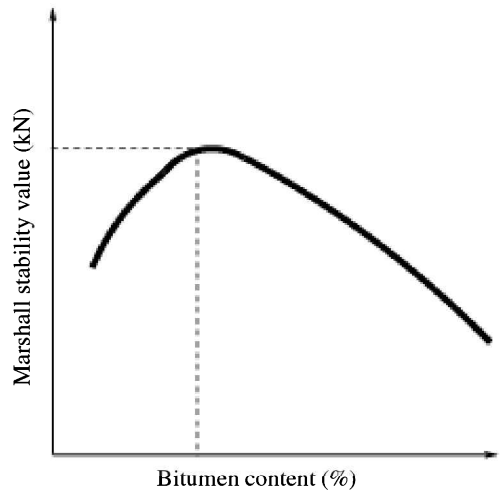


Figure 10.30 Schematic diagram of variation of Marshall stability value with bitumen content.

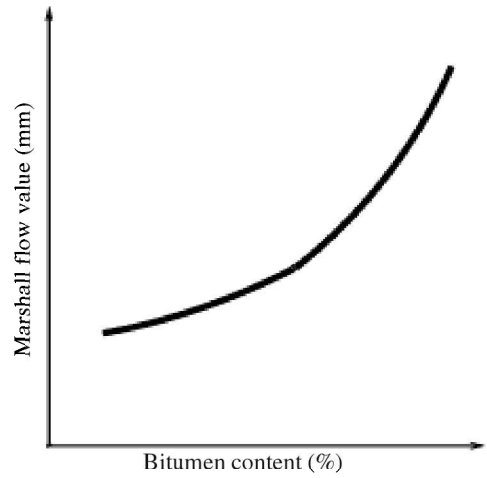


Figure 10.31 Schematic diagram of variation of Marshall flow value with bitumen content.

However, in this case, it is sometimes observed that, at low bitumen content, VMA slowly decreases with the increase in bitumen content, then remains constant over a range, and finally increases at a high bitumen content (see Figure 10.32). The initial fall in the VMA value is due to the re-orientation of the aggregates in the presence of bitumen. At a very high bitumen content, due to a thicker bitumen film, the aggregates move apart slightly, resulting in an increase in VMA.

With the increase in bitumen content, VA of Marshall sample decreases, as bitumen

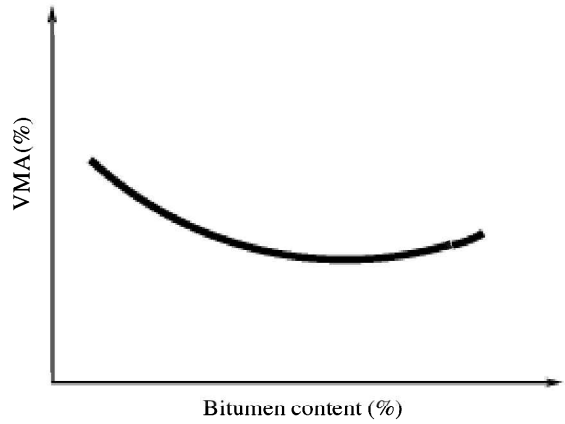
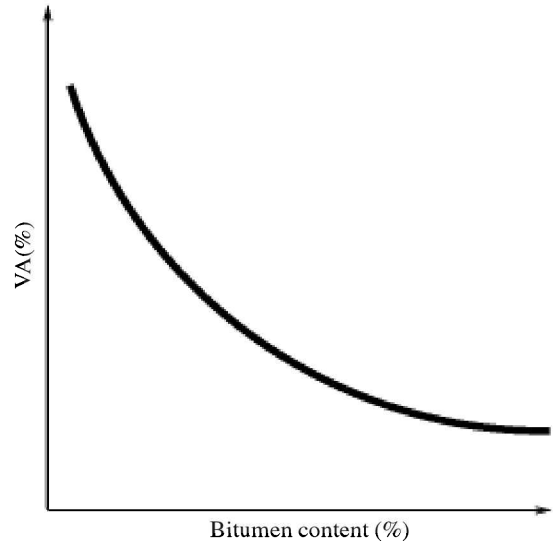


Figure. 10.32 Schematic diagram of variation of Marshall VMA value with bitumen content.



Figuree 10.33 Schematic diagram of variation of VA value with bitumen content.

replaces the air voids present in the mix (see Figure 10.33) and subsequently, VFB increases (see Figure 10.34) with the increase in bitumen content.

EXAMPLE 10.5

The following proportions of aggregates are used for the preparation of Bituminous Concrete (BC) as per the 1997 specification [217] .

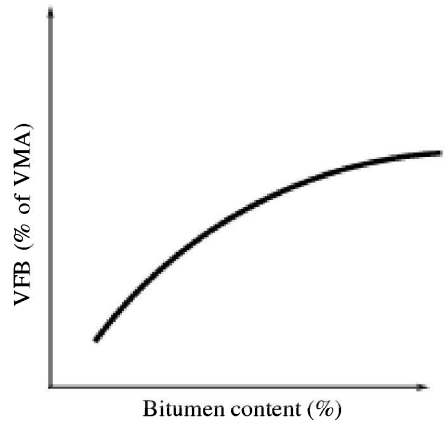


Figure 10.34 Schematic diagram of variation of VFB value with bitumen content.

Aggregate size (mm)		*Specific*	*Weight*
Passing	*Retained*	*gravity*	*fraction* (%)
26.5	19.0	2.60	5
19	9.5	2.60	27
9.5	4.75	2.55	18
4.75	2.36	2.65	14
2.36	0.30	2.55	24
0.3	0.075	2.36	7
0.075	—	2.42	5

Marshall test is conducted for different bitumen contents and the following values are obtained.

Bitumen content (%)	*Wt. of sample in air* (g)	*Wt. of sample in water* (g)	*Stability* (kN)	*Flow* (mm)
4.00	1380	755	09.5	1.5
4.25	1315	735	10.3	2.0
4.50	1315	740	13.5	2.4
4.75	1350	755	17.0	3.4
5.00	1355	765	19.0	3.6
5.25	1370	755	15.0	4.8
5.50	1380	785	12.3	5.6

Note:

1. All the values stated above are the mean of three samples.
2. Specific gravity of bitumen is 1.052.
3. Bitumen content is expressed as percentage of the total aggregate weights.

Mix design recommendations for BC as per 1997 specification [217] are as follows:

Minimum Marshall stability	8.20 kN
Marshall flow	2 to 4 mm
Percentage air voids in mix	3 to 5%
Minimum VMA per cent	11%
VFB	65 to 75%
Minimum binder content by weight of total mix	4.5%

Find the optimum bitumen content of the mix.

Solution

The bulk specific gravity G_{sb} is calculated as

$$\frac{100}{\frac{5}{2.60}+\frac{27}{2.60}+\frac{18}{2.55}+\frac{14}{2.65}+\frac{24}{2.55}+\frac{7}{2.36}+\frac{5}{2.42}} = 2.558$$

The parameters obtained from Marshall test are tabulated in the following table. The plots of variation of parameters with bitumen content are also shown in Figure 10.35. It may be noted that except for the variation shown by VMA (see Figure 10.32), all other trends match with the expected trend.

Bitumen (%)	G_{mm}	G_{mb}	VA (%)	VMA (%)	VFB (%)	*Stability* (kN)	*Flow* (mm)
4.00	2.425	2.21	8.93	17.0	47.48	9.5	1.5
4.25	2.417	2.27	6.19	15.0	58.65	10.3	2.0
4.50	2.409	2.29	5.08	14.4	64.80	13.5	2.4
4.75	2.402	2.27	5.54	15.3	63.82	17.0	3.4
5.00	2.395	2.30	4.10	14.5	71.72	19.0	3.6
5.25	2.388	2.30	3.56	14.5	75.41	15.0	4.8
5.50	2.380	2.32	2.56	14.1	81.76	12.3	5.6

A sample calculation for the table above for 4% bitumen is as follows:

As maximum theoretical specific gravity is not given, it can be approximately calculated as

$$\frac{100}{\frac{3.85}{1.0052}+\frac{96.15}{2.558}} = 2.425$$

$$\text{Bulk specific gravity} = \frac{1380}{1380-755} = 2.208$$

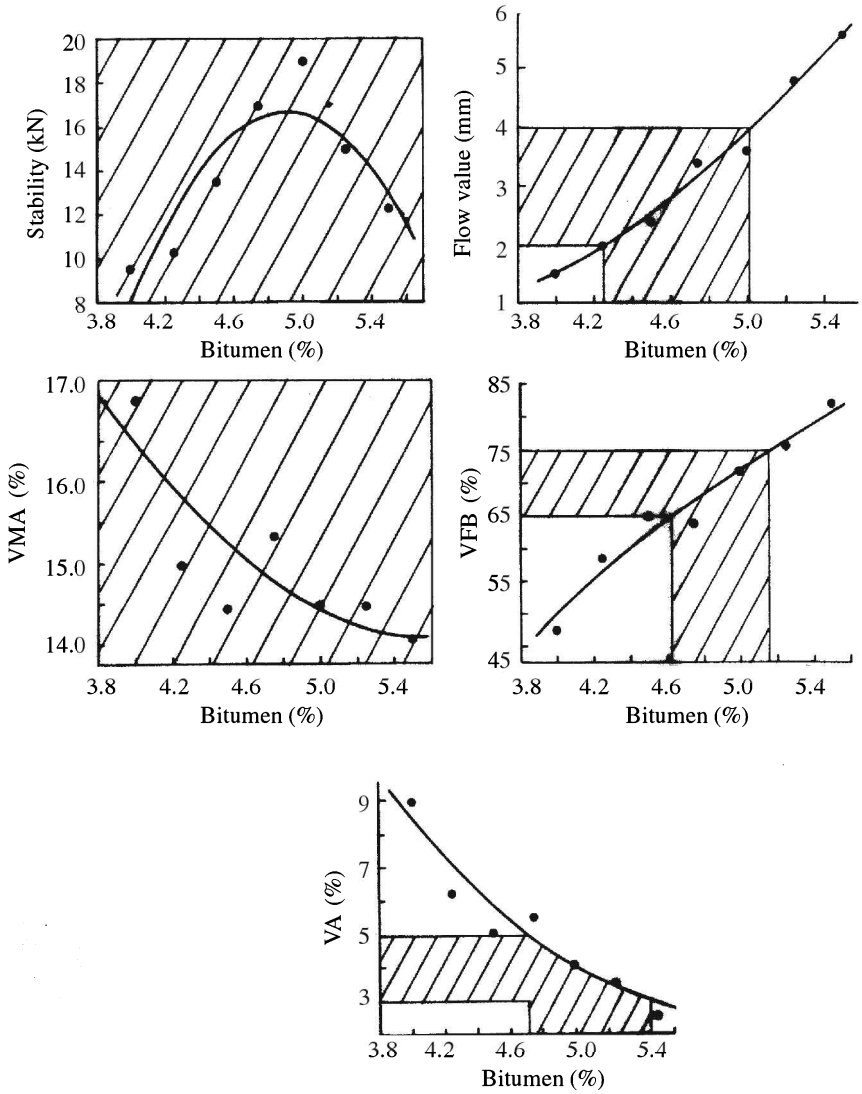


Figure 10.35 The plots of the Marshall test results.

$$\text{VA} = \frac{2.425 - 2.208}{2.425} \times 100 = 8.93\%$$

$$\text{VMA} = \left(100 - \frac{2.208 \times 96.15}{2.558}\right) = 17.0\%$$

$$\text{VFB} = \left(\frac{17 - 8.93}{17} \times 100 \right) = 47.48\%$$

The variations of the Marshall parameters are plotted in Figure 10.36 and the acceptable ranges, as provided by the recommendations in this Example, are marked with a different shade. The minimum bitumen content, which satisfies all the specifications, is (see Figure 10.36) 4.70%. It may be chosen [163] as the optimum bitumen content, being the most economical one. However, depending on the specific requirement of a project some other bitumen content may as well be chosen.

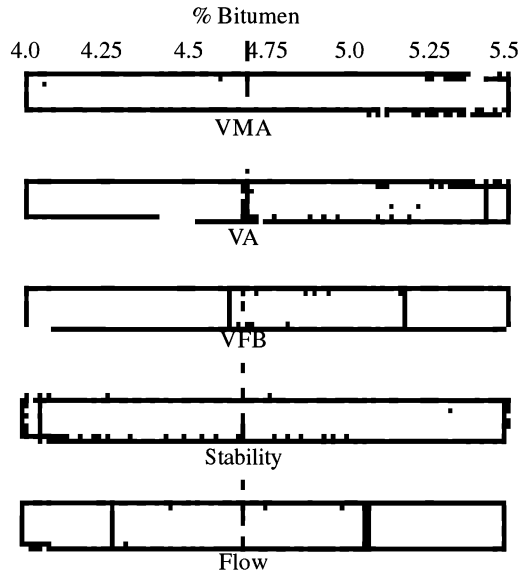


Figure 10.36 Determination of optimum bitumen content from Marshall test.

Mix specification. Marshall test is used for the bituminous mix design as per Indian recommendation. The various mix specifications are available in the MORT&H specifications for Road and Bridge Works [216, 217, 215] and other in IRC specifications.

Two things are of primary concern in a bituminous mix, namely the aggregate gradation and the mix design requirements. Various mixes have various gradations as discussed in Section 10.3.4. The acceptable volumetric parameters and Marshall stability requirements are different for different mixes (see Tables 10.5 and 10.6). Thus for various individual mixes a separate Marshall mix design needs to be carried out to find out the OBC value. The characteristics of locally available aggregates vary from region to region and so also of the bitumen derived from different refineries. Thus, the OBC value determined from Marshall test of the same mix specification (same aggregate gradation and same grade of bitumen), may even show different values.

Table 10.6 Mix design requirements of some bituminous mixes as per MORTH&H specification (fourth revised edition) [215].

	BC	SDBC	DBM
Marshall stability (kN at 60°C)	9.0	> 8.2	> 9.0
Marshall flow (mm)	2–4	2–4	2–4
Air void (%)	3–6	3–5	3–6
VMA (%)	as per Table 10.5		
VFB (%)	65–75	65–78	65–75
Binder content (%) as mass of total mix	5–6%	4.5%	4% minimum

Closing remarks. The Marshall test parameters are not directly related to the fatigue and rutting failure of the pavement. Also, the impact compaction used in this method does not simulate the nature of densification as it occurs in the field. Thus, this method, as such, may not truely represent the performance of the bituminous mix [180]. Still, the Marshall method of mix design is most popular because of its simplicity, low cost, and also because mix design by Marshall test, so far, has given satisfactory performance of the in-service roads.[5]

Hveem stability method

Hveem mix design procedure was developed by Francis Hveem of the California department of transportation. Hveem method has certain advantages over the Marshall method such as the following: (i) In the Hveem method, kneading compaction is adopted which is close to the filed compaction and (ii) Hveem stability gives an idea of the shear resistivity of the sample [223]. The Hveem method of testing (ASTM D1560 92) of bituminous mixtures involves three tests, namely the centrifuge kerosene equivalent (CKE) test, the stabilometer test, and the cohesiometer test.

In the kerosene equivalent test, the aggregates smaller than 4.75 mm are saturated with kerosene and centrifuged. Coarser aggregates (size between 9.5 mm and 4.75 mm) are saturated with a lubricating oil and allowed to drain out for 15 minutes at 60°C. The weights of kerosene and oil retained by the aggregates give an idea of the specific surface of the aggregate mass, therefrom, the tentative OBC value can be inferred.

[5]As discussed, there are a number of limitations posed by the Marshall method of mix design, though it is still the most popular mix design technique in India and other parts of the world. In this context, Strategic Highway Research Program (SHRP), USA, initiated a $150 million project in 1987, to evolve the Superior Performing Asphalt Pavement (Superpave) [224, 223]. A brief discussion on the concepts of Superpave can be found at the end of this chapter in Annexure II.

For the study of Hveem stability, bituminous mix is prepared at tentative OBC (as obtained by CKE) and at bitumen contents higher and lower than the tentative OBC. The bituminous mixture is compacted in a kneading compactor according to the specified procedure and the sample is heated to 60°C and tested in the Hveem stabilometer. In the stabilometer, the sample is kept enclosed in a rubber specimen which is surrounded by a liquid. Vertical load is gradually applied and the lateral pressure in the liquid is measured, as in a triaxial testing machine. The test result is expressed in the scale of 0 to 90 range. A value '0' of Hveem stability means that the sample is liquid (i.e. lateral pressure is equal to vertical pressure) and the value '90' means that the specimen is an incompressible solid. The Hveem stability test with some modifications is also used for finding the stability of other pavement materials, like soil or aggregates, and is used in pavement design (see Section 12.3.2). Figure 10.37 shows a schematic representation of the Hveem stabilometer test.

In the cohesiometer test, the compacted sample after the stability test, is used in the

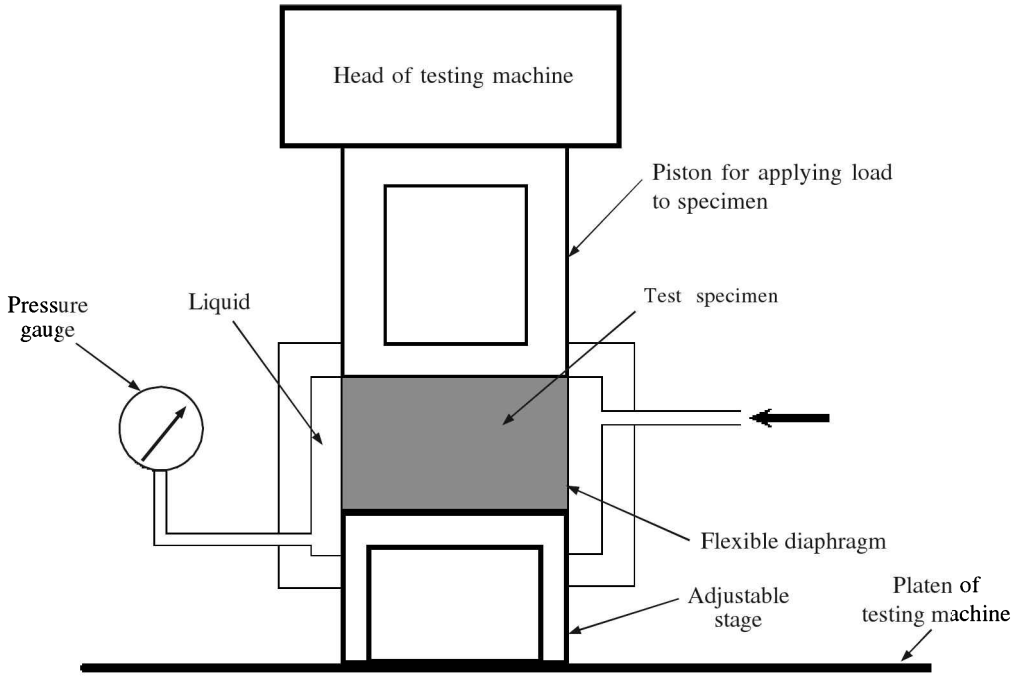


Figure 10.37 Schematic diagram of Hveem stabilometer test [128].

cohesiometer. Samples taken from in-service road can also be tested by cohesiometer. In this test, two parallel plates hold the sample from two sides. Vertical downward force is gradually applied from one side of the sample, and at some stage it fails. The failure load is used to calculate the Hveem cohesiometer value.

10.5.3 Stiffness Modulus and Fatigue Performance of Bituminous Mixes

Fatigue test on bituminous mix

Fatigue is a phenomenon of fracture under repeated cyclic or fluctuating load. The *fatigue life* (of a bituminous mix sample) is defined as the number of repetitions for which the initial stress or strain changes by an arbitrary factor. Figure 10.38 shows a schematic diagram of fatigue testing of bituminous mix. A rectangular sample of bituminous mix is placed on simple supports and a repetitive load is applied to the sample, and the sample is monitored till its failure. The supports are placed in both the directions because the cyclic (sinusoidal) loading pushes the beam for the duration of half-cycle, and then pulls it for the rest half. Figure 10.38 also shows two repetitive loads (separated by 1/3rd of the length of beam L) of equal magnitude P applied to the sample. Two loads are generally used because the bending moment in the middle third of the beam is constant and the shear force is zero; therefore, the failure of the beam takes place on account of the pure flexural fatigue.

Fatigue tests are generally conducted under two types of controlled loading, namely

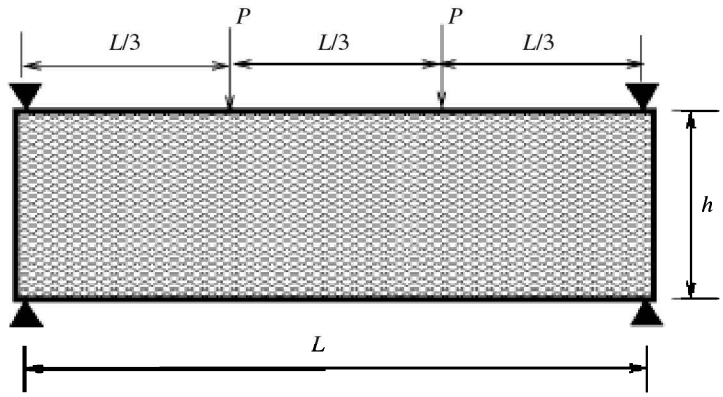


Figure 10.38 Schematic diagram of fatigue testing.

controlled stress amplitude and controlled strain amplitude. It is sometimes difficult, specially in controlled strain test, to establish the precise number of repetitions required for failure of the beam. As a consequence, arbitrary definitions of the failure condition of a specimen are adopted, for example, 50% reduction in the initial stress may be defined as the failure of the beam in constant strain testing. As can be justified from layered analysis of pavement (see Section 11.4.2), controlled stress fatigue testing is applicable to thick bituminous layer pavements while controlled strain fatigue testing is recommended for thin bituminous pavements [269, 56, 164]. Figure 10.39 shows a photograph of a fatigue test equipment. Here, the load is measured by a load cell and maximum central deflection is measured by an LVDT. By using the basic bending moment equation, the strain at the bottom fibre of the beam can be calculated. Also, the stiffness or dynamic modulus (discussed in the subsequent section) of the beam can be found out from the load and the maximum deflection.

Fatigue performance of bituminous mix

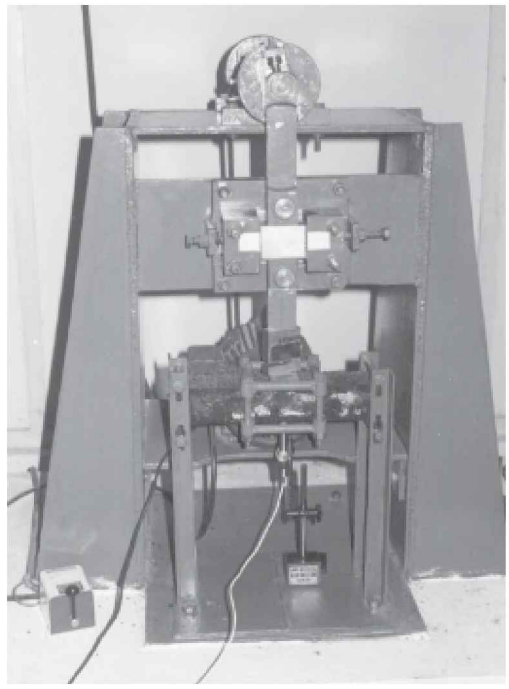


Figure 10.39 A beam fatigue set-up fabricated at IIT Kanpur.

Fatigue results obtained from the laboratory tests are generally expressed as a power law relation between the initial tensile strain ε_t at the bottom fibre of the beam, the dynamic modulus of the beam E_d and the number of load applications to failure N_f. It is derived from regression analysis of the experimental results. The general expression for fatigue test result on bituminous mix is given as

$$N_f = k_1\left(\frac{1}{\varepsilon_t}\right)^{k_2}\left(\frac{1}{E_d}\right)^{k_3} \tag{10.38}$$

where k_1, k_2, and k_3 are the regression constants.

Laboratory studies show that these constants are affected by material properties such as mixture stiffness, air voids, bitumen content, viscosity of bitumen, gradation of aggregates, dimensions of the test sample, temperature during the tests, and so on [165]. Figure 10.40 shows the results of a typical fatigue test carried out on Bituminous Concrete (BC) at different temperatures. Different temperatures resulted in different stiffness moduli (shown as E in Figure 10.40) of BC. As seen from the figure, the fatigue life of the mix for a given level of tensile strain is higher if the stiffness modulus E of

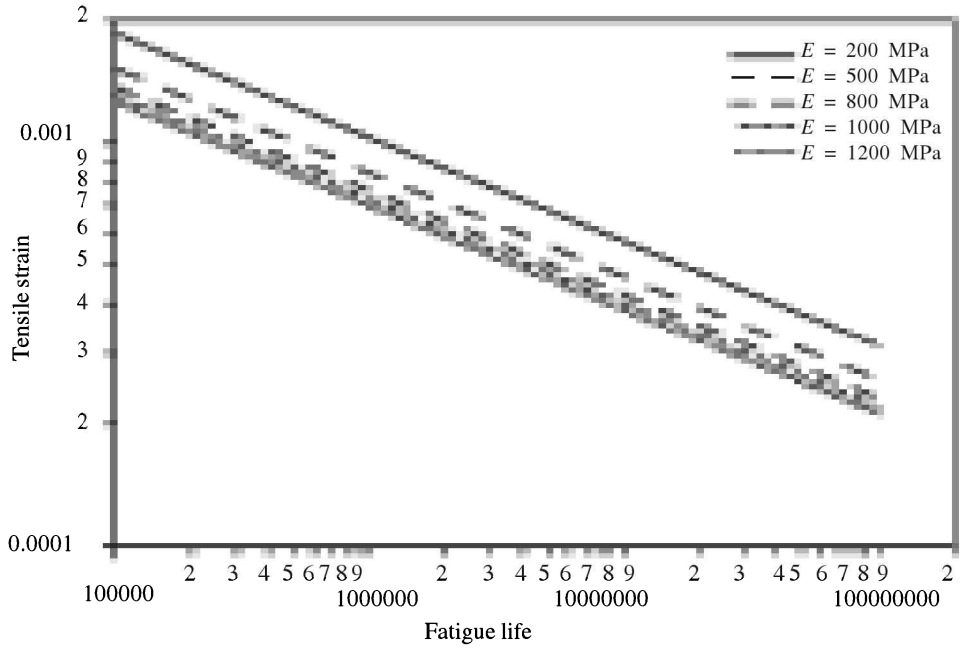


Figure 10.40 Laboratory fatigue curves on bituminous mixes [60, 217].

the mix is low. Thus, bituminous mixes with softer grades of bitumen, are expected to have better fatigue life.

Cumulative fatigue damage principle

M.A. Miner [160] while working on fatigue of aluminium metal, in 1945, developed a relationship concerning accumulation of fatigue damage. The same principle has been adopted for designing bituminous and concrete pavements [237, 238, 239]. The underlying principle is described as follows.

If at a strain level S_1, the sample is subjected to n_1 number of repetitions and at a strain level S_2, n_2 number of repetitions, and so on, . . ., up to the failure of the sample, then,

$$\frac{n_1}{N_1} + \frac{n_2}{N_2} + \frac{n_3}{N_3} + \cdots + = 1 \tag{10.39}$$

where N_1, N_2, N_3, . . ., are the number of repetitions for failure in respect of the individual fatigue load tests with strain levels S_1, S_2, S_3, . . ., respectively.

In the in-service road, axle loads of various magnitudes (i.e. different axle loads) undergo repetitions on the pavement. The above relationship [Eq. (10.39)] finds its application in such cases where the loads of different magnitudes can be considered together to find the cumulative fatigue damage.

Stiffness modulus of bituminous mix

Resilient modulus is generally used for strength characterization of soil and granular materials. The moduli used to characterize bituminous materials are complex modulus, dynamic modulus, and stiffness modulus. In bituminous mix, when a dynamic loading is applied, due to its viscoelastic nature, the load and the displacement do not go in the same phase. Therefore, the dynamic stress and strains can be expressed as real and imaginary parts. The dynamic stress divided by the dynamic strain gives the *complex modulus* of the mix. The magnitude of the complex modulus is known as *dynamic modulus*. The *stiffness modulus*, on the other hand, may be defined as the ratio of the maximum stress and maximum strain in a dynamic test, without giving any consideration to the phase difference between the stress and the strain.

As explained in this subsection, the load and the central deflection are measured in the fatigue test. Central deflection is a function of the magnitude of load, the loading configuration, and the stiffness modulus. Thus, the stiffness modulus can be found out as illustrated in the Example 10.6. Table 10.7 presents typical values of stiffness moduli of bituminous mixes, at various temperatures [61, 60, 89]. The bituminous mixes are taken as per the third revision of MOST specifications [217].

Table 10.7 Stiffness modulus of bituminous mixes at various temperatures

Mix type	*Temperature* (°C)				
	20	25	30	35	40
BC and DBM 80/100 bitumen	2300	1966	1455	975	797
BC and DBM 60/70 bitumen	3600	3126	2579	1615	1270
BC and DBM 30/40 bitumen	6000	4928	3809	2944	2276
BM 80/100 bitumen	—	—	—	500	—
BM 60/70 bitumen		—	—	700	—

EXAMPLE 10.6

In a fatigue test on bituminous mix, a single point sinusoidal load of 1500 N is applied with 10 Hz frequency to a specimen of size $(100 \times 100 \times 750)$ mm^3 (see Figure 10.41).

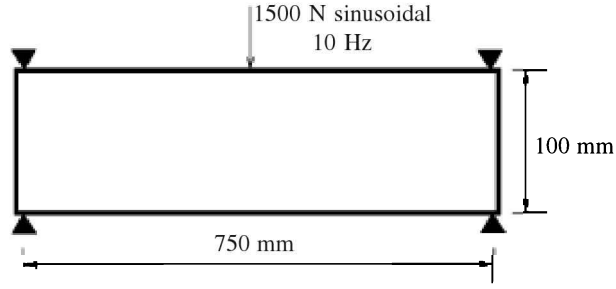


Figure 10.41 Example 10.6: fatigue test on bituminous mix.

A maximum deflection of 1.3 mm was noted and the specimen was found to sustain 5.3×10^7 repetitions till occurrence of failure (constant strain). Calculate the stiffness modulus of the bituminous mix at the test temperature.

Solution

The central deflection, $\delta = PL^3/48EI$, where P is the load, L is the length of beam, E is the stiffness modulus and I is the moment of area.

$$I = \frac{1}{12} \times 100 \times 100^3 = 8.333 \times 10^6 \text{ mm}^4$$

Substituting the known values

$$1.3 = \frac{1500 \times 750^3}{48E \times 8.33 \times 10^6}$$

Therefore, the stiffness modulus of the mix is

$$E = 1216.8 \text{ N/mm}^2 \approx 1217 \text{ MPa}$$

Closing remarks

Having discussed the concepts of fatigue life considerations of bituminous mixes, certain aspects of bituminous mix design and selection need reconsideration. The following remarks throw light on these aspects.

(i) From economic considerations, the lowest bitumen content may be chosen from the Marshall test. We have observed that the fatigue life improves if the bitumen content in the mix is increased. This is due to enhancement of inter-aggregate lubrication which leads to higher fatigue life. In the Example problem of the Marshall mix design, the OBC (from Figure 10.36) is obtained as 4.7%, which is the minimum bitumen content satisfying all the mix design parameters. However, if the fatigue is also considered as one of the parameters, an OBC of 5.1% will possibly be chosen, instead of 4.7% (see Figure 10.36).

(ii) Bituminous mixes made up of harder grade of bitumen exhibit high stiffness modulus. High stifness modulus causes less stress to the respective layers, and hence less distress in the pavement. But such bituminous mixes may have comparatively lower fatigue life, which is again an undesirable situation. This leads to a paradoxical conclusion where for better longevity of pavement, both high as well as low elastic moduli are needed [45]. A solution to this case could be thought, where a bituminous layer of higher stiffness (harder grade) is put above a bituminous layer of lower stiffness (softer grade). This concept is schematically presented in Figure 10.42, where the Alternative II is found to give the most economical design thicknesses [45] compared to the other two alternatives.

BC 80/100	BC 60/70	BC 60/70
DBM 80/100	DBM 80/100	DBM 60/70
Alternative I	Alternative II	Alternative III

Figure 10.42 Alternative bituminous surfacings with different bitumen grades.

(iii) For the aforesaid reasons, bitumen researchers are trying to evolve modified bitumen by adding admixtures to it, which enhance the fatigue life of the mix without compromising with its stiffness modulus. Some of the admixtures used for this purpose are natural rubber, crumb rubber, and polymers, (such as Styrene-Butadiene-Styrene (SBS), Ethylene-Vinyl-Acetate (EVA), Low Density Polyethylene (LDPE), etc.). The bitumen, thus derived, is called *modified binder*. A detailed discussion on modified binder, however, is beyond the scope of this book.

10.6 CEMENT

10.6.1 Composition

The basic constituents of cement are oxides of calcium, silicon, aluminium, iron, magnesium, sodium, and potassium in varying degrees of proportions. At a high temperature in a kiln where the cement is manufactured, the constituents combine and form complex compounds. The identification of these compounds is mainly based on Bogue's work and, hence they are called *Bogue's compounds*. Some of the Bogue's compounds are tricalcium silicate, dicalcium silicate, tricalcium aluminate, and tetracalcium aluminoferrite [175].

10.6.2 Manufacture

The raw materials used for manufacturing cement are limestone, chalk, shale, clay, etc. The raw materials are ground in suitable crushers, mixed in recommended proportions and then burnt in a kiln at about 1300–1500°C. There are two processes, namely the wet process and the dry process for the manufacture of cement. The dry process requires less fuel but mixing is difficult in this process. Clinkers formed after heating in the kiln, are ground into fine powder which is called cement.

Several types of cements can be formed depending upon the constituents used (additives and their relative proportions) and different processes of manufacture employed. Some of them can be enumerated as:

(i) Ordinary Portland cement

(ii) Low heat Portland cement

(iii) Air-entrainment cement
(iv) Rapid hardening cement
(v) Hydrophobic cement
(vi) Quick-setting cement
(vii) Sulphate resistant cement
(viii) Portland pozzolan cement
(ix) Expansive cement

10.6.3 Tests

Apart from the field tests, some of the laboratory tests for cement are:

(i) Fineness test
(ii) Setting time test
(iii) Strength test
(iv) Soundness test
(v) Heat of hydration test
(vi) Chemical composition test

10.7 CEMENT CONCRETE

Cement concrete is a homogeneous mixture of cement, coarse aggregate, fine aggregate, water, and any other admixture(s). The *characteristic strength* of concrete is defined as that compressive strength below which not more than 5% of the test results are expected to fall [186]. Concrete gradually gains its strength with time. However, as per convention, 28 days characteristic compressive strength, f_{ck}, is used for design purposes. It is advisable to perform tests on concrete to find its flexural strength, however, the flexural strength f_{cr} can also be estimated by the following formula recommended by the IS: 456 code [186]

$$f_{cr} = 0.7 \times \sqrt{f_{ck}} \tag{10.40}$$

where both the strength values are in N/mm^2.

Similarly, the modulus of elasticity of concrete E_c can be estimated as [186]

$$E_c = 5000\sqrt{f_{ck}} \tag{10.41}$$

where E_c is the short-term static modulus of elasticity, expressed in N/mm^2.

The Indian "Guidelines for the Design of Rigid Pavements for Highways" [91] assume the modulus of elasticity E of concrete as 3×10^4 MPa. The Poisson's ratio and the coefficient of thermal expansion are assumed as 0.15 and 10×10^{-6}/°C [91]

respectively. Conventionally, M40 concrete is chosen for the construction of concrete pavement, and its Modulus of Rupture (MR) value is assumed as 4.5 MPa [92]. A set of fatigue equations which can be adopted for M40 concrete used in the pavement construction [92] is

$$N_f = \text{unlimited} \qquad \text{for SR} < 0.45 \tag{10.42}$$

$$N_f = \frac{4.2577}{(\text{SR} - 0.4325)^{3.268}} \qquad \text{for } 0.45 < \text{SR} < 0.55 \tag{10.43}$$

$$\log_{10} N_f = \frac{0.9718 - \text{SR}}{0.0828} \qquad \text{for SR} > 0.55 \tag{10.44}$$

where SR is the stress ratio between the stress developed due to the wheel load and the flexural strength (modulus of rupture) of cement concrete. The largest size of aggregates should not be more than one-fourth of the thickness of the pavement [220] for concrete pavement construction. In case the pavement is reinforced, the maximum size of the aggregates should not exceed one-fourth of the minimum clear spacing between the reinforcing bars [220].

10.8 STABILIZED SOIL AND OTHER CEMENTED MATERIALS

Portland cement is the most commonly used cementing material for the construction of cemented base with soil, sand, and aggregate for pavements of roads and runways. Even low grade or marginal aggregates with suitable proportioning of coarse and fine fractions, and cement can be used as base and sub-base materials. Lime-clay, lime-laterite-sand-aggregate mixture, lime-granulated-blast-furnace-slag-soil mixture, etc. also form a strong cemented base or sub-base as indicated by laboratory tests. For such materials, it is the elastic modulus, flexural strength, and fatigue characteristics under repeated loading which are the guiding parameters for structural design [182].

Though lime is cheaper and initial reaction is quicker than cement in modifying clay soil. However, cement is a popular stabilizing material, possibly because of its gain in the strength level, applicability to wide range of soil and granular material, predictability in behaviour. After stabilization, carbonation of the lime may take place in the treated layer. Carbonation often commences on the surface of the layer and proceeds downwards, and may penetrate very rapidly causing loss of cementation. The way of reducing carbonation is the application of bituminous layer as soon as the base is cured [173]. Further discussion on various types of cemented materials is given in Sections 13.7.2 and 13.9.2.

Tables 10.8 and 10.9 summarize some Indian studies on structural properties and fatigue equations of cemented materials which may be used as input to the design of pavements with cemented base [13, 14, 181, 138, 156, 271, 148]. Design of bituminous pavements with cemented base/sub-base has been discussed in Section 12.4.1.

Table 10.8 Structural properties of some cemented materials

Cemented material	*E* (MPa)	*MR* (MPa)	*Remarks*
Lime-laterite soil	2240	1.250	OMC = 15%
Cement-laterite soil	2270	1.200	OMC = 16%
Lean concrete with morrum*	6000	1.111	—
Lean concrete with gravel*	7500	1.096	—
Lean concrete with dolerite*	8400	1.192	—
Soft Mizoram aggregates	3200	1.150	c:s:a = 1:7.5:7.5
Lime-flyash-sand	3600	1.0	OMC = 18%
Soil-lime-flyash	—	0.67	OMC = 12%

Note:
E = Elastic modulus, MR = Modulus of Rupture.
*cement: sand: aggregate (c:s:a) :: 1:6:12

Table 10.9 Fatigue equations of some cemented materials

Cemented material	*Fatigue equation*	r^2 *value*
Lime-laterite soil	$SR = 0.960 - 0.114 \times \log_{10}N$	0.762
Cement-laterite soil	$SR = 0.835 - 0.065 \times \log_{10}N$	0.830
Lean concrete with morrum	$SR = 1.045 - 0.144 \times \log_{10}N$	0.885
Lean concrete with gravel	$SR = 1.254 - 0.139 \times \log_{10}N$	0.857
Lean concrete with dolerite	$SR = 1.300 - 0.147 \times \log_{10}N$	0.884
Soft Mizoram aggregates	$SR = 1.000 - 0.065 \times \log_{10}N$	0.688
Lime-flyash-sand	$SR = 1.197 - 0.121 \times \log_{10}N$	0.846
Soil-lime-flyash	$SR = 0.990 - 0.026 \times \log_{10}N$	—

Note:
SR = stress ratio, r^2 = coefficient of determination.

EXERCISES

1. An unconventional granular material is planned to be used for road construction. How, as an engineer, will you accept or reject the material after testing in the laboratory?
2. Write a short note on resilient modulus of subgrade.
3. How can the Modulus of Subgrade Reaction of subgrade soil be estimated?
4. What is the difference between the crushing strength test and the impact test conducted on aggregates? How do you perform the shape test on an aggregate?
5. Suggest a suitable algorithm to determine the percentage mixing ratio of a number of aggregate batches to achieve a specified gradation.

6. What is the difference between bitumen and tar? How is asphalt produced?
7. How does a Brookfield viscometer work?
8. If the penetration test result of bitumen carried out at 50°C is 100, what is its viscosity?
9. What is the viscosity of bitumen at softening point?
10. Various bitumen samples obtained from different refineries are tested. Plot the expected nature of variation between softening points and penetration values (at 25°C) of the respective bitumen samples.
11. What is a cationic emulsifier?
12. What is Rolling Thin Film Oven test? What parameter of bitumen does it test?
13. How do you test for the purity of bitumen?
14. What are the desirable properties of a bituminous mix?
15. What is the difference between resilient modulus, elastic modulus, stiffness modulus, and dynamic modulus?
16. How is Optimum Bitumen Content (OBC) determined by the Marshall test?
17. Explain the physical significance of the theoretical maximum specific gravity used in the Marshall test.
18. Draw a curve between VMA and percentage bitumen in the total mix, and explain its variation.

ANNEXURE I

In this annexure, simple rheological models for spring, dashpot, and some combinations of spring–dashpot are presented. This forms a basis of formulation of a generalized model for bitumen where its performance due to static and dynamic loads can be predicted. The creep behaviour of a rheological model studies the response of the model when stress is kept constant with time, while the relaxation behaviour studies the response when strain is kept constant with time.

Spring

A spring model is made up of a simple spring, where stress is proportional to strain and the elastic modulus remains unchanged with respect to the level of stress and the strain rate.

Creep behaviour

$$\varepsilon = \frac{\sigma_0}{E} \tag{10.45}$$

Relaxation behaviour

$$\sigma = E\varepsilon_0$$

The response of a spring is shown in Figure 10.42.

Dashpot

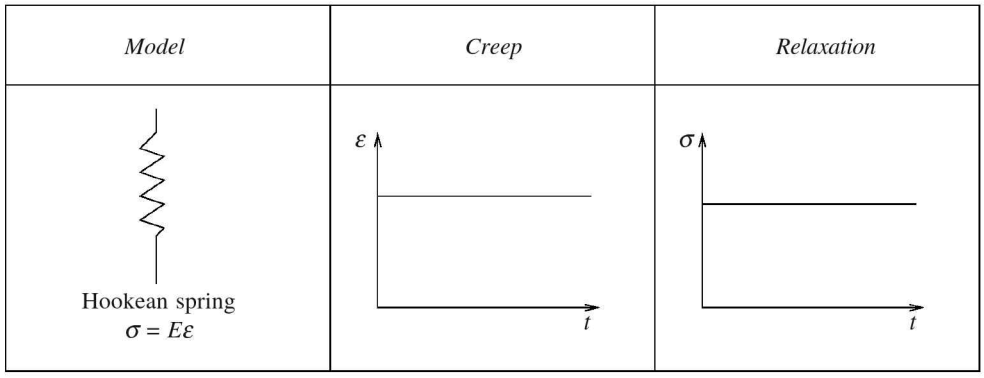


Figure 10.42 Response of a spring.

A dashpot causes damping in a structure. In a dashpot, the stress is proportional to the strain rate.

Creep behaviour

$$\eta = \frac{\sigma}{\varepsilon'}$$

or

$$\frac{d\varepsilon}{dt} = \frac{\sigma}{\eta}$$

or

$$\int_0^{\varepsilon} d\varepsilon = \int_0^{t} \frac{\sigma}{\eta} dt$$

or

$$\varepsilon = \frac{\sigma_0 t}{\eta} \tag{10.46}$$

Relaxation behaviour

$$\varepsilon' = 0$$

Therefore,

$$\sigma = 0$$

Figure 10.43 shows the response of a dashpot model.

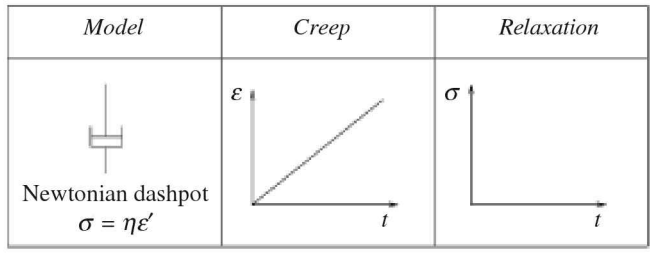


Figure 10.43 Response of a dashpot.

Maxwell Model

The Maxwell model constitutes a spring and a dashpot connected in series (see Figure 10.44). The equation used for the spring is $E = \sigma_1/\varepsilon_1$ and for the dashpot $\eta = \sigma_2/\varepsilon'_2$.

Now, for this model,

$$\sigma_1 = \sigma_2 = \sigma \quad \text{and} \quad \varepsilon_1 + \varepsilon_2 = \varepsilon$$

Putting the above conditions

$$\sigma + \sigma' \frac{\eta}{E} = \eta\varepsilon'$$

Creep behaviour

Here, $\sigma' = 0$
Therefore,

$$\sigma = \eta\varepsilon'$$

or

$$\int_0^t \sigma dt = \int_{\varepsilon_0}^{\varepsilon} \eta d\varepsilon$$

or

$$\sigma_0 t = \eta(\varepsilon - \varepsilon_0), \quad \text{as at } t = 0,\ \sigma = \sigma_0,\ \varepsilon = \varepsilon_0$$

or

$$\sigma_0 t = \eta\left(\varepsilon - \frac{\sigma_0}{E}\right), \quad \text{as at } t = 0,\ E = \frac{\sigma_0}{\varepsilon_0}$$

or

$$\eta = \frac{\sigma_0 t}{\eta} + \frac{\sigma_0}{E} \tag{10.47}$$

Relaxation behaviour

Here, $\varepsilon' = 0$
Therefore,

$$\sigma + \sigma' \frac{\eta}{E} = 0$$

or

$$\frac{d\sigma}{dt}\frac{\eta}{E} = -\sigma$$

or

$$\int_{\sigma_0}^{\sigma} \frac{d\sigma}{\sigma} = \int_0^t \frac{-E}{\eta} dt$$

or

$$\ln\frac{\sigma}{\sigma_0} = \frac{-Et}{\eta}$$

or

$$\sigma = E\varepsilon_0 e^{-\frac{Et}{\eta}}, \text{ as at } t = 0,\ E = \frac{\sigma_0}{\varepsilon_0} \tag{10.48}$$

Figure 10.44 shows the response of the Maxwell model.

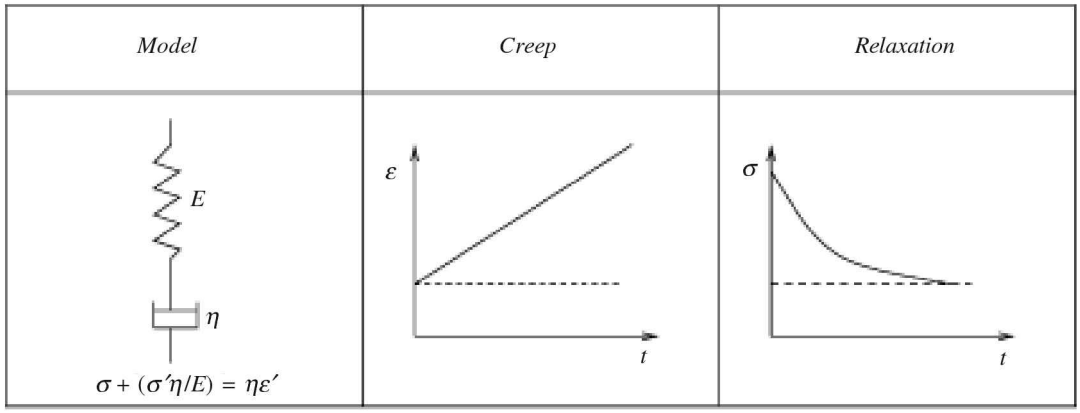


Figure 10.44 Response of the Maxwell model.

Kelvin Model

A spring and a dashpot connected in parallel constitute the Kelvin model. Derivations for the Kelvin model are carried out in a similar fashion as in the Maxwell's model, and the response is shown in Figure 10.45.

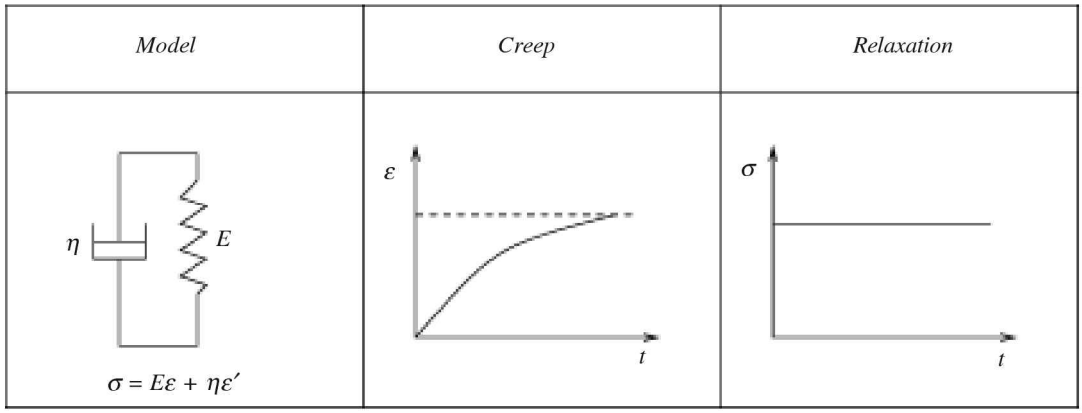


Figure 10.45 Response of the Kelvin model.

Standard Solid Model

The Kelvin and Maxwell models do not give true representative trend of creep and relaxation behaviour of a viscoelastic material, like bitumen. A standard solid model, which is a combination of these two models is, therefore, proposed. The creep and relaxation trends of this model are somewhat closer to the behaviour of bitumen. In the standard solid model, a spring and a dashpot are connected in parallel, and this arrangement is itself connected in series with

a spring, as shown in Figure 10.46. The relationships, which can be used are:

$$\sigma_1 + \sigma_2 = \sigma \quad \text{and} \quad \varepsilon_1 + \varepsilon_2 = \varepsilon$$

The second equation can be written as

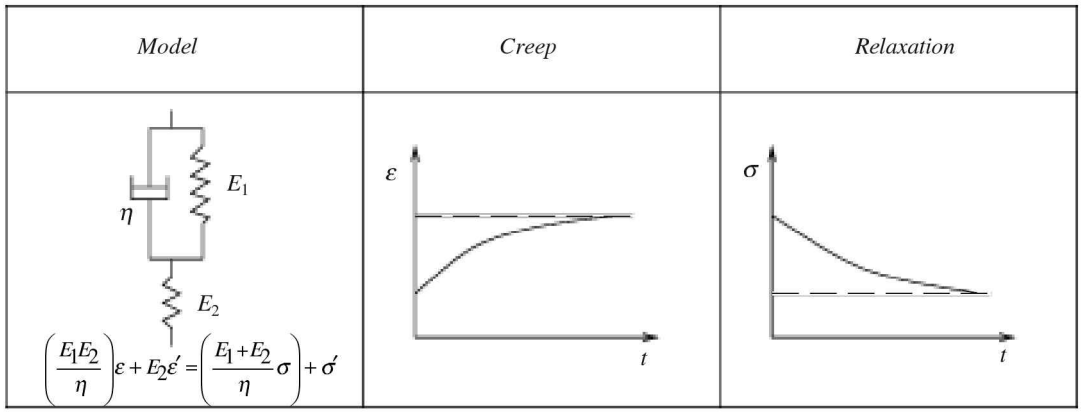


Figure 10.46 Response of the standard solid model.

$$\varepsilon_1 + \frac{\sigma}{E_2} = \varepsilon$$

or

$$\varepsilon'_1 = \varepsilon' - \frac{\sigma'}{E_2}$$

Now,

$$E_1\varepsilon_1 + \eta\varepsilon'_1 = \sigma$$

or

$$E_1\left(\varepsilon - \frac{\sigma}{E_2}\right) + \eta\left(\varepsilon' - \frac{\sigma'}{E_2}\right) = \sigma$$

or

$$\left(\frac{E_1E_2}{\eta}\right)\varepsilon + E_2\varepsilon' = \left(\frac{E_1 + E_2}{\eta}\right)\sigma + \sigma' \tag{10.49}$$

Equation (10.49) represents the standard solid model.

Creep behaviour

Here,

$$\sigma = \sigma_0 \quad \text{and} \quad \sigma' = 0$$

Substituting the above conditions in Eq. (10.49), multiplying both sides by $e^{\frac{E_1 t}{\eta}}$, and then integrating, the following equation is obtained:

$$\varepsilon = \frac{\sigma_0}{E_2} e^{\frac{-E_1 t}{\eta}} + \sigma_0 \frac{(E_1 + E_2)}{E_1 E_2} \left(1 - e^{\frac{-E_1 t}{\eta}} \right) \tag{10.50}$$

Relaxation behaviour

Similarly, for the relaxation case, the following equation is obtained.

$$\sigma = E_2 \varepsilon_0 e^{-\frac{E_1 + E_2}{\eta}} + \frac{E_1 E_2}{E_1 + E_2} \varepsilon_0 \left[1 - e^{\frac{E_1 + E_2}{\eta} t} \right] \tag{10.51}$$

Dynamic loading

If a dynamic loading, $\sigma = \sigma_0 e^{i\omega t}$ is assumed to be applied to the generalized solid model, then

$$\sigma' = \sigma_0 i\omega e^{i\omega t} = i\omega\sigma$$

The strain ε developed will have a phase difference of δ, therefore,

$$\varepsilon = \varepsilon_0 e^{i(\omega t + \delta)}$$

and then,

$$\varepsilon' = i\omega\varepsilon_0 e^{i\omega t} = i\omega t$$

Substituting these in the general equation of the standard solid model, the

$$i\omega\sigma + \frac{2E}{\eta}\sigma = E(i\omega\varepsilon) + \frac{E^2}{\eta}\varepsilon \tag{10.52}$$

Therefore, complex modulus (see Section 10.5.3 for definition)

$$E^* = \frac{\sigma}{\varepsilon}$$

$$= \frac{E^2 + i\omega\eta E}{2E + i\omega\eta}$$

$$= \frac{2E^3 + \omega^2\eta^2 E}{4E^2 + \omega^2\eta^2} + i\frac{\omega\eta E^2}{4E^2 + \omega^2\eta^2} \tag{10.53}$$

$$\text{Dynamic modulus, } E_d = |E^*| = \frac{[(2E^3 + \omega^2\eta^2 E)^2 + (\omega\eta E^2)^2]^{1/2}}{4E^2 + \omega^2\eta^2}$$

and

$$\text{Phase angle, } \delta = \tan^{-1} \frac{\omega\eta E}{2E^2 + \omega^2\eta^2}$$

ANNEXURE II

Concepts of Superpave

The concepts of Superpave (SUperior PERforming asphalt PAVEment) evolved from the Strategic Highway Research Program (SHRP) studies. The basic concepts involved in Superpave are briefly mentioned below [224, 223].

Binder selection

Temperature is an important parameter that affects binder behaviour. Superpave mix specification, thus, identifies that acceptability tests should be carried out at the prevalent field temperatures, and not in a laboratory specified temperature. This is an important and unique consideration because bitumen from two different sources may show the same physical properties at a particular temperature, and which may vary drastically at other temperatures. Thus in Superpave, only the acceptable test values and not the test temperatures are recommended. Rather, the temperatures are determined from the most prevalent maximum and minimum temperatures of the field, at a given probability level. As per Superpave designation, a binder is identified as PG T_1T_2, where PG stands for performance grade and T_1 and T_2 are the temperature limits under which the performance of that binder is tested to be acceptable.

Rolling Thin Film Oven Test (RTFO), Pressurized Aging Vessel (PAV), Dynamic Shear Rheometer, Rotational Viscometer, Bending Beam Rheometer, and Direct Tension Tester are some of the tests recommended in Superpave binder selection. Dynamic shear rheometer is used to measure viscosities of bitumen at high and intermediate temperatures. Rotational viscometer is used for high temperature viscosity measurements, like those at the time of mixing or pumping. Bending beam rheometer measures the stiffness of bitumen at low temperatures.

Aggregate selection

The criteria of specifying the 'consensus aggregate properties' mainly depends on the existing in-service traffic level and the position of the granular layer inside the pavement structure. These properties mainly include the coarse aggregate angularity, fine aggregate angularity, flakiness and elongation, clay content of aggregate, abrasion test,

soundness test, etc. Angularity ensures adequate shear strength due to aggregate interlocking, and limiting flakiness ensures that aggregates do not break during compaction and handling.

SHRP formed a 14 member expert task group to evolve an appropriate aggregate gradation to be used for Superpave. This group, after several experiments and discussions, decided to use a power gradation of 0.45 [see Eq. [10.17)] as the reference gradation with certain restriction zones and control points. The restriction zone and control points were incorporated in order to ensure certain proportion of fines for the following reasons.

(i) Proper interlocking of aggregates

(ii) To avoid the fall in shear strength of the mix due to excess of fines

(iii) To maintain the requisite VMA

These control points and restriction zones serve more as guidelines for selecting a gradation rather than a compulsion to be followed. Figure 10.47 represents the aggregate gradation recommended by Superpave. While specifying the aggregate gradation, Superpave used the following definitions for nominal maximum size and maximum size of aggregates.

- The nominal maximum size is defined as one sieve size larger than the first sieve to retain more than 10%.
- The maximum size is one sieve size larger than the nominal maximum size.

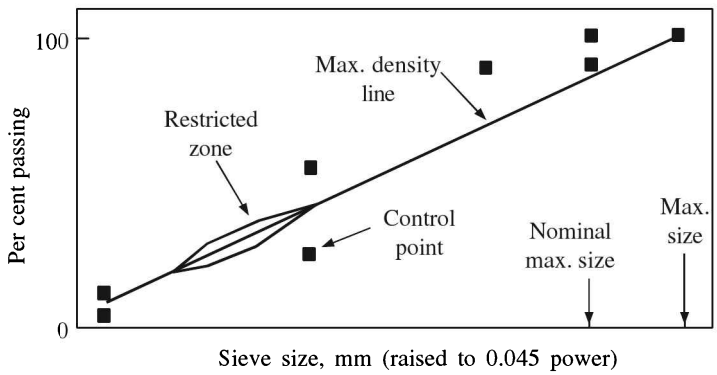


Figure 10.40 Recommended Superpave gradation [224].

Superpave volumetrics

The volumetrics already discussed in Section 10.5.1 is that used by Superpave. Superpave emphasizes the fact that some amount of bitumen is absorbed by the aggregates and therefore remains ineffective. Thus, the amount of bitumen which is effectively available in the mix should be considered for deciding the optimum bitumen content.

Compaction

Superpave developed a new compactor named Superpave Gyratory Compactor (SGC), which was developed keeping in mind the specific objectives as follows:

- The compaction should be such that it simulates the compaction in the field loading and climatic conditions.
- The compactor should be able to accommodate large aggregates.
- The device should be portable enough so that it can be used in field for quality control purposes.

The principles of French gyratory compactor were used to modify the Texas gyratory compactor, and finally the SGC evolved. The mould can accommodate maximum aggregate size of up to 50 mm (mould size 150 mm). In SGC, the test mould is placed on a tilted base (1.25°), rotated at a certain frequency (30 gyrations per minute) while being subjected to compaction pressure of 600 kPa (or more, depending on the field condition). It simulates the kneading action of rollers used to compact asphalt concrete pavements by applying a vertical load to a mixture with tilted mould. Height can be measured at any point of the test, which in turn gives the measure of the compaction level achieved. It may be noted that the gyratory compactor uses shear compaction effort whereas the Marshall compactor uses impact energy for compaction.

Evaluation of mix performance

The following are the two major tests recommended by Superpave to test the mixture performance.

(i) *Superpave Shear Tester.* It is used to evaluate the permanent deformation (rutting) and fatigue cracking susceptibilities in asphalt mixtures. It provides vertical and horizontal loads to a cylindrical specimen at different confining conditions and temperatures. The confinements and the test temperature are fixed so as to simulate the field conditions. Uniaxial strain test, repeated shear test, simple shear test, and frequency sweep test are some of the tests carried out by this equipment.

(ii) *Superpave Indirect Tensile Tester.* It measures the creep compliance and tensile creep of hot mix asphalt. It is also used to evaluate the thermal cracking susceptibility of asphalt mixtures. Load is applied to a cylindrical asphalt concrete specimen through its diametrical axis and the resulting deformations are measured for static creep load and repetitive dynamic loading.

Pavement Analysis

11.1 INTRODUCTION

A pavement mainly consists of a carriageway and shoulder. Figure 3.2 shows the various components of a typical road stretch. Pavements are generally classified into two major categories, namely bituminous and concrete pavements, which are conventionally known as flexible and rigid pavements. No real structure, however, can behave as perfectly rigid or flexible. Rather, both types of pavements may be considered as layered structures, in which the concrete pavement has a very high elastic modulus at the top layer compared to the bituminous pavement. In this book, the terms concrete and bituminous pavements are more often used than the terms rigid and flexible pavements.

This chapter is divided into four sections. The pavement compositions for bituminous and concrete pavements have been discussed in the first section. The second section discusses the various input parameters involved in pavement analysis. The third and fourth sections present an analysis of bituminous and concrete pavements.

11.2 PAVEMENT COMPOSITION

In Chapter 10, the engineering properties of various pavement materials have been discussed. Depending on the material constituents of individual layers, the layers possess varied strength characteristics, and this information is used as the input for analysis purposes. In general, the elastic modulus of the top layer of the pavement is the highest, and it gradually decreases downwards. There may be, of course, exceptions to this case. This section briefly discusses the composition and various layers of the bituminous and concrete pavements.

11.2.1 Bituminous Pavement

The bituminous pavement is generally composed of a subgrade, a sub-base, a base, a binder, and a wearing course. Subgrade is a compacted soil layer while base and sub-

base are granular or cemented layers. The binder course and the wearing course, for the bituminous pavement, are made up of bituminous material, and they are together called bituminous surfacing. The wearing course is the topmost layer which is in direct contact with the tyre, and provides structural strength (in some cases), protection to the surfacing, and impermeability to water percolation. Figure 11.1 shows the cross-section of a typical bituminous pavement.

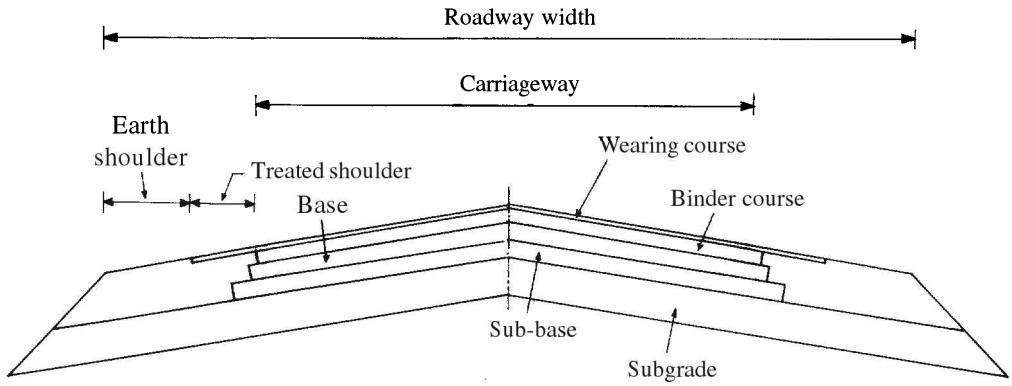


Figure 11.1 Cross-section of a typical bituminous pavement.

A detailed discussion on composition and construction of the individual layers of bituminous pavement has been covered in Section 13.7.

11.2.2 Concrete Pavement

The concrete pavement is constructed directly over subgrade, or a base layer made up of a stabilized material, lean cement concrete, or granular material is used. Steel reinforcement is sometimes put in the concrete pavement as temperature steel only. Composite pavements also exist where the bituminous overlay is put over the concrete surface, or otherwise.

Section 13.9 discusses in detail the composition and construction of the individual layers of concrete pavement. Figure 11.2 shows a representative cross-section of a typical concrete pavement. The number of layers, constituents, and even their order may vary from one section of the road to the other, depending upon the individual pavement design. Types of concrete pavements, joints in concrete pavements, and the use of steel in concrete pavements are briefly discussed in the following.

Types of concrete pavements

Plain concrete pavement. Plain concrete pavements are constructed without any reinforcement. The load transfer at the joints takes place through aggregate interlock.

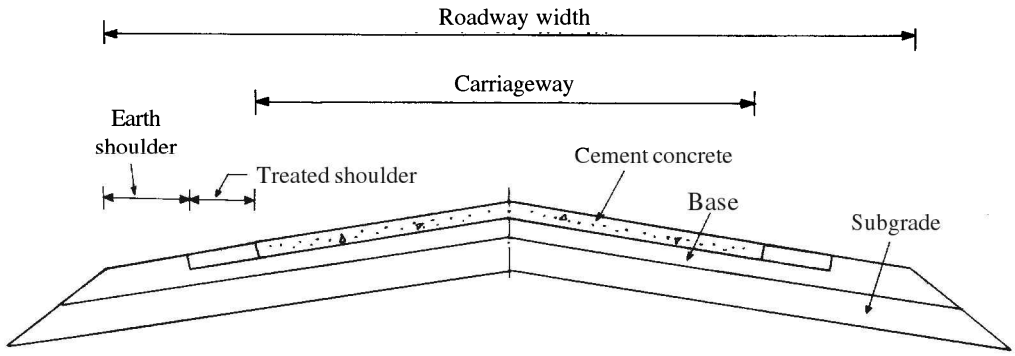


Figure 11.2 Cross-section of a typical concrete pavement.

Steel tie bars are sometimes used longitudinally for providing a warping joint between two lanes.

Plain dowelled concrete pavement. This pavement is constructed with plain concrete, except at transverse joints where steel dowel bars are provided for load transfer. Tie bars are used as longitudinal joints.

Continuously reinforced concrete pavement. This type of concrete pavement is reinforced throughout, and is without any contraction or expansion joint. The pavements develop transverse cracks, and the reinforcement bars of steel act as load transfer devices at these cracks [239].

Prestressed concrete pavement. The design thickness of the prestressed concrete pavements is less than that of the plain concrete pavements. Prestressed concrete pavements have been used on experimental basis at some airports, but are not popular for use on highways.

Joints in concrete pavements

Joints in concrete pavements are used to:

(i) Release stresses induced due to temperature variation

(ii) Provide proper bonding between two portions of concrete slabs, where there is a time lapse between the two phases of construction.

Joints are of various types, and can be classified on the basis of function, location, etc. Figure 11.3 displays longitudinal and transverse joints in a two-lane concrete pavement while Figure 11.4 is a schematic representation of the various types of joints. The following is a brief discussion on various types of joints.

Expansion joints. Concrete expands with increase in temperature. Unless there is a provision for expansion, the concrete slab may buckle outwards and break. Expansion

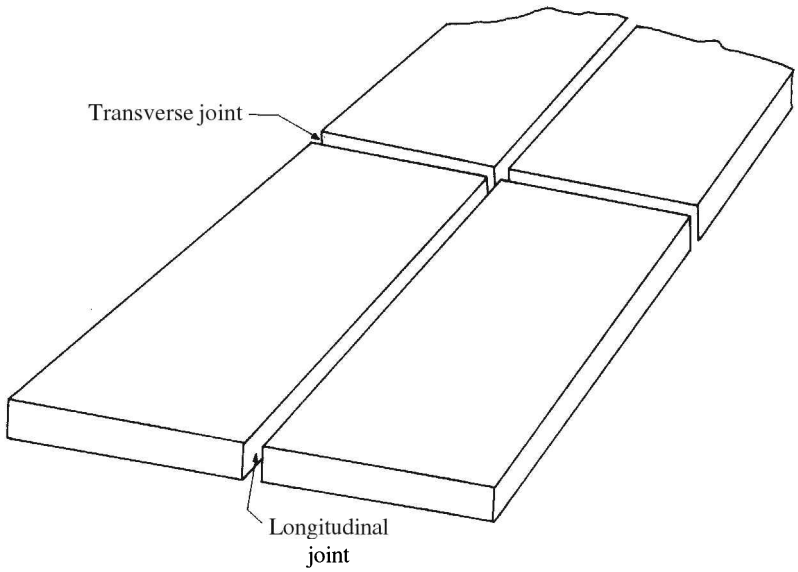


Figure 11.3 Longitudinal and transverse joints in a two-lane concrete pavement.

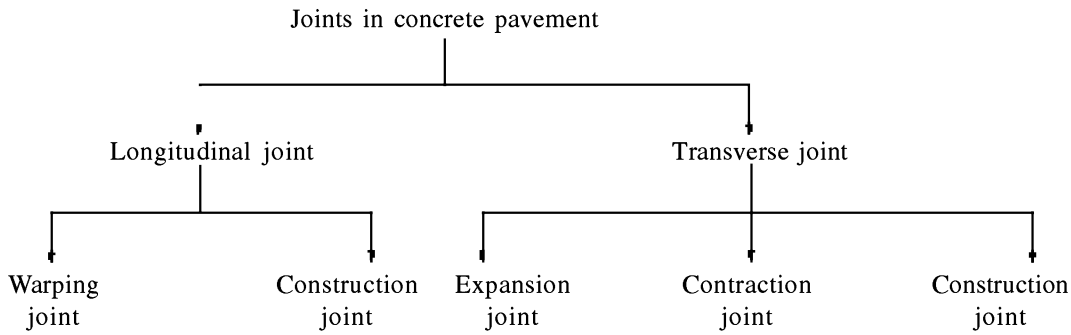


Figure 11.4 Various types of joints used in concrete pavement.

joints are full depth joints provided at specified intervals along the transverse direction of the pavement. Generally, dowel bars are placed across the expansion joints to take care of the load transfer. One side of the dowel bar is fixed with one concrete slab while the other side is generally lubricated and put inside an expansion cap placed within the other concrete slab. This allows free expansion of the concrete slabs. Figure 11.5 shows a typical expansion joint.

Contraction joints. Contraction joints are provided to take care of shrinkage of concrete slabs. They are provided along the transverse direction, at regular intervals. Dowel bars may be or may not be placed to assist the load transfer mechanism between

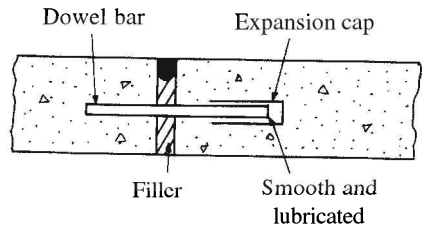


Figure 11.5 A typical expansion joint.

the two slabs. Sometimes, a dummy groove (or partial cut) is placed as a contraction joint [220]. Figure 11.6 shows a typical contraction joint.

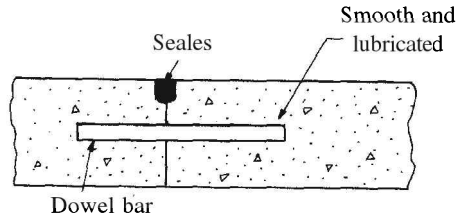
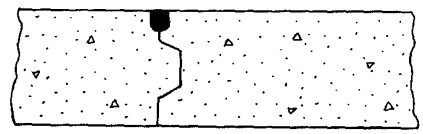
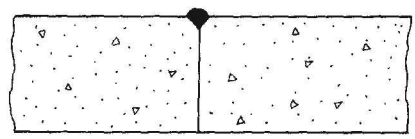


Figure 11.6 A typical contraction joint.

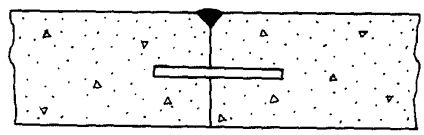
Warping joints. Warping joints relieve stresses due to warping. Figure 11.7 shows some warping joints. Warping joints are generally longitudinal joints.



A warping joint (hinged)



A warping joint (plain butt joint)



A warping joint (with tie bar)

Figure 11.7 Examples of some warping joints.

Construction joints. Construction joints are the joints between pavement sections that are constructed at different periods of time. A construction joint could be along the longitudinal direction or the transverse direction, and an effort is made such that it coincides with the locations of other joints. Thus, a construction joint may be an expension/contraction joint along the transverse direction or a warping joint along the longitudinal direction. As per the modern practice of concrete pavement construction, construction is carried out round the clock and thereby provision of construction joints is eliminated [218].

Steel in concrete pavements

As already discussed, concrete pavement is generally made up of plain concrete. Steel is only used for special temperature reinforcement purposes, or as dowel bars and tie bars. Steel in wire-mesh form, sometimes, is put within the concrete pavement as temperature steel. This temperature steel mesh impregnated within concrete does not prevent cracking of pavement, but controls the propagation of further cracks. Dowel bars are placed at the joints where load transfer takes place. The dowel bars provide flexural and bearing resistance and shearing across the joints. Tie bars hold two concrete slabs together and are not intended to participate in the load transfer.

11.3 PARAMETERS FOR PAVEMENT ANALYSIS

11.3.1 Elastic Modulus

As discussed in Chapter 10, elastic moduli of various pavement materials are obtained either through tests or through the recommendations available in the guidelines. Elastic moduli of subgrade soil and granular bases, are determined by repeated triaxial tests. Repeated flexure or indirect tensile tests are carried out to determine the dynamic modulus E_d of bituminous mixes.

11.3.2 Poisson's Ratio

For most of the pavement materials, the influence of Poisson's ratio μ, is usually small. This allows use of typical constant values of μ for analysis rather than direct testing [162]. The μ for clayey subgrade varies from 0.4 to 0.5 and a value of 0.5 is used for the wet condition [22]. The μ value of 0.35 may be assumed for sand [162]. For most of the cement treated materials (soil cement, cement treated base, lean concrete, and PCC), the value of μ normally lies between 0.10 and 0.25, the value of 0.15 being the most common [91, 52]. Typical values of μ for unbound granular material lie between 0.2 and 0.5 and those for bituminous mixes range from 0.35 to 0.50. A value of 0.50 is generally relevant at higher temperatures [183]. At low temperatures, the μ value of bituminous material is low, while it increases with the increase in temperature [183].

11.3.3 Wheel Load, Wheel Configuration, and Tyre Pressure

The weight of the vehicle is distributed among its axles, and the axles transmit the load to its wheels. These wheels constitute the loading configuration of a pavement structure, in which stresses are induced. Figure 11.8 shows some types of axles and wheel configurations of trucks. The concept of Equivalent Single Wheel Load (ESWL) is used to convert the equivalent effect due to dual wheels (or number of wheels with a given configuration) to a single wheel load. The effect of this load can be in terms of equivalent stress, strain, or surface deflection. ESWL is gradually becoming obsolete due to the fact that computational analysis of resultant response (stress, strain, or displacement) of any given wheel configuration can now be obtained computationally. Such an equivalence which involves further approximation (refer to the footnote discussion in Section 10.3.2), is not preferred nowadays.

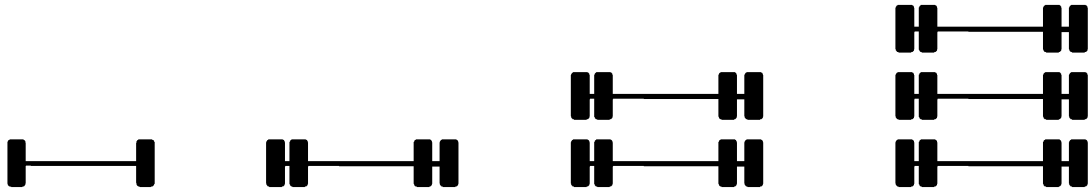


Figure 11.8 Axle types and wheel configurations.

Tyre pressure is another parameter, which determines the load contact area used in pavement analysis. Out of the two wheels having the same weight, the one whose tyre pressure (more precisely, tyre contact pressure) is less will have a larger area of tyre imprint. A minor variation in tyre contact pressure does not cause any substantial change in the stresses inside the pavement [209, 60]. Further discussions on these parameters have been covered in Section 12.2.2.

11.3.4 Temperature

Temperature affects the stiffness modulus of bituminous mixes. The variation of stiffness values with temperature for the bituminous mixes based on Indian specifications, has already been presented in Table 10.7. For a concrete pavement, the temperature gradient across its thickness induces warping stresses, and the same has been discussed in Section 11.5.2.

11.4 ANALYSIS OF BITUMINOUS PAVEMENT STRUCTURES

There are numerous techniques and subsequent modifications to the existing theories for the analysis of the bituminous and concrete pavements. The basic analysis techniques consist of analysis of the elastic multilayered structure pioneered by Burmister [25] (analysis of bituminous pavement) and analysis of finite slabs on elastic foundation (analysis of concrete pavement) pioneered by Westergaard [261]. Only the salient features of the analysis proposed by them are discussed in this chapter; detailed derivations and further modifications are beyond the scope of this book.

11.4.1 Elastic Half-Space Solution

Equilibrium and compatibility equations

The equilibrium equations of an element subjected to a non-uniform stress field, are obtained by equating the total forces in the respective coordinate axes to zero, that is, $\Sigma F_x = 0$, $\Sigma F_y = 0$, and $\Sigma F_z = 0$ in the rectangular coordinate system. The strain compatibility equations are obtained from the definitions of strain components and the generalized Hook's law. The concept of Airy's stress function shows that the solutions of the equilibrium and the compatibility equations become simpler and it converges to finding out the solution of only one equation, i.e. $\nabla^4 \phi = 0$, where ∇^2 is the Laplace's harmonic operator and ϕ is the Airy's stress function. The detailed derivation can be found in any book on elasticity [241] or geomechanics [131, 48].

In the cylindrical coordinate system, Love [147] had shown that the following equations [Eqs. (11.1)–(11.4)], are the stress variables (refer Figure 11.9) which can be expressed in terms of the Airy's stress function, ϕ. Body forces have been neglected here in these equations.

$$\sigma_r = \frac{\delta}{\delta z}\left[\mu\nabla^2\phi - \frac{\delta^2\phi}{\delta r^2}\right] \tag{11.1}$$

$$\sigma_\theta = \frac{\delta}{\delta z}\left[\mu\nabla^2\phi - \frac{1}{r}\frac{\delta\phi}{\delta r}\right] \tag{11.2}$$

$$\sigma_z = \frac{\delta}{\delta z}\left[(2-\mu)\nabla^2\phi - \frac{\delta^2\phi}{\delta z^2}\right] \tag{11.3}$$

$$\sigma_z = \frac{\delta}{\delta z}\left[(1-\mu)\nabla^2\phi - \frac{\delta^2\phi}{\delta z^2}\right] \tag{11.4}$$

These equations can satisfy the equilibrium equations and the compatibility equations, if the function ϕ is chosen as a biharmonic, i.e. $\nabla^4\phi = 0$. Similarly, in a two-dimensional polar

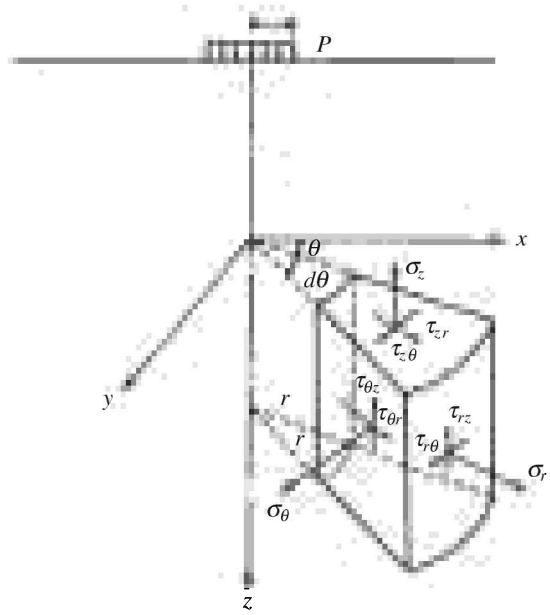


Figure 11.9 Stresses in an element in the cylindrical coordinate system.

coordinate system, the equations take the following form [131]:

$$\sigma_r = \frac{1}{r}\frac{\delta\phi}{\delta r} + \frac{1}{r^2}\frac{\delta^2\phi}{\delta\phi^2} \tag{11.5}$$

$$\sigma_\theta = \frac{\delta^2\phi}{\delta r^2} \tag{11.6}$$

$$\tau_{r\theta} = -\frac{\delta}{\delta r}\left(\frac{1}{r}\frac{\delta\phi}{\delta\phi}\right) \tag{11.7}$$

Boussinesq's solution

Boussinesq [17] gave the equations for computation of stress and strain when a concentrated force P acts on the horizontal boundary surface of an elastic, weightless, semi-infinite body, called *elastic half-space*. He assumed a logarithmic biharmonic stress function ϕ, and solved it for the values of stresses at a point within the soil mass. The boundary conditions assumed by Boussinesq were:

(i) All stresses vanish at infinity.

(ii) The shear stress component at any point on the ground surface is zero.

(iii) σ_z on surface is zero except at the point of application (where it is not defined).

(iv) The sum of the vertical force components along any imaginary hemisphere (with load

point as centre) is equal to the vertical load applied [131].

The expressions for stresses in polar coordinates as obtained by Boussinesq (refer Figure 11.10, assume that vertical point load P is applied at O) are:

$$\sigma_z = \frac{3P}{2\pi R^2}\cos^3\beta \tag{11.8}$$

$$\sigma_z = \frac{P}{2\pi R^2}\left(3\sin^2\beta\cos\beta - \frac{1-2\mu}{1+\cos\beta}\right) \tag{11.9}$$

$$\sigma_\theta = \frac{P}{2\pi R^2}(1-2\mu)\left(-\cos\beta - \frac{1}{1+\cos\beta}\right) \tag{11.10}$$

$$\tau_{rz} = \frac{3P}{2\pi R^2}\sin\beta\cos^2\beta \tag{11.11}$$

$$\delta = \frac{(1+\mu)P}{2\pi RE}(2-(1-\mu)+\cos^2\beta) \tag{11.12}$$

where

δ is the deformation

μ is the Poisson's ratio

E is the elastic modulus of the half-space

R is the radial distance

β is the angle as shown in Figure 11.10.

Circular loading

Boussinesq's [17] solution gives stresses and deformations due to a vertical point load, which can be suitably integrated to find out the response due to various kinds of loading, such as line loading, circular uniform loading, strip loading, etc. Separate solutions are derived by others for load along the horizontal direction, or for load within the half-space.

Figure 11.10 shows how the elastic half-space response due to circular uniform loading (of pressure p) can be estimated from Boussinesq's formulation. An infinitesimal vertical point load of $prd\theta dr$ is assumed on the circular disc, and its effect at point P is estimated. The distance l can be expressed as $(z^2 + r'^2)$ as shown in the diagram. Again r' can be expressed in terms of R', θ, and r (from the triangle property), where R' is the coordinate distance of point P and the remainder two parameters are the integration variables themselves. Thus, integration can be performed over the whole circular area to find the overall effect due to circular uniform loading on point P. For the σ_z computation, the components are directly added, but for computation of other stresses (σ_r, σ_θ, and τ_{rz}), proper directional component needs to be considered. Assuming that the loading plate is flexible and smooth, i.e. the pressure distribution is uniform and the horizontal displacement of soil in contact with plate is permitted, the following expression for

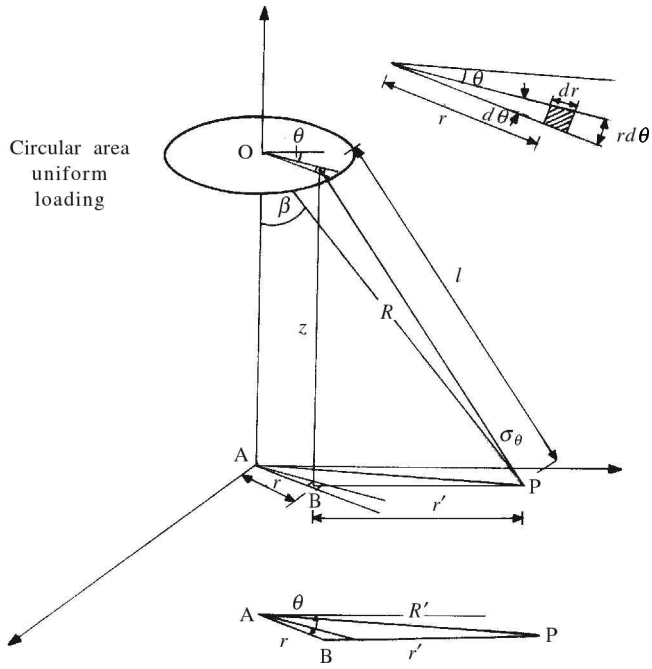


Figure 11.10 Stress computation in elastic half-space due to circular uniform loading.

the central vertical deflection δ below a flexible plate is obtained [266] as

$$\delta = \frac{(1+\mu)pa}{E}\left[\frac{a}{(a^2+z^2)^{0.5}} + \frac{1-2\mu}{a}((a^2+z^2)^{0.5} - z)\right] \tag{11.13}$$

Substituting the value of μ as 0.5 and the value of z as 0, the surface deflection is obtained as

$$\delta = 1.5 \times \frac{pa}{E} \tag{11.14}$$

Equation (11.14) is derived for the flexible plate where the pressure distribution is uniform on the plate but deflections at different points vary. In a rigid plate, on the other hand, the deflection at various points on the plate remains the same, but the pressure distribution is non-uniform. The equation for the rigid plate surface deflection can be written as

$$\delta = \frac{\pi(1-\mu^2)pa}{2E} \tag{11.15}$$

Assuming $\mu = 0.5$, this equation takes the form

$$\delta = 1.18 \times \frac{pa}{E} \tag{11.16}$$

EXAMPLE 11.1

A plate load test is carried out on subgrade soil using a 300 mm radius rigid plate. A load of 5 tonnes resulted in a deflection of 1.2 mm. Determine the elastic modulus of the soil, if the Poisson's ratio is 0.5.

Solution

The pressure applied is

$$\frac{5 \times 10^4}{\pi \times 300^2} = 0.176 \text{ N/mm}^2$$

Substituting the values in Eq. (11.16), we have

$$1.2 = \frac{1.18 \times 0.176 \times 300}{E}$$

Therefore,

$$E = 52.16 \text{ MPa}$$

11.4.2 Layered Elastic Solution

A pavement is a multilayered structure. If it is assumed to behave elastically under the application of fast moving wheel loads, each layer may be characterized by its elastic modulus E, Poisson's ratio μ, and the thickness h. Figure 11.11 shows a layered pavement structure, with uniform circular pressure of radius a acting on the top of it. Burmister [25, 26] gave a mathematical solution for a layered elastic structure. This analysis was based on certain assumptions such as:

(i) The first $(n - 1)$ layers are horizontally infinite.

(ii) The nth layer is a semi-infinite mass.

(iii) Inertial forces are negligible.

(iv) The stresses and displacements at infinite depths are zero.

(v) The interfaces could either be rough or smooth.

The following model briefly describes the approach proposed by Burmister [25, 26] and Verstraeten [253, 254]. The stress function adopted by Burmister [25] is given as

$$\phi_i = J_0(mr)[(A_i + C_i z)e^{mz} - (B_i + D_i z)e^{-mz}] \tag{11.17}$$

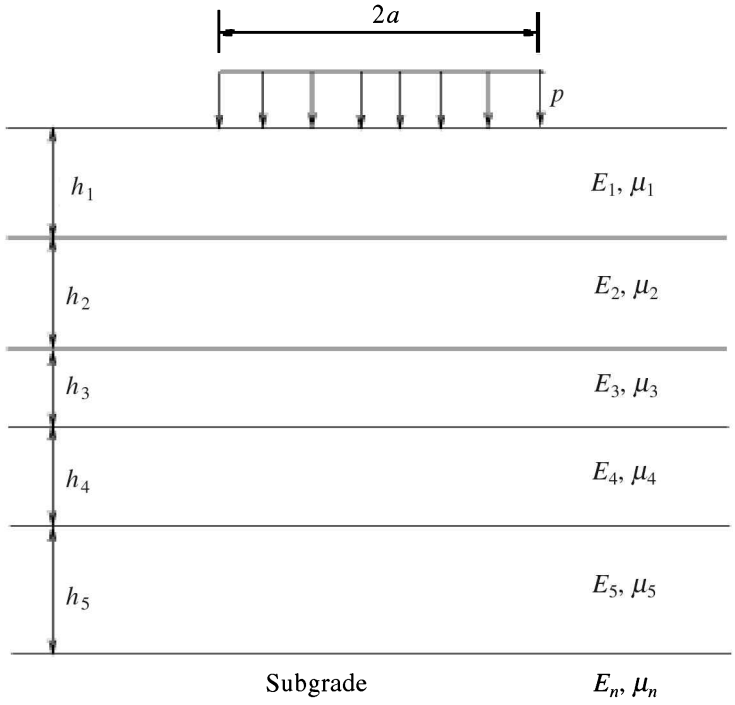


Figure 11.11 A layered pavement structure.

where,

i is the layer index

A_i, B_i, C_i, and D_i are constants for the ith layer

J_0 and J_1 are the Bessel functions of first kind of orders zero and one respectively

r and z are the radius and depth of the point under consideration.

Expression for σ_{z_1} is obtained from Eq. (11.4) as

$$\sigma_{z_1} = -mJ_0(mr)[m^2A_i - mC_i(1 - 2\mu_i - mz_i)]\, e^{-mz_i} + [m^2B_i + mD_i\,(1 - 2\mu_i + mz_i)]e^{-mz_i} \tag{11.18}$$

The values of A_i, B_i, C_i, and D_i are so selected that for $z = 0$, σ_{z_1} is given by

$$\sigma_{z_1} = -mJ_0(mr) \tag{11.19}$$

Thus, the response due to surface loading as given by Eq. (11.19) is obtained. However, in the present case, the loading configuration is as follows:

$$\sigma_{z_1}(r) = f(x) = -p \qquad \text{for } 0 < r \leq a$$
$$= 0 \qquad \text{for } r < a$$

Henkel transform[1] is applied for obtaining the stresses and displacements due to the wheel load acting over a circular area. The final expression for σ_{z_i} is given as

$$\sigma_{z_i} = (-p) \int_0^\infty J_0(m\rho)J_1(m)[[m^2A_i - mC_i(1 - 2\mu_i - m\xi)]\varepsilon^{m\varepsilon} + [m^2B_i + mD_i(1 - 2\mu_i + m\varepsilon)]\varepsilon^{-m\varepsilon}]dm \quad (11.20)$$

where

$\rho = r/a$

$\xi = z/a$

m is any number up to infinity.

Similarly, equations for other stress components and displacement can be obtained. As seen from these equations, it is necessary to evaluate $4 \times n$ integration constants for computing stresses and deflections. Since the stresses and deflections should vanish at infinite depths, $A_n = C_n = 0$. On the surface, $\tau_{rz} = 0$ and $\sigma_z = -mJ_0(mr)$. For rough interface, the two layers are in contact without any vertical separation and complete shear transfer takes place. Therefore, the conditions are

$$\sigma_{z_i} = \sigma_{z_{i+1}}, \qquad \tau_{r_i} = \tau_{rz_{i+1}}, \qquad U_{z_i} = U_{z_{i+1}} \qquad \text{and} \qquad U_{r_i} = U_{r_{i+1}}$$

where U denotes the displacement components.

For a smooth surface, the conditions are

$$\sigma_{z_i} = \sigma_{z_{i+1}}, \qquad \tau_{rz_i} = 0, \qquad \tau_{rz_{i+1}} = 0 \qquad \text{and} \qquad U_{z_i} = U_{z_{i+1}}.$$

The basic formulation for the analysis of the elastic layered structure has been discussed till now. Based on this, a number of computer programs such as CHEVRON, BISAR, ELSYM, WESLEA, FPAVE [60] are available, which can be used to determine the stress–strain displacement of any point within a pavement structure, due to a given loading configuration. Another approach for pavement [refer Eq. (12.18)] analysis is

[1]If a function is defined as $f(m) = \int_0^\infty xf(x)J_0(mx)dx$, then according to Henkel transform [211], $f(x)$ can be obtained by the inverse transform of $f(m)$ as, $f(x) = \int_0^\infty mf(m)J_0(mx)dm$.

known as *Odemark transformed section approach*. This approach comprises converting pavement layers into an approximate equivalent single layer and then analyzing it. Modifications to the layered elastic approach were proposed where nonlinearity of subgrade soil and granular layer, as well as the viscoelasticity of bituminous mix were incorporated. But these are computationally exhaustive and hence not so popular among practising engineers. These situations can possibly be handled by the use of a large number of pavement layers or a pre-analyzed response database.

11.5 ANALYSIS OF CONCRETE PAVEMENT STRUCTURES

11.5.1 Slab on Elastic Foundation

As per the principle of slab on elastic foundation, it is considered that the reaction from the elastic foundation (subgrade in this case, with modulus of subgrade reaction k) is proportional to the deflection w of the plate. The basic equation for the plate, for loading q is given by [242]

$$\left(\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr}\right)\left(\frac{d^2w}{dr^2} + \frac{1}{r}\frac{dw}{dr}\right) = \frac{q - kw}{D} \tag{11.21}$$

where $D = \dfrac{Eh^3}{12(1-\mu^2)}$.

The first slab theory was developed by Hertz, and later it was expanded by Westergaard [267, 261]. Deflection w at any point due to a concentrated load P on the edge of the concrete slab is given by Westergaard as [185]

$$\delta = \frac{2P}{\pi k l^2}\int_0^{\alpha} \frac{\gamma \cos(\alpha x/l)[\cos(\beta y/l) + (1-\mu)\alpha^2 \sin(\beta y/l)]e^{-\gamma y/l}d\alpha}{1 + 4(1-\mu)\alpha^2\gamma^2 - (1-\mu)^2\alpha^4} \tag{11.22}$$

Bending moment, at any point, can be computed from

$$M_x = D\left[\frac{\delta^2 w}{\delta x^2} + \mu\frac{\delta^2 w}{\delta y^2}\right] \tag{11.23}$$

where h is the slab thickness and $l = (D/k)^{0.25}$. α, β, and γ are the parameters with relationships, $\alpha^2 + \beta^2 = \gamma^2$ and $2\beta\gamma = 1$.

Equation (11.23) can be used for calculating the stresses at the edge (using the reciprocity theorem) due to wheel loads of various configurations. Westergaard had done extensive research on analysis of concrete pavements between the years 1926 and 1939 [266]. He calculated stresses in a concrete slab at three locations, namely the edge, the corner, and the interior—the critical stress govern the design. Westergaard assumed full contact with the slab and the subgrade at all loading conditions. Pickett developed a semi-empirical formula where he assumed that the slab loses contact when it is warped outwards, and his formula on corner stresses gave results close to the field observations.

11.5.2 Stresses in Concrete Pavements

The stresses in concrete pavements are due to both temperature and load. In this subsection, the temperature and load stresses are first separately discussed, followed by a discussion on the combined effect of both.

Stresses due to temperature

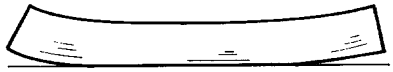
The concrete pavement undergoes temperature changes throughout the day. The temperature at the top surface is maximum during daytime. Similarly during night-time, the bottom of the pavement has the highest temperature. There always exists a temperature gradient across the thickness of the concrete pavement. Studies by inserting thermocouples at different depths, as well as by the theoretical analysis, showed that the temperature gradient is not linear. Two factors are responsible for the development of temperature stress in the concrete pavement slab—the expansion and contraction due to temperature change, and the temperature gradient across the thickness of the slab. During daytime, the slab tends to warp convex upwards [Figure 11.12(b)]. *If, ideally, the slab is weightless and does have any other restraint (like dowelled or tied joints, or interlocking) to move, no temperature stress would develop.* Now, there can be the following two considerations regarding the temperature stress:

(i) If the slab is assumed to have self-weight, it will not be free to take a shape as shown in Figure 11.12(b), rather it will have a shape like the one shown in Figure 11.12(d). *Therefore, it will experience a tensile temperature stress at the bottom and compressive temperature stress at the top during daytime.* Similarly, the development of temperature stress during night-time can be explained from Figures 11.12(a) and (c).

(ii) The edge of the slab, specially the corners may be somewhat free to move, expand, or contract [as shown in Figures 11.12(c) and (d)]. Therefore, if the variation of temperature stress along any direction (longitudinal or transverse) is considered, it will show maximum temperature stress in the interior of the slab, which will gradually decrease outwards. This has been shown schematically in Figure 11.13.

(a) Warping of free and weightless slab during night-time: Temperature stress is zero

(b) Warping of free and weightless slab during daytime: Tempeature stress is zero



(c) Warping of free slab, having self-weight, during night-time: Temperature stress is compressive at the bottom and maximum in the interior

(d) Warping of free slab, having self-weight, during day-time: Temperature stress is tensile at the bottom and maximum in the interior

Figure 11.12 The concept of temperature stress in concrete slab.

For the same reason, as discussed, warping stresses may also be caused due to differential moisture content in the concrete, but these stresses usually are opposite to those caused by temperature warping and hence need not be considered. It can be shown [266] that for a slab, infinite in two directions, the temperature stress is given by

$$\sigma = \frac{E\alpha\Delta T}{2(1-\mu^2)}(1+\mu) = \frac{E\alpha\Delta T}{2(1-\mu)} \tag{11.24}$$

where ΔT is the temperature differential.

For a finite slab, the temperature stress can be expressed as

$$\sigma = \frac{E\alpha\Delta T}{2(1-\mu^2)}(C_x + \mu C_y) \tag{11.25}$$

where C_x and C_y are the correction factors found by Bradbury based on Westergaard's analysis [20]. The above equation is based on the assumption that the temperature gradient is linear and the concrete slab experiences a downward pull by the imaginary springs (Winkler's foundation) when it is curled up.

Stresses due to loading

Due to loading, the maximum stress is induced at the corner, being discontinuous in two directions. The edge stress is lower than that at the corner, being discontinuous in one direction, while the interior stress is the minimum among all these.

Combined stress due to temperature and loading

Figure 11.13 conceptually shows how the corner stress, edge stress, and the interior stress vary due to temperature and load. For example, it is seen that the corner stress is

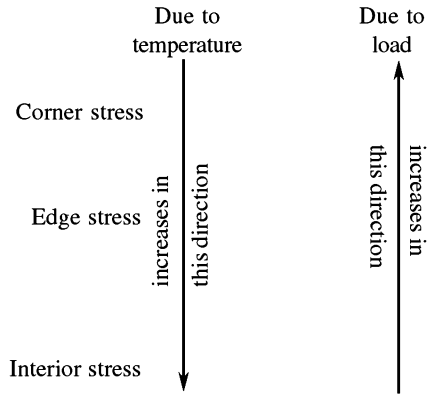


Figure 11.13 Stress levels due to temperature and loading.

minimum due to temperature but it is maximum due to load. Thus for design considerations, the combined effect needs to be considered. Moreover, the bottom stress, due to temperature, during night-time is compressive, and therefore subtractive in nature with reference to the tensile bottom stress due to load. Thus, the temperature stress situations at night-time are not considered while determining the combined stress.

EXERCISES

1. Explain the structural difference between the concrete and bituminous pavements.
2. Explain the various types of joints used in concrete pavements.
3. Two pavement sections (I and II) show the same deflection when subjected to the same standard load, tyre contact pressure, etc. However, the radius of curvature of the deflection bowl in section I is larger than that in section II. Which one of these you think will have less longevity and why?
4. What are the parameters required as the input for pavement analysis?
5. What is non-uniform stress field? What is Airy's stress function?
6. Do you think that a concrete pavement analyzed as a multilayer elastic structure and as a slab on elastic foundation will give the same stress values at a point within the pavement structure? Justify.
7. You have a software for multilayer elastic analysis, which can give stress–strain displacement at any given point for circular uniform loading. How will you validate that the software gives the correct results?
8. Write down the boundary conditions of a rough interface for an elastic multilayered structure.
9. How is the combined effect of temperature stress and load stress considered in concrete pavements?

Pavement Design

12.1 INTRODUCTION

Pavement design deals with the techniques of determining thickness values, and laying configuration for the chosen pavement materials. The design of pavement may consider the construction of a new pavement or rehabilitation of an existing one. The latter is more commonly known as *strengthening by overlay*, which will discussed in detail in Section 14.4. The topic of pavement design considers three types of design aspects, namely the structural design, the functional design, and the drainage design. The structural design is concerned with the structural adequacy of the pavement to sustain traffic load and its repetitions. Unless a pavement is grossly underdesigned, it does not fail suddenly, rather slowly progresses towards failure. The task of a structural pavement designer, therefore, is to ensure that the pavement does not suffer from a premature failure. Functional design emphasizes the surface characteristics of the pavement. The functional parameters of a pavement are roughness, rolling resistance, skid resistance, colour, reflectivity, tyre-pavement interaction, noise, etc. A pavement which may be structurally adequate to sustain load repetitions, may not be functionally serviceable; for example, it is extremely uncomfortable for a driver or passengers to travel over a road having big potholes, ravelling, or corrugations all over its stretch. A pavement with a seriously damaged top surface may not be functionally serviceable, though it may be adjudged structurally sound. On the other hand, a pavement which has failed structurally must also show the characteristics of functional failures, and in that case, any repair work done for restoration of only the functional status would remain effective for just a short duration. Also, a pavement designer must give due consideration to the drainage features of the pavement such that the entry of water is restricted and percolated water or water from capillary rise is quickly drained off from the pavement before causing any distress, and subsequent failure.

There exist a number of design methodologies for structural design of pavements. These pavement design methodologies can be subdivided into two major groups, namely the empirical design methods and the mechanistic design methods. The empirical design methods are evolved through experiences concerning pavement

performance, while the mechanistic design methods try to relate the stress–strain parameters with the expected life of the pavement. The mechanistic method is more theoretical in approach, though it needs calibration based upon the performance of in-service pavements. Due to the inherent complications involved in analyzing a pavement structure and the observed deviations from the predicted response obtained through various pavement models, the empirical methods have not yet lost their importance. Some empirical design methods, like the one suggested by AASHTO [2] are being successfully used to design pavement structures.

This chapter mainly emphasizes the structural design of pavements and has seven sections. In the first section, general design parameters are discussed. Various design philosophies for pavement design have been presented in the next section. The third section deals with the existing methodologies of bituminous pavement design. Similarly, various design methodologies for the latest practices in concrete pavement design have been explained in the next section. The fifth section discusses the drainage provisions while a brief note on frost damage is added in the next section. In the last section, various other advanced concepts of pavement design have been discussed.

12.2 DESIGN PARAMETERS

12.2.1 Material Properties

The properties of various materials used for highway construction have been discussed in detail in Chapter10. Also, various specifications of highway materials and the soil stabilization aspects are discussed in Chapter13. The engineering parameters derived from material characteristics are used in pavement design and are different for different materials. Their respective specifications also vary with materials, for example, the stiffness modulus value of Bituminous Concrete (BC) is different from that of Semi Dense Bituminous Concrete (SDBC), and so on. The elastic modulus, Poisson's ratio, fatigue life, and modulus of rupture are some of the engineering parameters used for the structural design of the pavement. The permeability and void ratio are the parameters used for drainage considerations.

The input parameters are either obtained experimentally or estimated from the formulae or recommendations provided in the design guidelines. In case there are variations in the input parameters, statistically suitable values should be adopted. The following example explains how a design CBR value is estimated out of a number of CBR values available in a given stretch.

EXAMPLE 12.1

The CBR values are calculated after every kilometre on a selected stretch of 10 kilometres having the same type of soil. The values obtained are:

3.8, 2.8, 4.5, 3.9, 4.2, 2.8, 4.7, 4.3, 4.0, and 4.5%.

Find the design CBR value.

Solution

The mean CBR value is 3.95% and the standard deviation obtained is 0.668%. For a 50% reliability in design, the design CBR can be chosen as

$$3.95 - 0.0 \times 0.668 = 3.95\%$$

Similarly, for an 85% reliability in design, the design CBR could be chosen as

$$3.95 - 1.04 \times 0.668 = 3.25\%$$

This procedure assumes that the distribution of CBR values is normal. As an alternative procedure [237], percentile values (percent equal to or greater than) can be plotted against the CBR values and, say, 60 percentile CBR can be chosen for a design reliability of 60%. Design specifications provide a guideline [237, 2] to choose a suitable reliability value for various design situations. If the project stretch is long, it is segmented into a number of statistically homogeneous sections, and a separate pavement design is carried out with the corresponding design CBR values of the individual sections. A discussion on delineation of pavement database into suitable homogeneous sections can be found in Section 14.5.2.

12.2.2 Traffic Characteristics

Axle load

A vehicle can have a number of axles. A standard truck has two axles, namely the front and the rear. The weights of individual axles are called the axle load which may be assumed as approximately half the total weight of a standard truck. In practice, the weights of the rear and the front axles are not equal, and, they depend on the position of the load the vehicle carries. Generally for design purposes, it is the weight of the rear axle of a vehicle which is taken into account.

The axle load of various vehicles are different. The damages caused by them are, therefore, of different magnitudes. If a damage is caused by N_1 number of repetitions for an axle weight of W_1, and the same extent of damage is caused by N_2 number of repetitions by another axle load W_2, then AASHO's experimental data [1] gives the following equation:

$$\frac{N_1}{N_2} = \left(\frac{W_2}{W_1}\right)^4 \tag{12.1}$$

This equation developed by AASHO is a universal equation, used in pavement design in many countries, and much later was explained theoretically. It is popularly known as the *fourth power damage formula* in pavement engineering. This equation is used to convert the number of repetitions of vehicles of various axle loads plying on the road to an equivalent standard axle load repetitions (termed *Equivalent Single Axle Load* or ESAL repetitions in various design guidelines). The standard axle load in India is

8.16 tonnes and 14.968 tonnes for single axle and tandem axle respectively [89]. A standard single axle load of 18,000 lb (or 18 kip) is assumed by other countries [2, 237], and this, in fact, is equal to 8.16 tonnes. Legal axle load is that maximum axle load, any value beyond which is not permitted to move over the road. In India, the legal axle load for the single axle is 10.2 tonnes. The legal axle loads for various axles in India are specified by the Motor Vehicles Act. Table 12.1 gives the legal single axle loads in some countries [251].

Table 12.1 Legal single axle loads in some other countries

Countries	*Legal single axle load*
Japan, Netherland and Sweden	10.00 tonnes
UK	10.17 tonnes
European Community proposed limit	11.50 tonnes
Italy	12.00 tonnes
Belgium, France, Greece, Luxembourg	13.00 tonnes

Standard axle load is that axle load based on which all the calculations related to pavement damage have been standardized. This means that, in design charts, the thickness values are read against equivalent standard axle load repetitions. Therefore, the various axle load repetitions on in-service road need to be converted to equivalent standard axle load repetitions, before proceeding to use a pavement design chart.

Traffic volume

The *traffic volume count* is the total traffic which the pavement to be designed is expected to experience. Generally, the daily traffic count is measured as the average of 7 days 24 hours classified traffic count in accordance with the IRC:9–1972 [244]. When this data is not available, the average of three days traffic count may be used as an approximation [91]. The traffic estimates for new roads can be made on the basis of potential land use and traffic on existing routes in that area [89]. If the average daily traffic count and the traffic growth rate (compound growth rate) are A and $r\%$ respectively, then the volume of traffic in the nth year can be calculated as

$$\text{traffic}_{n\text{th year}} = 365 \times A(1 + r)^n \tag{12.2}$$

Therefore, the total traffic flowing in the selected stretch in n years, can be obtained as the summation of the terms of a geometric progression series. Thus,

$$\text{design traffic} = 365 \times A \times \frac{(1+r)^n - 1}{r} \tag{12.3}$$

While designing a pavement for a given design period, Eq. (12.3) is used to predict the total number of axle load repetitions expected. Here, n is known as the *design period.*

The traffic growth rate can be estimated using extrapolation from the trend of traffic growth or from various econometric models. IRC:37–2001 [89] and IRC:58–1988 [91]

recommend that if adequate data is not available, the traffic growth rate may be taken as 7.5%. Asphalt Institute [237] suggests various growth rates from 4 to 10% depending on the type of the road.

As pointed out earlier, different axle loads have different damaging effects on the pavement. The results of the axle load survey give the distribution of axle load (i.e. axle load and the corresponding traffic volume) in a given road stretch. Thus, a factor needs to be derived from the axle load distribution data which when multiplied by the total number of repetitions, would convert the commercial vehicle repetitions into standard axle load repetitions. The weighted average of the damages caused by the individual axle load group with respect to the corresponding volume of traffic of each group is called the *Vehicle Damage Factor* (VDF), which can be represented by the following equation:

$$\text{VDF} = \frac{V_1\left(\frac{W_1}{W_s}\right)^4 + V_2\left(\frac{W_2}{W_s}\right)^4 + V_3\left(\frac{W_3}{W_s}\right)^4 + \ldots}{V_1 + V_2 + V_3 + \ldots} \tag{12.4}$$

where

W_1, W_2, W_3, $\cdots$ are the median values of the various axle load groups

V_1, V_2, V_3, $\cdots$ are the respective traffic volumes

W_s is the standard axle load.

Here, the fourth power damage formula [Eq. (12.1)] has been used to determine the relative damage of various axle loads, using the standard axle load as the basis. Precisely speaking, it is not only the axle load, but also the axle configuration, terrain, type of road, condition of road, and so on, which determine the extent of damage to a pavement. However, for general purposes, it is the axle load survey data, which is used for evaluation of VDF. IRC:37–2001[89] has adopted the axle load equivalency factors (which are indicative of the relative damages to the pavement caused by them) as recommended in the AASHTO guidelines [2]. These factors are used for the calculation of VDF, for single and tandem axles. A partial list of load equivalency factors adopted by IRC:37–2001[89] is presented in Table 12.2. Generally, the fourth-power damage formula is used to calculate VDF when load equivalency factors are not available.

Table 12.2 Partial list of load equivalency factors from AASHTO [2]

Gross axle weight (tonne)	*Load equivalency factors*	
	single axle	*tandem axle*
6350	0.35	0.024
7260	0.61	0.043
8160	1.00	0.070
9070	1.55	0.110
9980	2.30	0.166
10890	3.27	0.242

According to the Indian code, the axle loads of those vehicles, whose gross weight is more than 3 tonnes, are the only ones used for calculating VDF [89]. This VDF is multiplied by the total number of commercial traffic within the design period obtained from Eq. (12.3) so as to convert it to the number of equivalent standard axle repetitions. The traffic repetitions are generally expressed in *million standard axle load* or 'msa'. Another factor called the *load safety factor* (LSF) is multiplied by the total equivalent standard axle load repetitions, to take care of the variations in traffic axle load measurements and unpredicted heavy truck load. The VDF value may be more than or less than one, but LSF is always greater than one. In some countries, VDF is known as the 'truck factor' [237]. Load safety factor is generally used in design of concrete pavements [239] to take into account the sudden nature of failure in concrete pavements. It is to be noted that the VDF factor becomes redundant if the individual damages of various axle load groups are considered in design. This is the reason, why VDF is not considered, when pavement design is done based on cumulative damage principle [see Eq. (10.39)].

EXAMPLE 12.2

The axle load spectrum from a survey data is presented in the following table. Find the values of VDF.

Axle load range (tonnes)	*Percentage frequency*
17–15	04
15–13	19
13–11	24
11–09	37
09–07	12
07–05	04

Solution

As the axle load equivalency table is not provided, the fourth power damage law is assumed. Taking the median axle load as the representative of the axle load group, VDF is calculated from Eq. (12.4), as follows:

$$\text{VDF} = \frac{4\left(\frac{16}{8.2}\right)^4 + 19\left(\frac{14}{8.2}\right)^4 + 24\left(\frac{12}{8.2}\right)^4 + 37\left(\frac{10}{8.2}\right)^4 + 12\left(\frac{8}{8.2}\right)^4 + 4\left(\frac{6}{8.2}\right)^4}{100}$$

Thus,

$$\text{VDF} = 4.23$$

Lane distribution factor

Traffic survey reports average total traffic count in both the directions. However, for design purposes, the traffic along a particular lane needs to be only considered. Thus, a factor, called the *Lane Distribution Factor* (LDF) is introduced which is multiplied by the total number of commercial vehicles in both the directions to obtain traffic along a single lane.

IRC:37–2001 [89] gives some recommendations while choosing the LDF value, for example, for a single-lane road, the factor is one; for a two-lane road with single carriageway, the factor is 0.75; for a four-lane single carriageway it is 0.40, and so on. In all these cases, the LDF is applied to the total traffic count along both the directions. AASHTO design guidelines [2], however, use two factors in this regard, namely the directional distribution factor and the lane distribution factor. As is obvious from the names, the directional distribution factor finds traffic along a direction while the latter finds out the traffic along a lane. Together, these two factors, are equivalent to the lane distribution factor as recommended in IRC:37–2001 [89].

Lateral distribution of wheel path in a lane

Even on a same lane, all the vehicles do not follow the same line of traverse along a road. If the road section is viewed transversely, there exists a distribution of vehicle wheel path of each lane. Figure 12.1 presents a survey data where the maximum travelled path of the outer wheel was found to be 2.15 m away from the pavement edge.

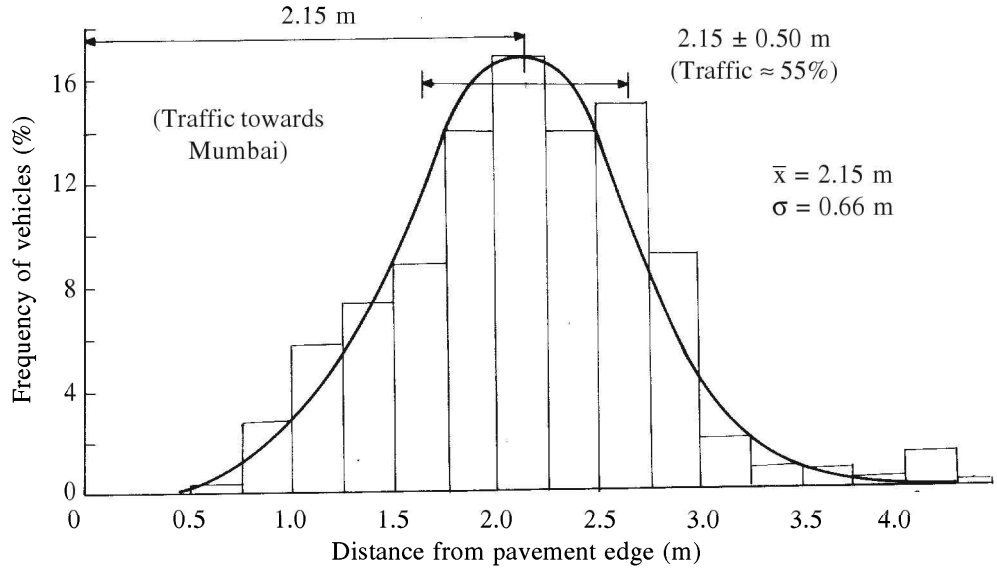


Figure 12.1 Lateral distribution of traffic, data collected on NH–6, in January, 1983, at km 132 [61].

Obviously, the maximum traversed path is distressed to the highest level. Thus, pavement design should take into account the distress along this critical line. However, the other wheel paths, which are not coincident with the maximum traversed line, also have their contributions in terms of the distress along the critical line. Thus, another conversion factor is needed, that would convert all the vehicle repetitions along their respective traverse line to the equivalent repetitions along the critical line. In other words, the *Lateral Distribution Factor* is a factor which is to be multiplied by the total traffic repetitions in a lane to convert it to equivalent repetitions along the maximum distressed (or traversed) path. It is possible to find out analytically the value of lateral distribution factor if the lateral traffic distribution is available (as is the case above), which discussion, however, is beyond the scope of the this book.

Wheel configuration

Figure 12.2 shows the axle and wheel configuration of a standard truck, derived from the survey on trucks plying on Indian roads [61]. The stress created at a point inside the pavement due to traffic load is the sum of the effect of individual wheels taken together. Generally, the linear superposition principle is used to calculate the overall contribution of the individual wheels. It is found that the length of the axle is such that the effect of

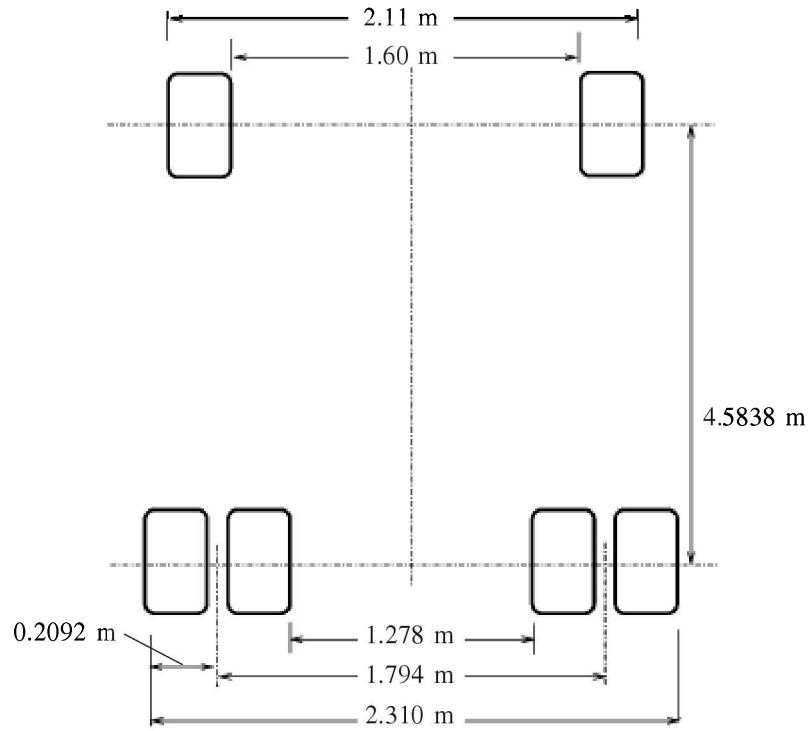


Figure 12.2 Axle and wheel configuration of a standard Indian truck [61].

the wheels on one end is not felt appreciably at the other end. Therefore, for practical purposes of pavement design, generally, only two wheels are considered and the system is called the *dual-wheel system*. Figure 12.3 shows a dual-wheel system used for the design of pavements as per Indian recommendations [89].

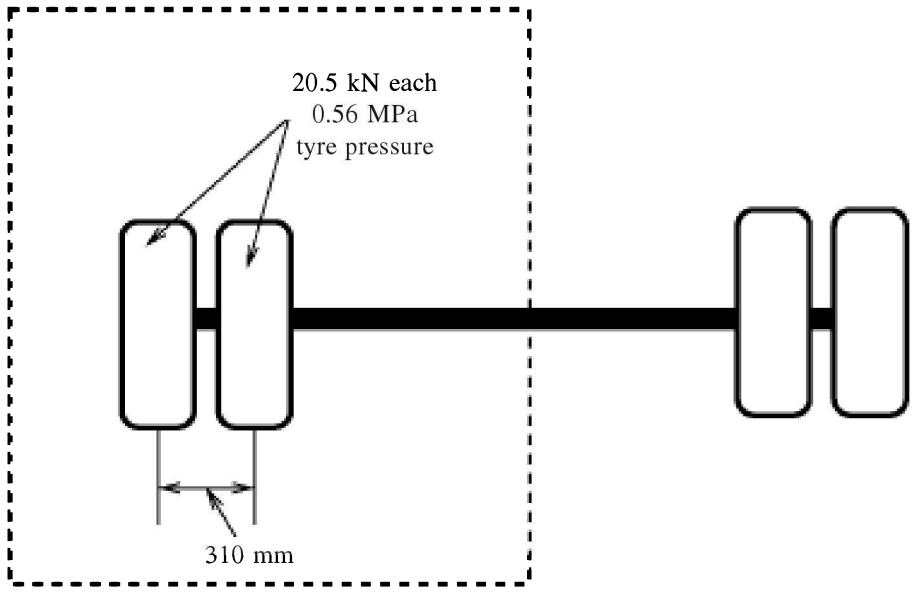


Figure 12.3 A dual-wheel system of the rear axle of a truck used in pavement design.

Tyre contact pressure

The tyre contact pressure is the pressure exerted by the tyre on the ground. The tyre contact pressure could be more than, equal to, or less than the tyre pressure but generally it is assumed to be equal to the tyre pressure and uniform in distribution. A survey on tyre pressures of commercial vehicles in India indicated that the pressure ranged from 0.77 to 0.84 MPa [61]. Another study showed [252] that 65% of the tyres had pressures ranging from 0.70 to 0.90 MPa, while 13% of the trucks had tyre pressures as high as 0.95 MPa [252].

Tyre imprint area

For ease in computation, tyres are approximated to various idealized shapes. As an example, for multilayered elastic analysis, tyre imprint is generally approximated to circular in shape and for concrete pavement design by PCA [239] method, it is approximated to rectangular in shape.

EXAMPLE 12.3

Two closely separated wheels of load 20.5 kN each and tyre pressure 0.7 MPa are acting on a pavement section. If the two wheels are replaced by a single wheel with the same tyre pressure, calculate the radius of the tyre imprint (idealized as circle) of the single wheel.

Solution

The total load acting on pavement is

$$2 \times 20.5 \text{ kN} = 41 \text{ kN}$$

and

$$\text{tyre pressure} = 0.7 \text{ MPa}$$

If the radius of the equivalent single wheel is r, then

$$\pi r^2(0.7 \times 10^6) = 41 \times 10^3$$

or

$$r = 0.136 \text{ m}$$

EXAMPLE 12.4

The radius of a tyre imprint is approximated to a circle of 150 mm. What is the maximum loading duration on a particular point of pavement by a truck moving at a speed of 60 kmph?

Solution

Duration is the time required to cross the diameter of the tyre imprint. Therefore, the time required is

$$\frac{2 \times 150}{60 \times 10^6/3600} = 0.018 \text{ s}$$

Closing remarks

Various load parameters used in pavement design have been discussed so far. The traffic parameters used in pavement design vary from zone to zone. Also, the design thickness depends on the axle load, tyre pressure, and the wheel configuration, which are taken as standard. However, these standard values are different in different codes of practices. Table 12.3 shows such variation in these values in the four codes of practices, all of which are based on mechanistic pavement design principles.

Table 12.3 Standard value of single-wheel load, centre-to-centre distance between dual wheels and tyre pressure used in various guidelines

Methodology/guideline	*Single-wheel load* (kN)	*Centre-to-centre distance* (mm)	*Tyre pressure* (kPa)
IRC:37–2001 [89]	20.5	310	560
Shell [206]	20.0	315	600
Austoards [182]	20.5	330	550–700
South African Mechanistic Design Method [236]	20.0	350	520

12.2.3 Environmental Characteristics

Temperature

As discussed earlier, the stiffness modulus of bituminous mixes changes with temperature (refer Table 10.7), and the variation in temperature and the temperature differential induce temperature stresses in the concrete pavements (see Section 11.5.2). Temperature variation is responsible for the expansion and contraction of pavement joints. The phenomenon of freezing and thawing at subgrade level is also caused by temperature.

The temperature of an in-service pavement varies during the day and also seasonally. For design purposes, generally the Average Annual Pavement Temperature (AAPT) is used as a parameter. Some empirical as well as theoretical relationships between AAPT and Annual Average Air Temperature (AAAT) are suggested by the researchers in other countries [262, 42, 204, 24]. The bituminous pavement design charts in India are calibrated as per constant AAPT of 35°C [89] throughout the design period. However, pavement design which takes into account temperatures prevalent in various seasons as well as various regions, would definitely give more reliable results compared to the pavement design that assumes a single value AAPT throughout the whole design period. This provision is incorporated in the Asphalt Institute [237] manual for design of bituminous pavements.

For concrete pavement design, the temperature differential values are recommended in the Indian guidelines [91]. Table 12.4 is a partial table showing the same.

Table 12.4 A partial table for recommended temperature differentials in concrete roads in India [91]

Zone	*States*	*Temperature differential in* °C *in slabs of thickness* (cm)				
		10	15	20	25	30
I	Punjab, UP, Rajasthan, Gujarat, Haryana and North M.P., excluding hilly regions and coastal areas	10.2	12.5	13.1	14.3	15.8

Subgrade moisture

Subgrade moisture affects the subgrade modulus. For evaluation of CBR value of subgrade soil, proper consideration for the moisture content is necessary. As per the present Indian recommendations, the CBR sample is prepared at OMC, and soaked for four days prior to testing in order to take into account the most severe moisture condition encountered in the subgrade soil. According to IRC:37–1984 [88], for strengthening of the existing pavement, the CBR sample should be prepared at the field moisture content immediately after the monsoon period. In the Australian guidelines [182], the most probable moisture content is estimated and the subgrade strength is determined at that moisture content.

Frost action

Frost action can be divided into two phases, namely (i) freezing and (ii) thawing of the soil water. The freezing phase may cause heaving of the road surface due to increase in volume on account of formation of ice lenses. The thawing phase causes softening of the subgrade as ice melts during the spring-break. Often, thawing affects the pavement more seriously than heaving.

For ice lens formation and subsequent frost heaving, the temperature should be below the freezing temperature for a sufficient period of time, and there should be supply of water through capillary action for the ice lenses to grow. The ice lenses have strong attraction towards water and water solidifies when it comes in contact with them. Thus, ice grows in size, until there is a capillary cut-off from the water source, or the ice lens touches the level where the temperature is just above the freezing temperature.

Some types of soil, such as silty clay, are susceptible to frost action. It is seen that 0.02 mm is the critical particle size which is susceptible to frost [65, 2]. If soil has 10% fraction of particles having size 0.02 mm or less, it is expected to be frost susceptible, whereas if the fraction is less than 1%, the soil is least expected to be frost susceptible [65].

The *freezing index* is a term used to measure the severity of frost damage and is expressed in degree days. It is defined as the difference of degree days between the highest and the lowest points on a curve of cumulative degree days plotted against time (days) in a given season (i.e. from the beginning of freezing to the beginning of thawing). If the average daily temperature for five days is5.2, –4.3, –2.9, –3.5 and –3.8°C, it can be expressed as, –19.7 degree days. Degree days are calculated with reference to freezing point of water, which is 0°C in this case. The following example shows how to calculate the freezing index.

EXAMPLE 12.5

The degree days of individual months (say, October to March) are +10, +82, –114, –213, –171, and +27. Calculate the freezing index.

Solution

The cumulative degree days are plotted in Figure 12.4.

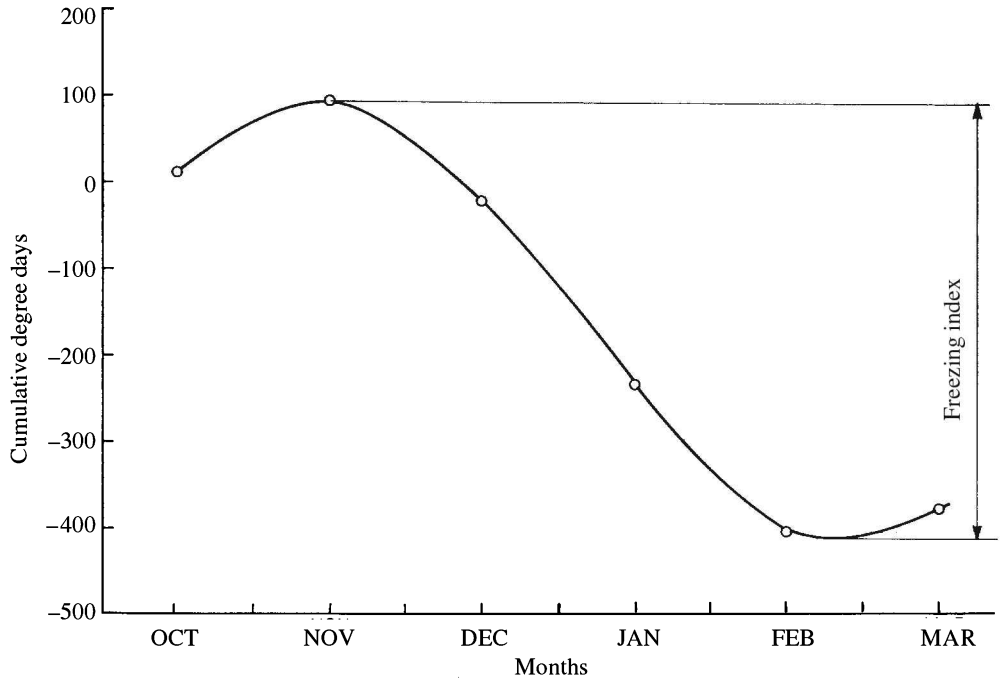


Figure 12.4 Plot of the cumulative degree days of the example problem.

From the graph, the difference between the maximum and minimum degree days is

$$[92 - (-406)] = 498$$

Thus, the freezing index is 498 degree days.

It may be noted that the value of the freezing index should not change, no matter how many months have been considered in the calculation—though the cumulative degree day values will change if, say September, is also included in the above example. However, the freezing index being the difference between the maximum and minimum values, its value will remain unchanged, provided the maximum and the minimum values have been included.

12.2.4 Design Life

Design life is the number of years for which the pavement is designed. A pavement is expected to give satisfactory service over the period of its design life. The recommended design life is 15 years for National Highways and State Highways, and 20 years for

Expressways as per IRC: 37–2001 [89], for bituminous pavements. For concrete pavements 20–40 years may be assumed as the design period [92].

When the cost of pavement construction project comes out to be higher than the available funds, the pavement is designed for a reduced design life, or for stage construction (where the construction of the pavement is done in stages). Stage construction has been discussed at the end of this chapter. If a pavement loses its service life, prematurely or naturally towards the end of its design life, it may be strengthened by putting another layer over the existing pavement. This is called *overlay*. The design principle of the overlay has been discussed in Section 14.5.2.

12.3 PHILOSOPHIES OF DESIGN

12.3.1 CBR Method

The CBR method is one of the earlier empirical methods for pavement design developed during 1928–29 [247]. This method involves determination of the CBR value of subgrade for the most critical moisture condition. The design thickness of pavement is read against the CBR value of subgrade from the design charts which were developed from experience on pavement performance after noting and analyzing a number of failed pavement sections and the corresponding subgrade CBR values. The failure of a pavement was defined as lateral displacement of subgrade soil, differential settlement of pavement, and excessive deflection. Based on this apparent relationship between the CBR value and the thickness of pavement, the U.S. Corps of Engineers in 1940 adopted the CBR method of design for airfield pavements [247]. The same basic equation is still being used for the design of airport runways [106]. It may be noted that this attempt at designing pavement did not initially involve the number of load repetitions that a pavement can sustain and hence later, correction factors related to the number of load repetitions were introduced.

12.3.2 California (Hveem) Method

This method was developed in 1940 based on experience, theory, and test road results, and is still used widely in some countries [65]. The pavement design by this method involves three major considerations, namely:

(a) Traffic load
(b) Strength of subgrade material
(c) Strength of the construction material

The traffic index (TI) is calculated as follows:

$$TI = 9.0\left(\frac{ESAL}{10^6}\right)^{0.119} \tag{12.5}$$

where ESAL is the equivalent single axle load repetition.

The subgrade strength, R (Hveem resistance value), is obtained from the Hveem stabilometer test (see Section 10.5.2). The total thickness of the pavement, in terms of the gravel equivalent GE, is obtained as:

$$\text{GE} = 0.0032 \times \text{TI} \times (100 - R) \tag{12.6}$$

The unit of gravel equivalent is foot. Different types of layer compositions carry various values of gravel equivalent. Thus, after getting the total design gravel equivalent of the pavement, a suitable pavement layer composition can be arrived at.

12.3.3 Limiting Shear Failure Method

In this method, it was proposed that the layer thicknesses could be designed using the bearing capacity approach, in which the stress developed in respective layers must be less than the corresponding bearing capacity of the individual layers. This approach was first proposed by Barber in 1946, where he applied Terzaghi's bearing capacity approach for pavement design [266]. This is also one of the considerations in the South African pavement design method [52].

12.3.4 Limiting Deflection Method

This method used limiting deflection as a criterion for pavement design. Deflection is easy to measure, however, failure theories indicate that failures are due to excess stresses and strains, not due to the deflection. Therefore, deflection solely may not be considered as a parameter for pavement design [266]. However, for determination of thickness of strengthening for an in-service pavement, the deflection method is still used successfully (refer Section 14.4.1).

12.3.5 Regression Method Based on Pavement Performance

In the regression method, various input parameters for pavement design (as discussed in Section 12.2) and the design thickness are related by regression equations. These equations can be developed through experience of performance of various in-service pavements, or from full-scale pavement testing (discussed in Section 12.3.6). One of the successful examples of the use of regression method in pavement design is the AASHTO method (see Sections 12.4.6 and 12.5.3). The drawback of the regression method is that this design methodology is applicable to that area in particular, wherefrom performance data have been collected, hence, may not be applicable to other areas with different climatic, traffic, and material properties [266].

12.3.6 Mechanistic Empirical Method for Bituminous Pavement Design

Mechanistic empirical pavement design is popularly being used in varius countries or organizations. In India too, the pavement design guidelines IRC:37–1984 [88] have recently been updated to IRC:37–2001 [89], where the design methodology has changed from empiricism to mechanistic pavement design principles.

Mechanistic pavement design procedure owes its genesis to the landmark work reported by Dorman [50] way back in 1962. He postulated two failure criteria for the design of bituminous pavements, which even today are regarded as the basis of mechanistic empirical pavement design. Dorman's work was based upon two other works done by Pell in conjunction with Shell Laboratory [21], in 1960, and Monismith, et al. [164] in 1961, individually yet almost simultaneously. It was then named rational pavement design which is now known as mechanistic empirical pavement design.

Figure 12.5 shows a layered bituminous pavement structure subjected to a set of standard dual wheel load system. The horizontal tensile strain ε_t at the bottom of the bituminous layer and the vertical compressive strain ε_z on the subgrade are identified as the critical parameters for fatigue and rutting failures respectively. The concepts of fatigue and rutting are described in the following sections.

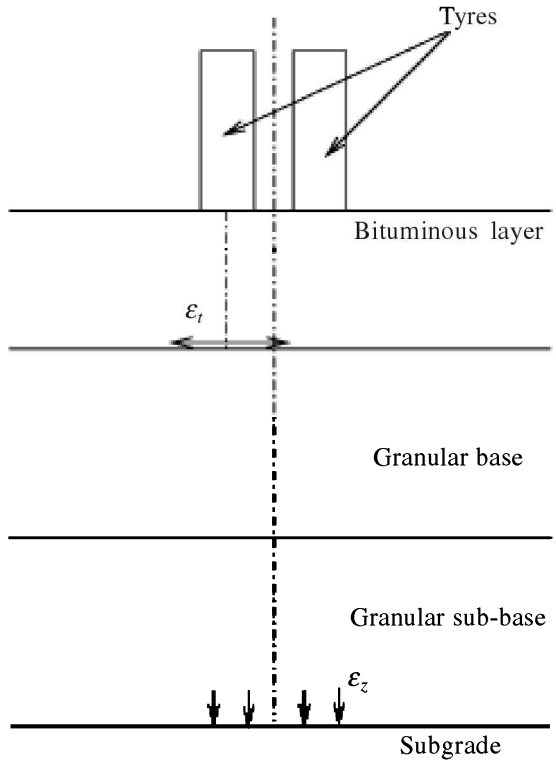


Figure 12.5 Critical strains in a bituminous pavement structure.

Fatigue failure

The bituminous layer undergoes fatigue due to repetitive load application by the traffic. Cracks initiate at the bottom (see Figure 12.6) and finally propagate upwards to the top. A stage is reached when the layer fails completely due to fatigue. The fatigue characteristics of the bituminous mix can be tested in the laboratory (see Section 10.5.3) and the fatigue life of the particular bituminous mix can be found out. As already mentioned, the initial tensile strain at the bottom fibre of the bituminous mix is assumed (see Figure 12.5) to be the index of fatigue life of the bituminous layer, based on which the relationship between the laboratory and field fatigue performance is established.

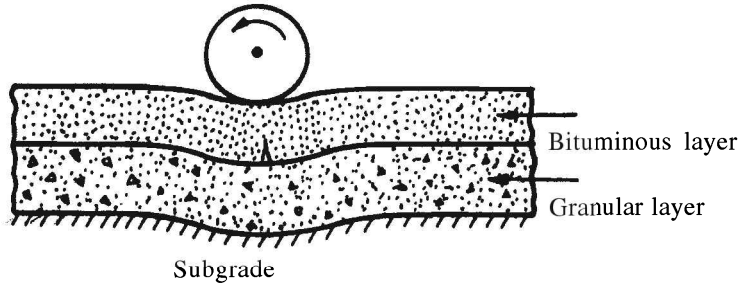


Figure 12.6 Initiation of fatigue cracks in bituminous layer due to repetitive traffic loading.

Rutting failure

Rutting is the permanent deformation of pavement along its wheel path. It is a manifestation of two different phenomena: (i) densification and (ii) shear deformation of the various layers [269]. Figure 12.7 explains the concepts involved in rutting. Figure 12.7(a) shows pavement layers immediately after construction. After a number of traffic repetitions, the initial thicknesses h_1, h_2, and h_3 reduce to h'_1, h'_2, and h'_3 due to compaction. The better the compaction during the construction phase, the less will be the reduction in thickness due to traffic. It is also seen in Figure 12.7(b) that individual layers have undergone some permanent deformation. These two phenomena constitute rutting of pavement.

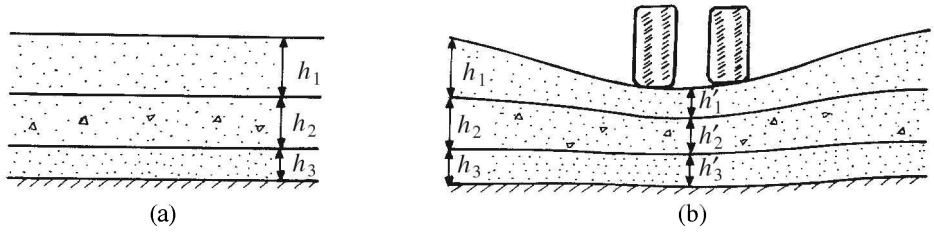


Figure 12.7 Schematic diagram showing the concept of rutting.

The contribution to rutting from various layers could be different. During AASHO road test [1], it was found that 46% of the rutting took place in the bituminous surface

and granular base course, while the sub-base and subgrade contributed 54% of the total rutting [1]. TRRL studies indicated that 54% of permanent deformation was contributed by the surface and base layers while the rest was from the sub-base and the subgrade. The vertical strain on subgrade is assumed as the index of rutting to occur in a pavement. Figure 12.8 shows the variation of vertical strain, due to dual wheel standard loading, along the depth of an arbitrarily chosen pavement section (the values are obtained from elastic multilayer analysis), as per Indian practice [89]. Maximum vertical strain can occur either below one of the wheels or between both the wheels, thus both these are plotted in Figure 12.8. It is seen that the vertical strain shows the maximum value on subgrade and between both the wheels. Generally, this trend is observed for all pavement sections with practical dimensions and loading configuration. Thus, vertical subgrade strain between the dual wheels (see Figure 12.5) is assumed as the index to the rutting failure of the pavement.

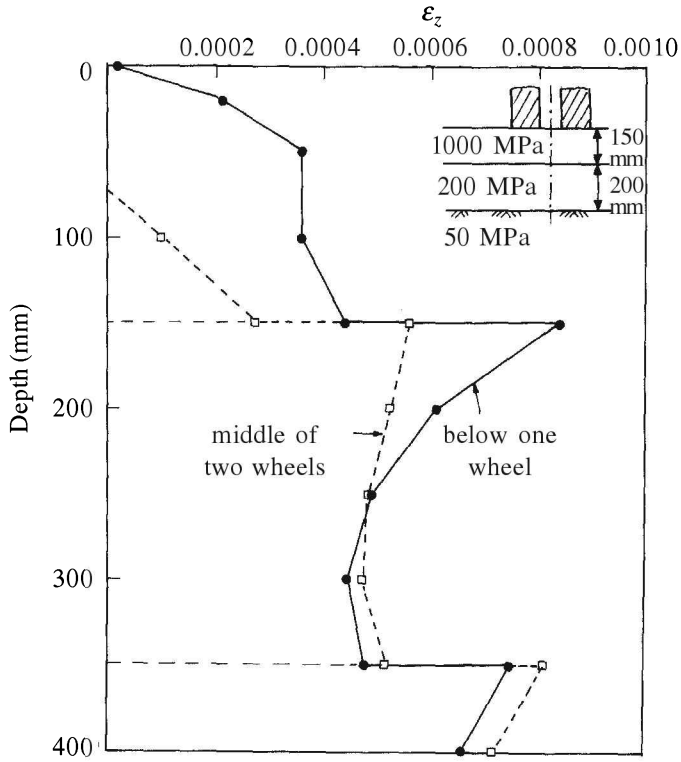


Figure 12.8 Variation of vertical strain along the depth of pavement.

Field performance

Mechanistic pavement design methodology methodology requires calibration using the field performance. Generally, the field failure due to fatigue is estimated in terms of the percentage fatigue cracks on the surface. If the percentage fatigue cracks exceed a predefined value, the pavement is deemed to have failed due to fatigue. The initial tensile

strain of the pavement section is measured or calculated, and then monitored till failure. The corresponding number of equivalent single axle load repetitions (ESAL) are obtained from the traffic survey data. Similarly, the relationship between the initial tensile strain, number of repetitions, and fatigue life for that particular bituminous mix can be obtained from laboratory fatigue testing (see Section 10.5.3). The fatigue life of the bituminous mix determined from the laboratory fatigue test is usually lower than that observed in the field. Thus, a *shift factor* is used to convert the laboratory fatigue life to the field fatigue life. The estimation of shift factor is, therefore, the purpose of the field calibration. The laboratory and field fatigue life do not match each other (the laboratory life is lower than that of field) because of the following reasons [246]:

(i) Boundary conditions of the bituminous mix sample for fatigue testing and bituminous surfacing laid on the in–service pavement are different.

(ii) In the field, random rest periods between load applications allow healing of the cracks in the pavement material. In the laboratory, cyclic loading is generally applied continuously with no or small and equal rest periods [246] which is not the case with the field.

(iii) Residual stresses may remain in the bituminous surface layer even after passage of each load. These stresses relax with time and after sufficient lapse of time, the residual stress does not remain [265]. In the laboratory, the magnitude and duration of residual stresses developed in fatigue samples are much different from those of the field. Random rest periods occurring in the field are difficult to simulate in the laboratory [246].

(iv) The lateral wander of traffic is also an important consideration. The wheel paths of different vehicles are not the same. Therefore, all the wheels of the vehicle do not stress the same point repeatedly.

Asphalt Institute's DAMA [42] program has recommended a rutting shift factor of 20. Finn et al. [63] suggested a shift factor of 13.03, whereas the British experiences [23] indicate a shift factor in the range 100–750. Brunton et al. [24] recommended a factor of 440 for fatigue failure. Figure 12.9 shows the field calibrated fatigue line, based on road performance study in India, which has been obtained by parallely shifting the laboratory fatigue line.

The field calibrated fatigue equation is expressed in the same form as Eq. (10.38), except the use of shift factor SF as follows:

$$N_f = \text{SF} \times k_1 \left(\frac{1}{\varepsilon_t}\right)^{k_2} \left(\frac{1}{E_d}\right)^{k_3} \tag{12.7}$$

In a similar fashion, rutting studies are also carried out in the field and in the laboratory. In the laboratory, the wheel tracking test is performed where repetitive load is applied to the sample and its densification and permanent deformation are measured. In the field, rutting is estimated by putting a straight edge of specified length on the

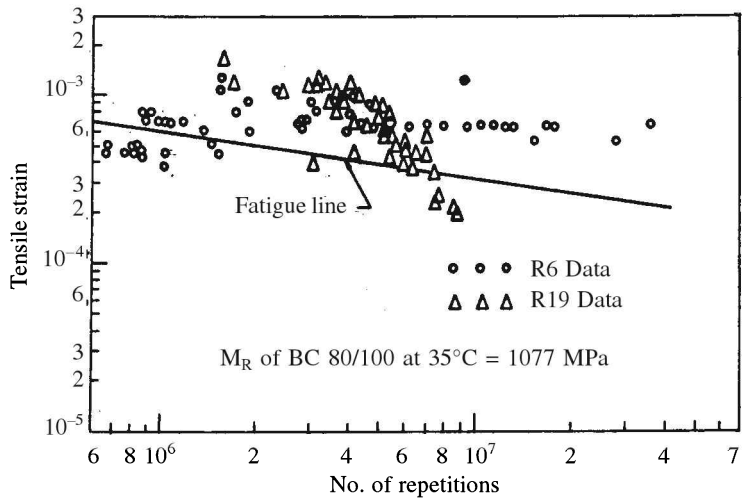


Figure 12.9 Field calibrated fatigue curve [60].

pavement, and then measuring the maximum depression along the most traversed wheel path. Acceptability level of rut depth is different in different countries. TRRL recommends 10 mm as the allowable rut depth [15] and 20 mm as the failure rut depth [24]. France [167] allows 15 mm and in South Africa the limiting rut depth is 20 mm [52]. As per Indian recommendations, a pavement having 20% surface fatigue cracking or 20 mm rutting measured by a 3 m straight edge is assumed to have failed [89]. These criteria have been used to calibrate the fatigue and rutting equations for the design of Indian roads [89]. Equations (12.8) and (12.9) represent the field calibrated fatigue and rutting equations recommended in the IRC:37–2001 guidelines [89].

Data collected from in-service roads due to variations in traffic flow, mixed flow, and other constraints may not give a very precise performance data. For this reason, instrumented test tracks, dedicated to pavement performance studies, are made. AASHO [1], Maryland, WASHO, Arlington, and many other road tests were conducted in the same way. Figure 12.10 shows such an example of a test track recently constructed by National Center for Asphalt Technology, Auburn University, USA [7].

Pavement design curves

The design curves by the mechanistic pavement design method can be developed with the help of the field calibrated fatigue and rutting equations and the multi-layered pavement analysis program. For example, first, some tentative granular and bituminous layer thicknesses are assumed and the fatigue and rutting strains at critical locations are found out. Keeping the granular layer thickness fixed (say), the bituminous layer thickness is varied, and the design thickness of the bituminous layer for pavement, being individually safe from fatigue and rutting failure, is evaluated. That means, for a given granular layer thickness, two corresponding thicknesses of bituminous layers are

Figure 12.10 NCAT test track [7].

obtained. Similarly, for various granular layer thicknesses, other points are found out and plotted in the form of a design chart as shown in Figure 12.11.

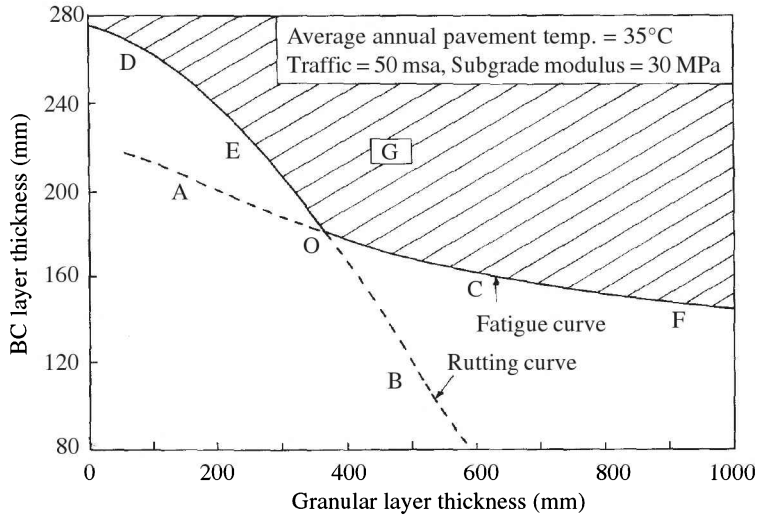


Figure 12.11 A typical bituminous pavement design chart obtained by mechanistic pavement design method.

A bituminous pavement design chart (refer Figure 12.11) is constituted with fatigue and rutting curves. Any thickness combination, which is within the shaded zone, is over-safe from the point of view of fatigue and rutting. The points lying exactly on the line DEOCF are just safe. Different aspects of this pavement design chart can be described as follows:

(i) Point A: The fatigue strain developed is equal to the allowable fatigue strain but the rutting strain developed is more than that allowed. Therefore, for this combination of thicknesses, the pavement is safe against fatigue but not against rutting failure.

(ii) Point B: The rutting strain developed is equal to the allowable rutting strain but the fatigue strain is more than that allowable. Therefore, this pavement section is safe with respect to rutting alone and not against fatigue failure.

(iii) Point C: This point lies in the feasible zone because it is safe with respect to rutting and fatigue failures, but is over-designed with respect to rutting failure.

(iv) Point G: This point is over-safe both from the rutting and fatigue points of view.

(v) Point O: Both the rutting and fatigue strains developed are equal to the allowable rutting and fatigue strains respectively, and after the expiry of the design period, the pavement is expected to fail simultaneously due to fatigue and rutting. Thus, this point may be called *structurally optimum point* where the materials are utilized to the fullest extent.

Closing remarks

Figure 12.12 presents a basic flowchart of the mechanistic pavement design. The material parameters are fed into the analysis program to obtain the developed strains. The allowable strains are obtained from the field performance equations, for a given design life.

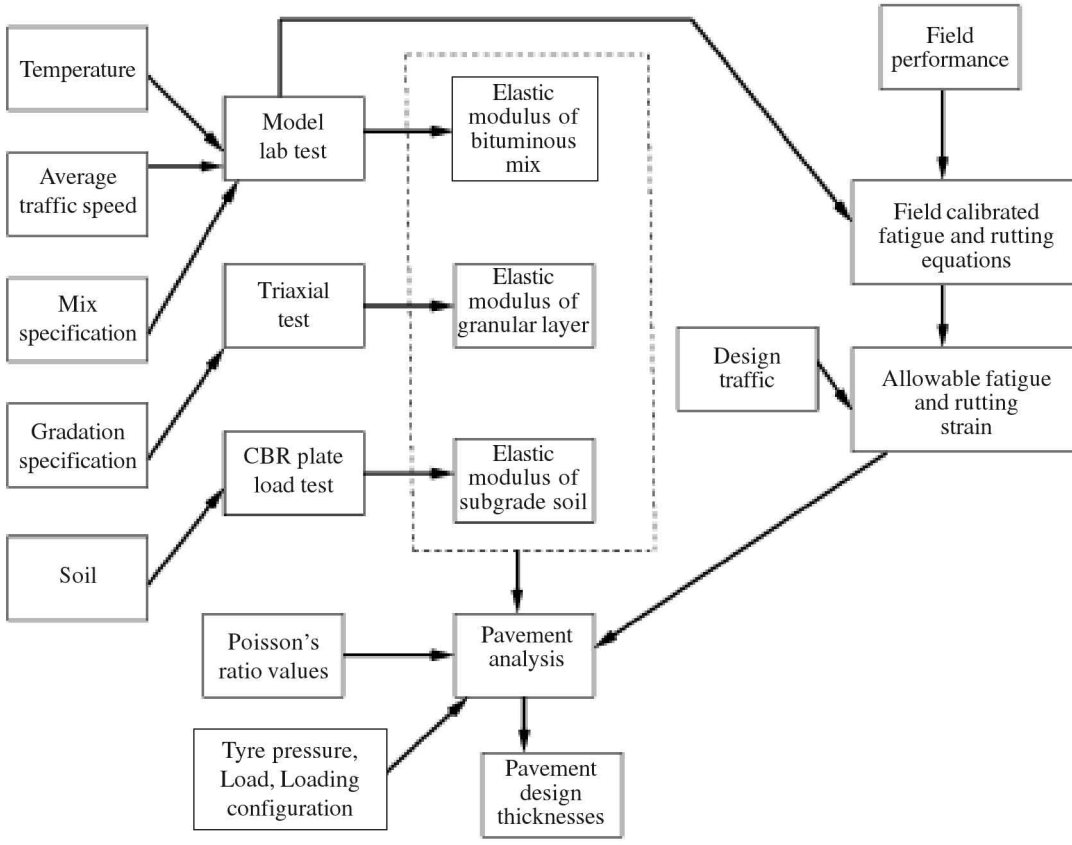


Figure 12.12 Flowchart of mechanistic pavement design.

The thicknesses are optimized in such a way that the developed strains are nearly equal to the allowable strain. Though the design parameters, material properties, traffic, and environmental conditions may vary from country to country, the basic principle of mechanistic pavement design is universal in nature. Thus, the principle of pavement design followed in Shell [38, 251], Austroads [182], Asphalt Institute [237], South Africa [236, 52], and India [89] etc. is somewhat similar.

EXAMPLE 12.6

A cross-section of bituminous pavement is depicted in Figure 12.13.
The pavement section is analyzed by multilayered elastic analysis and the results are tabulated

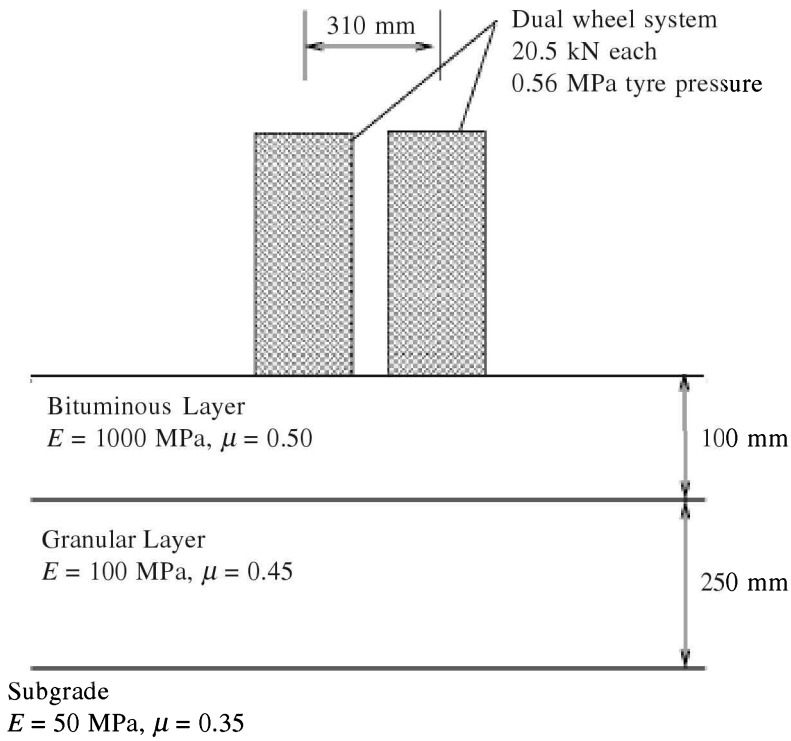


Figure 12.13 Cross-section of a pavement in the example problem.

as follows.

Depth	*Radial*	*Vertical*	*Tensile*	*Radial*

of the point, Z (mm)	*distance, R* (mm)	*strain,* ε_z	*strain,* ε_t	*strain,* ε_r
0.00	0.00	0.8686E–03	–0.5318E–03	–0.3368E–03
50.00	0.00	–0.1891E–03	0.1053E–03	0.8383E–04
100.00	0.00	–0.1116E–02	0.6723E–03	0.4437E–03
100.00L*	0.00	–0.1409E–02	0.6723E–03	0.4437E–03
150.00	0.00	–0.1182E–02	0.6090E–03	0.3773E–03
200.00	0.00	–0.9932E–03	0.5344E–03	0.3279E–03
250.00	0.00	–0.8651E–03	0.4809E–03	0.3028E–03
300.00	0.00	–0.8016E–03	0.4606E–03	0.3018E–03
350.00	0.00	–0.8117E–03	0.4882E–03	0.3252E–03
350.00L	0.00	–0.1129E–02	0.4879E–03	0.3255E–03
400.00	0.00	–0.9710E–03	0.4135E–03	0.2901E–03
450.00	0.00	–0.8447E–03	0.3561E–03	0.2591E–03
0.00	155.00	0.6630E–03	–0.5426E–03	–0.1204E–03
50.00	155.00	0.8755E–04	0.8214E–04	–0.1697E–03
100.00	155.00	–0.5980E–03	0.6456E–03	–0.4766E–04
100.00L	155.00	–0.8694E–03	0.6456E–03	–0.4766E–04
150.00	155.00	–0.9743E–03	0.6255E–03	0.1497E–03
200.00	155.00	–0.9578E–03	0.5690E–03	0.2503E–03
250.00	155.00	–0.9079E–03	0.5194E–03	0.3013E–03
300.00	155.00	–0.8771E–03	0.4988E–03	0.3373E–03
350.00	155.00	–0.9020E–03	0.5283E–03	0.3778E–03
350.00L	155.00	–0.1247E–02	0.5283E–03	0.3778E–03
400.00	155.00	–0.1075E–02	0.4465E–03	0.3403E–03
450.00	155.00	–0.9316E–03	0.3816E–03	0.3047E–03

* response at the lower interface.
negative value indicates compressive strain.

The field calibrated fatigue and rutting equations are as under [89]:

Fatigue equation

$$N_f = 2.21 \times 10^{-4} \left(\frac{1}{E}\right)^{0.854} \left(\frac{1}{\varepsilon_t}\right)^{3.89} \tag{12.8}$$

where

N_f is the number of repetitions (of standard axle) causing fatigue failure

E is the elastic modulus of the bituminous layer

ε_t is the horizontal tensile strain at the bottom of the bituminous layer.

Rutting equation

$$N_r = 4.1656 \times 10^{-8} \left(\frac{1}{\varepsilon_z} \right)^{4.5337} \tag{12.9}$$

where

N_r is the number of repetitions (of standard axle) causing rutting failure,

ε_z is vertical subgrade strain.

If the average elastic modulus of bituminous mix is assumed to be 1000 MPa, how many standard axle repetitions will the pavement survive?

Solution

From the given table, the critical fatigue strain is identified as 0.6723×10^{-3} and the critical rutting strain is identified as 0.1247×10^{-2}.

The fatigue life is calculated from Eq. (12.8) to be 1.328×10^6 standard axle repetitions. The rutting life is calculated from Eq. (12.9) to be 0.611×10^6 standard axle repetitions.

The minimum of the above two is to be chosen. Therefore, the pavement will survive 0.611 msa standard axle repetitions, and will first fail due to rutting.

12.4 PRESENT TRENDS IN BITUMINOUS PAVEMENT DESIGN

The pavement design practice in vogue in different countries can be classified into two major categories, namely (i) the empirical or semiempirical design procedure and (ii) the mechanistic design procedure. The former methods are based solely on the study of performance of in-service pavements, or pavements constructed specially for monitoring the pavement behaviour. Design methods proposed by AASHTO [2], Road Note 29 [245], Japan Roads Association [151] belong to this category. The methods given by Asphalt Institute [237, 208], Shell [206, 38], Austroads [182], South African [236, 52], and the present Indian practices [89] can be put in the second category in which a pavement is analyzed as a multilayered structure, stresses and strains at various critical points are found out, and the pavement performance is evaluated in terms of the critical strains in pavements. The special features of the various design methods are reviewed in the following sections.

12.4.1 Shell Method

Shell [38] pavement design method is based on the concepts of the mechanistic pavement design. On the basis of fatigue and rutting criteria, the thickness design curves are evolved with bound surfacing layer and unbound base layer for various subgrade moduli, asphalt mix type, traffic or design period, and temperature. The pavement

structure is regarded as a linear elastic multilayered system and the layers are assumed to be homogeneous, isotropic, and horizontally infinite in dimensions. The traffic is expressed in terms of repetitions of standard axle loads. The adopted standard design single axle load is 80 kN. Each wheel of the dual wheel system carries a load of 20 kN at a tyre pressure of 0.6 MPa and with the radius of the contact area 105 mm [38].

Shell recommends the laboratory triaxial test for the determination of elastic modulus of the subgrade soil. In case when it is not possible to carry out such a test, the CBR value at the field moisture content and field density should be used to estimate the elastic modulus empirically [refer Eqs. (10.5) and (10.6)]. Elastic modulus of the unbound granular layer is computed from the subgrade modulus and the layer thickness from Eq. (10.15).

The design charts are based on the output of computer program 'BISAR'. In a standard design, the pavement is assumed to consist of three layers, namely a bituminous top layer, a bound or unbound granular layer, and the subgrade. The design criteria using the Shell method are as follows [206]:

(a) Large horizontal tensile strain at the bottom of the asphalt layer causes cracks in the bituminous layer.

(b) Large compressive strain at the top of the surface of the subgrade causes permanent deformation.

On the basis of these criteria, design curves are developed by selecting different combinations of layer thicknesses of asphalt and base layers for a given subgrade modulus, asphalt mix type, and weighted mean annual air temperature (w–MAAT). The number of standard axle load applications expected during the design life are computed so that the critical strains do not exceed the permissible values [38]. Typical Shell pavement design charts are shown in Figure 12.14.

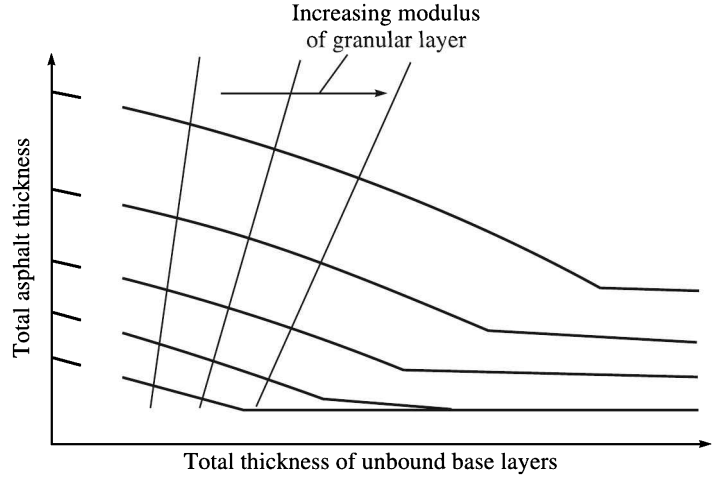


Figure 12.14 An example of Shell pavement design chart [206].

12.4.2 Asphalt Institute Method

In this method too, the bituminous pavement is characterized as a multilayered elastic system. Established theory, experience, test data from AASHO Road test, WASHO test, British and U.S. Army Corps of Engineer's experience (including those obtained in the laboratory), and a computer program, DAMA [42], were used by the Asphalt Institute to develop a comprehensive procedure for pavement design. Maximum tensile strains induced by wheel loads at the bottom of the bituminous layer and vertical compressive strain induced at the top of the subgrade layer are used as the design criteria for development of the design charts.

This method [237, 208] gives alternative sets of solution for the thicknesses of bituminous pavements in terms of:

(i) Full depth Asphalt Concrete (AC) on subgrade

(ii) Asphalt concrete surface and asphalt emulsion treated base on subgrade

(iii) Asphalt concrete surface over granular base.

Elastic modulus of asphalt mixes is estimated from pavement temperature, loading frequency, and the mix properties. Any pavement structure consisting of AC, emulsified asphalt mixtures, untreated aggregate materials, and subgrade soil can be analyzed by DAMA [42], provided the maximum number of layers does not exceed five. Strains at three transverse points, namely at the centre of one wheel, at the edge of one wheel, and at the mid-point of the dual-wheel system are calculated to find out the maximum tensile strain for damage evaluation. Cumulative damage is computed on monthly basis for the given traffic repetitions taking into consideration fatigue as well as rutting. The pavement is said to have failed when the cumulative damage in either mode reaches a value of unity (refer Section 10.5.4 for the cumulative damage principle). Asphalt Institute [237] also recommends a minimum necessary thickness for a particular construction. Mean Monthly Air Temperature (MMAT) data is used to account for the effect of temperature on the modulus of asphalt mix. The effects of freeze and thaw or variable moisture conditions throughout a pavement's life are also taken into account in the DAMA program.

For pavements with emulsified asphalt stabilized materials (see Section 13.10.1 for a discussion on emulsified bituminous mix), the DAMA program accounts for the effects of modulus changes over the specified cure time. Because of the added analytical complexities brought about by emulsified bituminous mixes in the damage computations, damage is separately computed on a monthly basis for the cure time. The time of cure is zero for a pavement structure not having an emulsified asphalt layer [42]. Design charts in accordance with the Asphalt Institute method are shown in Figure 12.15.

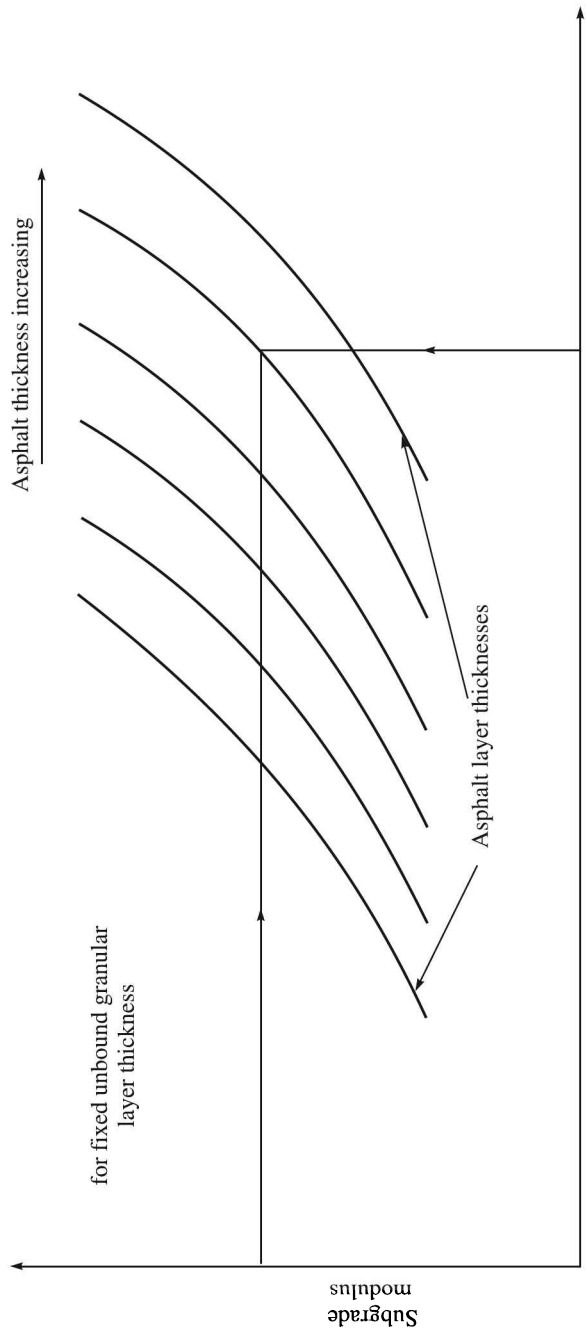


Figure 12.15 An example of Asphalt Institute pavement design chart [237].

12.4.3 Austroads Method

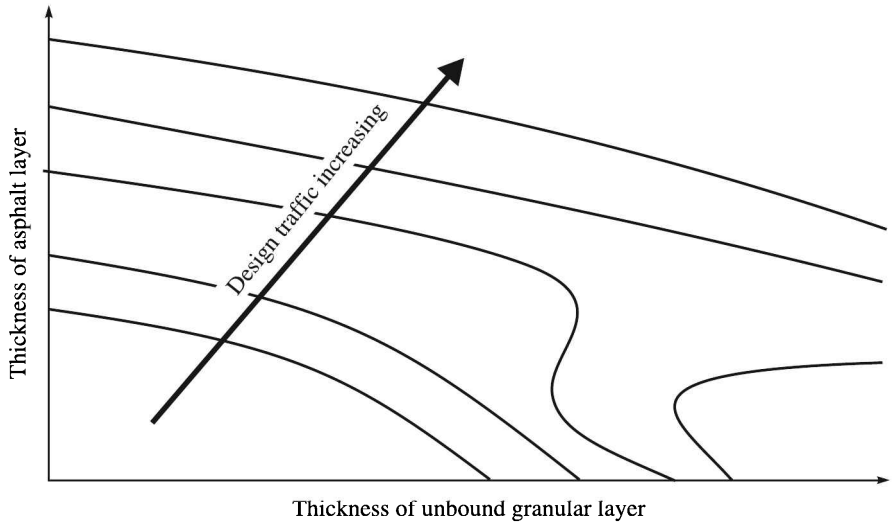
The Australian method is also based on a mechanistic approach. It presents a number of design charts with varying subgrade strength, asphalt mix stiffness, cumulative standard axle loading for bound and unbound bases and sub-bases. The charts are developed from the output of a finite element method (FEM) program, entitled 'CIRCLY'. Axle load repetitions, traffic wheel path distribution, loading rate, and tyre pressure, and so on, are the inputs from traffic considerations. Equations are proposed to determine the structural properties of various highway materials used in Australia. The design process consists of selecting trial pavement thicknesses and analyzing the performance. A pavement is assumed to be consisting of numerous layers which show elastic or viscoelastic behaviour. The damages caused by loads of different magnitudes and durations are computed and if the cumulative damage is less than unity, the pavement section is taken as adequate (see Section 10.5.3 for the cumulative damage principle). Australian design recommendations include extensive use of cemented bases in pavements. Figures 12.16(a) and (b) represent two sample design charts recommended as per the Australian practice.

12.4.4 South African Method

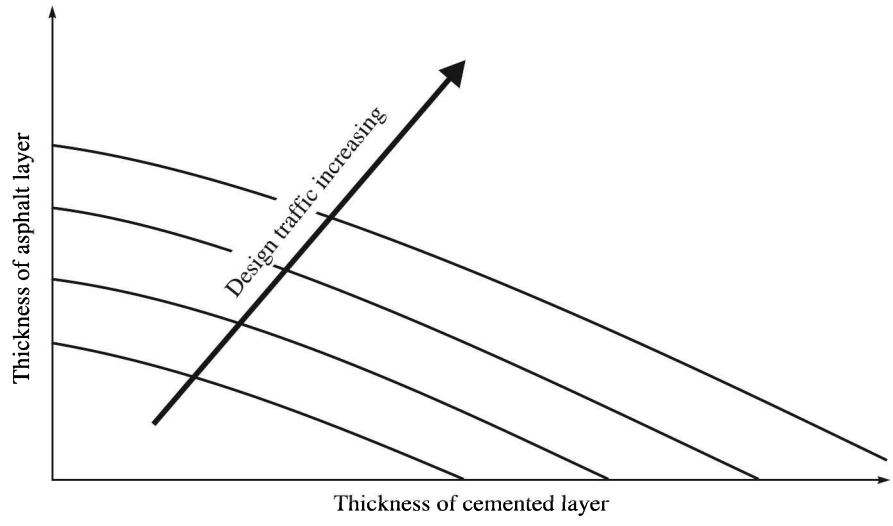
Structural analysis in the South African method involves linear elastic and static analysis of a multilayer system. The maximum horizontal tensile strain of asphalt layers and the maximum tensile strain at the bottom of the cemented layers are used as the critical parameters determining the fatigue life of these two materials [236]. Exhaustive work has been reported on use of cemented material as base and sub-base in the South African practice. The pavement is also designed in such a way that various pavement layers are stressed to the same level of the maximum bearing capacity. It is assumed that the bearing capacity of a structurally balanced pavement section increases evenly with depth up to the subgrade. For the design purpose, the thickness of the top layer is increased until the target bearing capacity of the pavement is achieved [52].

12.4.5 Road Note 29 Method

Current design practices for bituminous and concrete pavements in the UK are based on the Road Note 29 [245] guidelines. Road note 29 is developed based on the research carried out on various full-scale tests in the UK. It is also to some extent, based on the findings of AASHO road test and CBR method of pavement design [247]. The manual contains experience-based thickness design charts for sub-base, base, and bituminous surfacings plotted against cumulative standard axles. The aspect of drainage, weather protection, and frost susceptibility are also given due consideration. The required thickness of sub-base is determined from the construction traffic. For base course, a



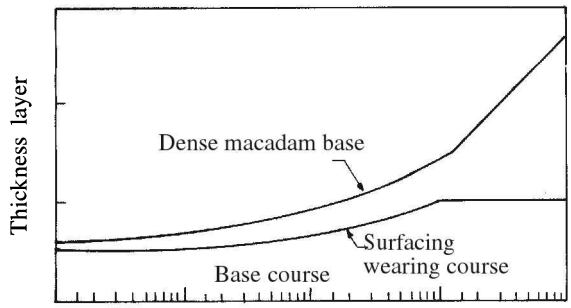
(a) Pavement design chart with granular layer



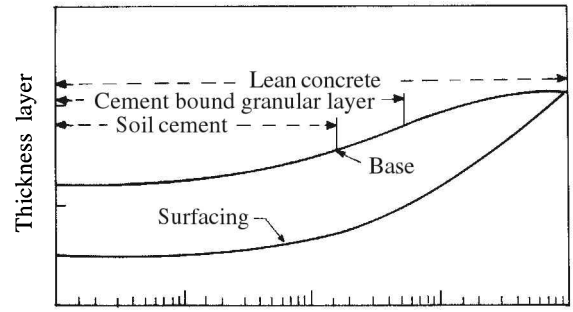
(b) Pavement design chart in cemented layer

Figure 12.16 Examples of Austroads pavement design chart.

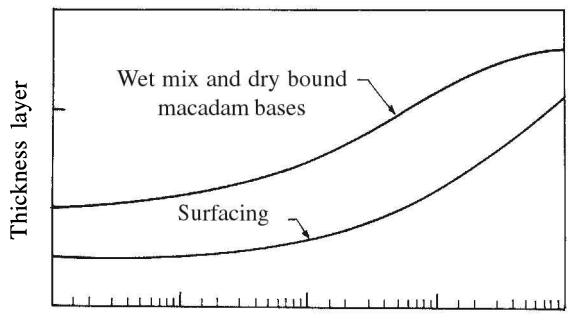
number of alternatives are suggested which include lean cement, concrete, dense tar macadam, dense bituminous macadam, rolled asphalt, wet mix macadam, dry bound macadam, soil cement, and cement bound granular material based on the expected traffic. Sample design charts recommended by Road Note 29 are illustrated in Figure 12.17.



Dense macadam base: minimum thickness of surfacing and base: road note 29



Lean concrete, soil cement and cement bound granular bases: minimum thickness of surfacing and base: road note 29



Wet mix and dry bound macadam bases: minimum thickness of surfacing and base: road note 29

Figure 12.17 Example of Road Note 29 Design Chart [245].

12.4.6 The AASHTO Design Method

The AASHTO method was originally developed based on the results of AASHO [1] road test conducted in Ottawa, Illinois, in late 1950s in the USA. Although the pavement design code was revised many times based on the latest experiences (in 1972, 1981, 1985, and 1993), the basic empirical equations evolved from AASHO test are still somewhat similar. Forty-nine USA States, District of Columbia, Puerto Rico, the Bureau of Public Roads, and several industry groups participated in the AASHO road test. Ten different axle load and axle configurations were used in this test [65].

According to the AASHTO [2] method, failure of a bituminous pavement is subjectively defined in terms of the quality of ride experienced by average road users. It was expressed as the Present Serviceability Rating (PSR) interpreted by a panel of road users using a scale of 0–5. The serviceability of a pavement was later converted to an equivalent parameter, called the Present Serviceability Index (PSI) which is measurable in terms of cracking, patching, and rut depth in the case of bituminous pavements. A brief discussion on PSI is covered in Section 14.3. A serviceability of 4.2 is assumed for bituminous pavements in the AASHO [1] test, immediately after construction; the terminal serviceability value can be assumed as 2.5, or 2.0, or even 1.5 depending upon the type of the road. When the pavement reaches its terminal serviceability level, it is assumed to have failed. The basic design equation used in this guide is given by

$$\log (N_{18}) = Z_R S_0 + 9.36 \log (SN + 1) - 0.20 + \frac{\log\left[\frac{\Delta PSI}{(4.2 - 1.5)}\right]}{0.40 + \frac{1094}{(SN+1)^{5.19}}} + 2.32 \log (M_R) - 8.07 \tag{12.10}$$

where

N_{18} is the predicted number of 18 kip equivalent single-axle load applications
S_0 is the combined standard error of the traffic and performance predictions
Z_R is the standard normal deviate
PSI is the Present Serviceability Index
ΔPSI is the change in PSI
M_R is the resilient modulus of subgrade (psi)
SN is the total Structural Number of the pavement

The values of the reliability factors Z_R and S_0 depend on the variability of traffic and materials' data used as inputs to the equation and also on the degree of desired reliability in design. The concept of Structural Number (SN) was first developed from the AASHO road test. The test used various combinations of surfacing thickness D_1 (AC thickness), the base thickness D_2 (e.g. wet-mix macadam, crushed stone, bitumen coated aggregate,

or cement coated aggregate thickness), and thickness D_3 of sub-base (e.g. sandy gravel). The *SN* value is computed from the equation

$$SN = a_1D_1 + a_2D_2m_2 + a_3D_3m_3 \tag{12.11}$$

where

a_i is the ith layer coefficient

D_i is the ith layer thickness (in inches)

m_i is the ith layer drainage coefficient.

The use of the layer moduli for computation of a_1, a_2, and a_3 made the 1986 AASHTO [2] guidelines more scientific. A reference elastic modulus E of 3000 MPa was chosen for bituminous mix and cement bound materials while it was taken as 160 MPa for unbound materials. The drainage coefficients were also introduced in the 1986 AASHTO guidelines [2]; these factors are based on the time required for an individual layer to become saturated as well as its quality of drainage.

For solving a pavement design problem in accordance with the AASHTO pavement design method [2], for a given number of standard axle load (as per AASHTO it is 18,000 lb) repetitions, the required value of *SN* is found out from Eq. (12.10), either iteratively or by nomographs. Finally, a pavement designer chooses a suitable pavement composition, such that the *SN* provided is close to what is required.

12.4.7 Japan Roads Association Method

The asphalt pavement design method, currently in use in Japan, is based on a semi-empirical approach. For determining the effective subgrade CBR value, a depth of 100 cm is considered. This is specially valid when the existing subgrade soil has been replaced by soil with a higher CBR value. In Indian specifications [89, 215] too, a poor quality subgrade needs replacement up to a depth of 500 mm, however.The recommendation of the Japan Roads Association method [151], is to use the following formula to find out the effective CBR value at a particular location:

$$\mathrm{CBR}_m = \left(\frac{h_1\mathrm{CBR}_1^{1/3} + h_2\mathrm{CBR}_2^{1/3} + h_3\mathrm{CBR}_3^{1/3} + \cdots + h_n\mathrm{CBR}_n^{1/3}}{100} \right)^3 \tag{12.12}$$

where

CBR_m is the effective CBR value of the location under consideration, in (%)

CBR_1, CBR_2, $\cdots$, CBR_n are the CBR values of soil of 1st, 2nd, 3rd, $\cdots$, nth layers in (%)

h_1, h_2, $\cdots$, h_n are the thicknesses of 1st, 2nd, 3rd, ..., nth layers, in cm.

As already mentioned, an effective depth of subgrade of 100 cm is considered in Eq. (12.12), and the CBR value of subgrade soil below 100 cm is ignored.

Thus, the effective CBR values of the other locations can be obtained in a similar manner. These effective CBR values, in a selected stretch will definitely show local variations. The design CBR value, CBR_{design}, is evaluated from these average, maximum, and minimum effective CBR values (i.e. $CBR_{average}$, $CBR_{maximum}$, and $CBR_{minimum}$) and a constant C (depending on how many CBR values are available), by the following formula:

$$CBR_{design} = CBR_{average} - \frac{CBR_{maximum} - CBR_{minmum}}{C} \tag{12.13}$$

The total thickness requirement of the pavement is obtained as follows [151, 133]:

$$H = \frac{28.0N^{0.1}}{CBR_{design}^{0.3}} \tag{12.14}$$

Similarly, the requirement of minimum equivalent full-depth bituminous thickness of the pavement, 'required T_A', is evaluated from the following equation [133]:

$$\text{required } T_A = \frac{3.84N^{0.16}}{CBR_{design}^{0.3}} \tag{12.15}$$

The individual depth compositions are then assumed and the T_A value is calculated using Eq. (12.16), similar to the SN [Eq. (12.11)] of the AASHTO method [2]. The 'calculated T_A' is expressed as the sum of the products of the coefficients of the relative strength of materials and the thicknesses of the individual layers, i.e.

$$\text{calculated } T_A = a_1h_1 + a_2h_2 + \cdots + a_nh_n \tag{12.16}$$

where

$a_1, a_2, a_3, \cdots, a_n$ are the coefficients of relative strength

$h_1, h_2, h_3, \cdots, h_n$ are the thicknesses of the individual layers.

The coefficients are different for different materials and are recommended in the *Manual for Asphalt Pavement* published by the Japan Road Association [151]. The thickness compositions are chosen in such a way, that the 'calculated T_A' is nearly equal to the 'required T_A'; the minimum requirement of the total pavement thickness H should also match simultaneously.

12.4.8 Indian Roads Congress Method

Indian Roads Congress first brought out a guideline for designing bituminous pavements in 1970 [87]. It was based on an empirical method where the thickness value of the

pavement used was read from the CBR of the subgrade. The guidelines were revised in 1984 [88] where traffic was expressed in terms of cumulative standard axle load. Figure 12.18 presents the pavement design chart as given by IRC:37–1984 [88]. In this design chart, the total pavement thickness could be read for a given CBR value and cumulative standard axle load. For example, the pavement thickness needed for design CBR value of subgrade as 3.5% and traffic level as 3 msa, is read from the chart as 572 mm.

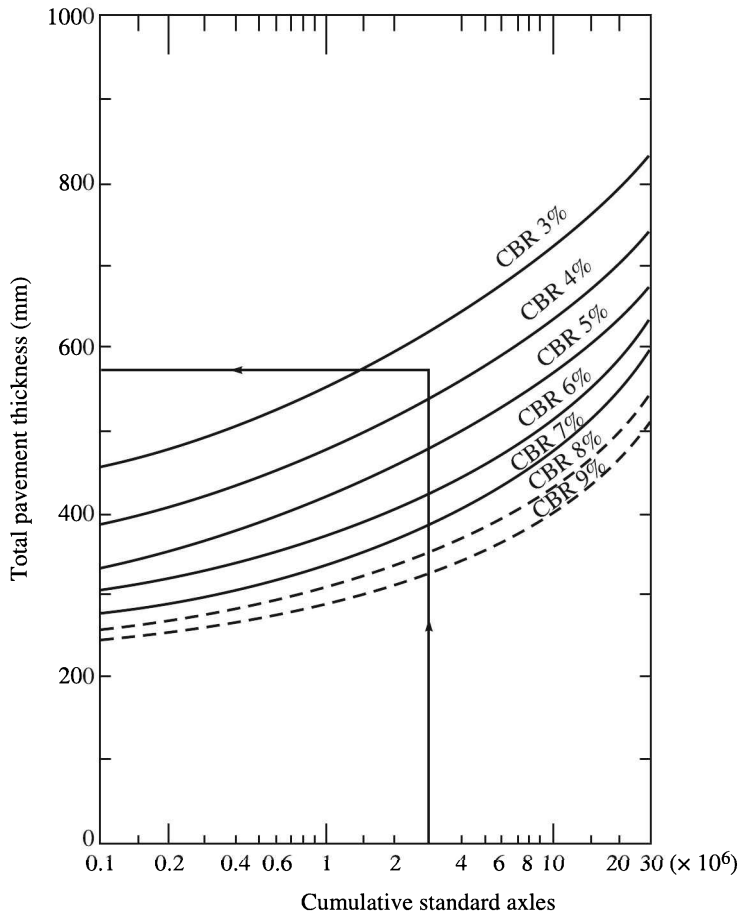


Figure 12.18 Pavement design curve as per IRC:37–1984 [88].

The new design guidelines [89] brought out in 2001, changed the design methodology completely. The present-day pavement design recommendations as per the IRC:37–2001 [89] are based on the mechanistic empirical pavement design principle, which has evolved from theoretical, laboratory and pavement performance studies [61, 62, 60] on Indian pavement materials and pavements constructed in India. In IRC:37–2001, for analysis, DBM layer with 60/70 bitumen and an annual average pavement temperature (AAPT) of 35°C have been used. The vertical strain on the subgrade between the dual

wheels and the horizontal tensile strain at the bottom of the DBM layer below one of the wheels (see Figure 12.5), are assumed to be the design criteria for fatigue and rutting failure of the pavement. A brief discussion on various steps to be followed for the design of bituminous pavements as per IRC:37–2001 [89] is presented below.

Traffic calculation

The design traffic in terms of the cumulative standard axle is obtained using the equation

$$N = 365 \times A \times \frac{(1+r)^n - 1}{r} \times \text{VDF} \times \text{LDF} \tag{12.17}$$

where

N is the design traffic
VDF is the vehicle damage factor
LDF is the lane distribution factor.

Subgrade strength

The CBR value of the subgrade is determined for the compaction level and moisture conditions as close as possible to the weakest condition likely to occur in the pavement. Four-day soaked CBR value may be an underestimation of subgrade strength, specially when the water-table is at a deep level, or climate is arid throughout the year. In such cases, the CBR value may be obtained at the natural moisture content at the subgrade, immediately after monsoon [89]. The CBR test procedure and other details have already been discussed in Section 10.12.

Design thickness

Design charts are available in IRC:37–2001 [89] for CBR values ranging from 2%–10% and traffic ranging from 1 msa–150 msa. Figure 12.19 gives an example design chart

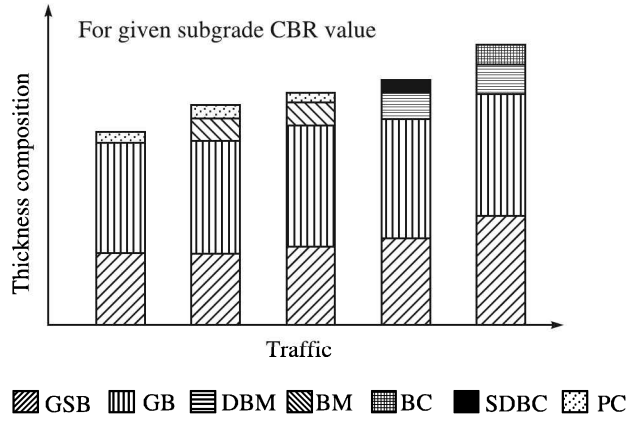


Figure 12.19 An example pavement design curve as per IRC:37–2001 [89].

from IRC:37–2001 [89]. From these charts, for the given traffic (msa) and CBR values, the recommended thickness composition can be determined.

Design tables are also available which give the same pavement thickness composition in a tabular form. Interpolation is suggested when the value lies somewhere in between. Table 12.5 presents a typical design thickness value based on IRC: 37–2001[89].

Table 12.5 Recommended designs for traffic range 2 to 10 msa for subgrade CBR 7% [89]

Cumulative traffic (msa)	*Pavement composition*			
	Bituminous surfacing		*Granular base* (mm)	*Granular sub-base* (mm)
	wearing course	*binder course* (mm)		
2	20 PC	50 BM	225	150
3	20 PC	50 BM	250	160
5	25 SDBC	50 DBM	250	180
10	40 BC	60 DBM	250	230

Various other considerations

For values of traffic or subgrade where the CBR does not exactly match with the design charts or tables, the designer is required to interpolate linearly between the values. There are further restrictions to the design thicknesses recommended by the guidelines [89]. Some of these are as follows:

(i) The thickness of sub-base should not be less than 150 mm for design traffic less than 10 msa.

(ii) The CBR value of the subgrade should have a minimum value of 2%. If it is less than that, a capping of 150 mm thickness of material of CBR value of at least 10% should be provided.

(iii) If WBM (specifications for various pavement compositions are discussed in Chapter 13) construction is adopted for base course for traffic more than 10 msa, the thickness of WBM should be increased from 250 mm to 300 mm, with equal reduction in the sub-base thickness keeping the overall pavement thickness same.

(iv) If the granular layer is manually laid, the DBM layer may be preceded by 75 mm BM layer, with an equivalent reduction in the DBM thickness value.

(v) Thickness of Premix carpet (PC), if used as wearing course, up to 25 mm, should not be counted as structural layer.

For finding layer equivalency, IRC:37-2001 [89] recommended the following formula

$$\frac{E_1 h_1^3}{12(1-\mu_1^2)} = \frac{E_2 h_2^3}{12(1-\mu_2^2)} \tag{12.18}$$

where E_1, h_1, and μ_1 are the elastic modulus, height, and Poisson's ratio values respectively for bituminous material type 1.

EXAMPLE 12.7

Design a suitable bituminous pavement section for a new pavement construction with the following available information.

The pavement will be a two-lane road with a single carriageway. The traffic expected is 500 commercial vehicles per day in both the directions with average vehicle damage factor of 1.9. Design subgrade CBR is 7% and the assumed design life of the pavement is 12 years.

Solution

The lane distribution factor, as per recommendation, is chosen as 0.75. Traffic growth factor is assumed as 7.5%. The value of cumulative standard axle repetitions is calculated from Eq. (12.17) as

$$N = 365 \times 500 \times \frac{(1+0.075)^{12}-1}{0.075} \times 1.9 \times 0.75$$

$$= 4.79 \text{ msa} \approx 5 \text{ msa}$$

The thickness values read from Table 12.5 for 5 msa, are obtained as 25 mm SDBC, 50 mm DBM, 250 mm granular base of WBM, and 180 mm granular sub-base with CBR value not less than 30% [89].

12.4.9 Closing Remarks

We have briefly described some of the important pavement design guidelines. The recommendations by different guidelines show wide variations. As a separate exercise [60], design thicknesses of IRC:37–2001 are compared with other design guidelines. Thicknesses of bituminous pavement layers by the Austroads [182] and the AASHTO [2] methods are by and large, in agreement with the suggested design. But thicknesses for base as well as surfacing by other methods are generally found larger compared to those

obtained by the Indian method, for a given traffic and the subgrade strength. This may be because of the fact that the acceptable level of serviceability is higher in other countries than what is accepted in India. In the current IRC design method [89], the fatigue and rutting criteria are based on 20% surface cracking and 20 mm as the limiting rut depth, whereas these values may be lower in other countries [15, 24, 167, 52].

12.5 PRESENT TREND IN CONCRETE PAVEMENT DESIGN

Since concrete pavements fail due to bending stresses, it is necessary that their design is based on the flexural strength of concrete. Concrete pavement is sometimes placed directly over the subgrade, but the chances of mud pumping and frost heave (wherever applicable) being more, designers prefer to use a base course over the concrete pavement. Provision of base course may not bring any economy in terms of the thickness of the concrete pavement, the strength of concrete being very high compared to other layers. The following are the reasons why a cemented base construction is preferred to concrete pavement just over the subgrade [182].

1. It provides a working platform for the construction equipment.
2. Gives a uniform support to the pavement.
3. Reduces deflections at the joints, thereby ensures better load transfer.
4. To some extent, checks the effect of shrinkage and swelling of subgrade.
5. Resists erosion of subgrade due to mud pumping action.

As mentioned in Chapter11, it is generally the plain concrete slab that is used for the construction of concrete pavements, which is jointed at regular intervals with the next concrete slab. The following sections contain brief discussions of concrete pavement design as suggested by various design practices, namely the PCA method, Austroads method, AASHTO method, and the Indian Roads Congress method.

12.5.1 PCA Method

The PCA method is developed by the Portland Cements Association, USA. The PCA method is based on Westergaard, Picket, and Ray's work and further theoretical analysis by the finite element method (FEM). The data generated from various road tests, such as the AASHO road test, the Arlington test (conducted by PCA), the Bates road test, and the Maryland road test were used to develop the PCA concrete pavement design methodology [239]. The PCA design method is based on the following two considerations [239]:

(a) Consideration of fatigue damage to the concrete slab due to repetitive application of traffic load. The fatigue characteristics of the concrete are used to develop this criterion.

(b) Considerations of possibilities of erosion of pavement materials placed below the concrete slab. The rate at which the slab is deflected due to axle load is used as a criterion for erosion. The computed corner deflection, pressure, and the radius of relative stiffness are used in the analysis. Theoretically, it was shown that a thin pavement with a smaller deflection basin is subjected to a faster rate of deflection, compared to a thicker slab, hence the former is more susceptible to erosion.

The steps involved in the design of concrete pavement by the PCA method are:

(a) The stresses developed for a trial thickness of the concrete slab are calculated for various axle loading configurations and the critical one is chosen.

(b) The ratio between the developed stresses and the modulus of rupture is calculated. Recall Section 10.8 where it is defined as stress ratio.

(c) The allowable repetitions from fatigue considerations are obtained from the fatigue characteristics of concrete for a given stress ratio (refer Section 10.7). Similarly, the erosion factor is calculated from the stress ratio wherefrom the allowable repetitions from erosion considerations are obtained.

(d) The individual damage fraction is calculated by dividing the actual traffic repetitions by the allowable repetitions of that particular axle load.

(e) The process is repeated for various axle loads and the cumulative damage is calculated. The cumulative damage should be equal to one for pavement to be just safe (see Section 10.5.4), otherwise the trial thickness is changed and the design process repeated.

12.5.2 Austroads Method

The Austroads [182] method encourages the use of cemented sub-base course laid below the concrete pavement. The thickness of the sub-base course can be chosen from the design traffic and the CBR value of the subgrade soil. The sub-base course can be cement stabilized granular material, lean rolled concrete, or lean mix concrete. The characteristic 28-day compressive strength must attain a minimum value of 5 MPa.

Austroads [182] has adopted the Portland Cement Association (1984) method [239] for the design of the thickness of concrete pavements.

12.5.3 AASHTO Method

The AASHTO design method [2] for design of concrete pavements was evolved from the AASHO road test [1]. Pavement performance, subgrade and sub-base strength,

traffic, properties of concrete, drainage, and reliability were the aspects considered in the pavement design. The following equation for design of rigid pavements is suggested by the AASHTO code [2].

$$\log (W_{18}) = Z_R S_0 + 7.35 \log_{10} (D + 1) - 0.06 + \frac{\log_{10}\left[\frac{\Delta PSI}{4.2 - 1.5}\right]}{1.0 + \frac{1.624 \times 10^7}{(D+1)^{8.46}}}$$

$$+ (4.22 - 0.32 p_t) \log_{10} \left[\frac{S'_c C_d (D^{0.75} - 1.132)}{215.63 J (D^{0.75} - \frac{18.42}{(E_c/k)^{0.25}})}\right] \quad (12.19)$$

where

W_{18} is the predicted number of 18 kip equivalent single-axle load applications
S_0 is the combined standard error of the traffic and performance predictions
Z_R is the standard normal deviate
p_t is the terminal pavement serviceability
PSI is the Present Serviceability Index
ΔPSI is the change in PSI
D is the thickness of the slab (inches)
S'_c is the modulus of rupture (psi) for the cement concrete
C_d is the drainage coefficient
E_c is the modulus of elasticity (psi) for cement concrete
k is the modulus of subgrade reaction (psi)
J is the load transfer coefficient

The design of concrete pavement thickness D is found out by substituting various trial thicknesses in the above equation, and then checking at which value the allowable repetitions are close to the expected repetitions.

12.5.4 Indian Roads Congress Method

As mentioned earlier (see Section 10.3.2), the modulus of subgrade reaction k is the necessary subgrade strength parameter used for the design of concrete pavements. The k value is derived from the plate load test. This procedure being relatively costly and time consuming compared to CBR testing, the k value is sometimes estimated from the CBR value (as shown in Table 10.1) recommended by IRC:58 [91].

As discussed in Section 11.5.2, the load stress is the highest in the corner of pavement, less at the edge, and the least in the interior of the pavement. The order in which the temperature stress varies, is just the reverse of this. Therefore, a check is required to see where the overall effect is maximized and that the maximum combined stress is considered for design of thickness. IRC:58–1988 [91] recommends that both the combined corner stress and the combined edge stress should be checked for finding out the most critical stress between these two. The following are the various expressions used for estimation of stress as per the IRC:58-1988 [91] guidelines.

Edge stress due to load

IRC:58-1988 has adopted the analysis by Westergaard (modified by Teller and Sutherland). The load stress in the critical edge region is given by [91]

$$\sigma_{le} = 0.529\frac{P}{h^2}(1 + 0.54\mu)\left(4\log_{10}\frac{l}{b} + \log_{10} b - 0.4048\right) \tag{12.20}$$

where

σ_{le} is the load stress in the edge region (kg/cm^2)
P is the design wheel load (kg)
h is the pavement slab thickness (cm)
μ is the Poisson's ratio of concrete
E is the modulus of elasticity of concrete (kg/cm^2)
k is the modulus of subgrade reaction (kg/cm^3)
l is the radius of relative stiffness (cm) $\left[= \left(\frac{Eh^3}{12(1-\mu^2)k}\right)^{1/4}\right]$
b is the radius of equivalent distribution of pressure

$$\left[\begin{array}{l} = a \text{ for } \frac{a}{h} \geq 1.724 \\ = \sqrt{1.6a^2 + h^2} - 0.675h, \text{ for } \frac{a}{h} \leq 1.724 \end{array}\right]$$

a is the radius of circular contact area (cm)

Edge stress due to temperature

The edge stress due to temperature considered by IRC:58–1988 [91] is given by (see Section 11.5.2 for review),

$$\sigma_{te} = \frac{E\alpha\Delta T}{2}C \tag{12.21}$$

where

σ_{te} is the temperature stress in the edge region (kg/cm^2)

ΔT is the maximum temperature differential during day time between the top and bottom of the slab

α is the coefficient of thermal expansion of concrete

C is the Bradbury's coefficient

L is the slab length between consecutive construction joints

W is the slab width

l is the radius of relative stiffness.

The Bradbury's coefficients are obtained from Table 12.6 as follows:

Table 12.6 Table for Bradbury coefficient [91]

L/l or *W/l*	1	2	3	4	5	6
C	0.000	0.040	0.175	0.440	0.720	0.920
L/l or *W/l*	7	8	9	10	11	12 and above
C	1.030	1.075	1.080	1.075	1.050	1.000

Corner stress due to load

The corner stress due to load is obtained by Westergaard's analysis (modified by Kelley) as

$$\sigma_{lc} = \frac{3P}{h^2}\left[1 - \left(\frac{a\sqrt{2}}{l}\right)^{1.2}\right] \tag{12.22}$$

Corner stress due to temperature

The corner stress due to temperature is negligible as corners are free to warp.

Steps followed in pavement design

The following are the steps recommended by the IRC:58–1988 [91] guidelines for the design of concrete pavements.

(i) The input parameters are obtained to formulate the design problem. The joint spacing and the lane width are also decided. If there is a bound base layer put over the subgrade, a suitable value of effective k is chosen from theoretical or other considerations [92].

(ii) A trial thickness of the concrete slab is assumed.

(iii) The critical temperature stress at the edge is evaluated from Eq. (12.21) and the available residual strength of concrete is found out.

(iv) Critical load stress at the edge is calculated from Eq. (12.20) and the factor of safety is calculated. The factor of safety in this case could be defined as load stress divided by the residual stress. The factor of safety should either be equal to one or slightly more than that. If the factor of safety is less than 1 or far in excess of 1, the design steps (ii) and (iii) are repeated by changing the slab thickness.

(v) Adequacy of the corner stress is checked and the design is finalized.

EXAMPLE 12.8

Design a suitable concrete pavement (4.5 m × 3.5 m) as per IRC:58-1988 [91], situated at Kanpur, for design wheel load of 4100 kg and tyre pressure of 7 kg/cm^2. The CBR value of the subgrade soil is found to be 4.5%. The forecasted traffic intensity at the end of design life is 1000 CV/day (assume other parameters wherever necessary).

Solution

The pavement thickness h is assumed to be 19 cm.

From Table 12.4, the temperature differential (ΔT) is found to be 12.98°C. Table 10.1 gives the coefficient of subgrade reaction as 3.81 kg/cm^3.

$$\text{The radius of relative stiffness, } l = \left[\frac{Eh^3}{12(1-\mu^2)k}\right]^{1/4} = \left[\frac{3\times10^5\times19^3}{12(1-0.15^2)\times3.81}\right]^{1/4} = 82.37 \text{ cm}$$

$$L/l = 5.46, \text{ and } W/l = 4.25$$

From Table 12.6, the Bradbury coefficient, $C = 0.812$.
From Eq. (12.21), the edge stress due to temperature

$$\sigma_{te} = \frac{E\alpha\Delta T}{2}C = \frac{3\times10^5\times10\times10^{-6}\times12.98}{2} \times 0.812 = 15.81 \text{ kg/cm}^2$$

Therefore

$$\text{Residual concrete strength} = 40 - 15.81 = 24.19 \text{ kg/cm}^2$$

where 40 kg/cm^2 is the modulus of rupture of concrete (assumed)

$$\text{Radius of circular contact area, } a = \left(\frac{4100}{7\times\pi}\right)^{1/2} = 13.65 \text{ cm}$$

$$a/h = 0.718 < 1.724$$

Therefore

$$b = \sqrt{(1.6\times13.65^2 + 19^2)} - 0.675 \times 19 = 12.85 \text{ cm}$$

Load stress in the edge region is calculated from Eq. (12.20) as

$$\sigma_{le} = 0.529 \frac{4100}{19^2}(1 + 0.54 \times 0.15)\left(4\log_{10}\frac{82.37}{12.85} + \log_{10}12.85 - 0.4048\right)$$

$= 25.53$ kg/cm^2 which is more than the residual stress; hence the design is unsafe.

Thus, a new pavement thickness h of 20 cm is assumed. The calculations are repeated in the same way and the values obtained are as follows:

$\Delta T = 13.1°C$	$l = 85.60$ cm
$L/l = 5.25$	$W/l = 4.08$
$C = 0.77$	$\sigma_{le} = 15.13$ kg/cm^2
residual stress $= 24.86$ kg/cm^2	$a/h = 0.683$
$b = 12.92$	$\sigma_{le} = 23.39$ kg/cm^2

Therefore, the factor of safety $= \dfrac{24.86}{23.39} = 1.06$. Hence, the design is fine. The corner stress is calculated as

$$\sigma_{le} = \frac{3\times4100}{20^2}\left[1 - \left(\frac{13.65\sqrt{2}}{85.6}\right)^{1.2}\right] = 25.6 \text{ kg/cm}^2$$

This is less than the modulus of rupture.

From Table1 of IRC:58–1988 [91], the traffic category is 'E'. Therefore according to Table 9 of IRC:58–1988 [91], no further thickness adjustment is necessary.

Hence the design thickness is 20 cm.

Design of dowel bar

The following equations are adopted by IRC:58–1988 [91] for estimating the load transfer capacity by a single dowel bar, from the point of view of shear, bending (in the bar) and bearing (on concrete).

$$P_{\text{shear}} = 0.785d^2 f_{ss} \tag{12.23}$$

$$P_{\text{bending}} = \frac{2d^3 f_{sf}}{r + 8.8z} \tag{12.24}$$

$$P_{\text{bearing}} = \frac{f_{cb} r^2 d}{12.5(r + 1.5z)} \tag{12.25}$$

where

P is the load transfer capacity of a single dowel bar
d is the diameter of the dowel bar
r is the length of embedment of the dowel bar
z is the joint width
f_{sf} is the permissible flexural stress in the dowel bar
f_{ss} is the permissible shear stress in the dowel bar
f_{cb} is the permissible bearing stress in concrete.

The length of the dowel bar can be obtained by solving Eqs. (12.24) and (12.25) and can be expressed in the following form:

$$r = 5d\left[\frac{f_{sf}}{f_{cb}}\left(\frac{r+1.5z}{r+8.8z}\right)\right]^{0.5} \tag{12.26}$$

From Eq. (12.26), the length of the dowel bar is determined. It is assumed that dowel bars up to $1.8l$ (where l is the radius of relative stiffness) in both the directions participate in the load transfer, and the load carried by the dowel bars varies linearly [91]. The following example shows how to determine the load carried by a single dowel bar. The load on a single dowel should not exceed its capacity. This forms the basis of finding the interval of placing of the dowel bars.

EXAMPLE 12.9

If the radius of relative stiffness is 55 cm, and if the dowels are placed at every 30 cm, calculate the maximum load carried by a single dowel which is just below the wheel. Assume that a wheel of 4100 kg is placed at the joint corner, and 50% of the load is transferred through the joints.

Solution

As the wheel is placed at the joint corner, only the dowel groups of one side are available to participate in the load transfer.

Now,

$$1.8l = 1.8 \times 55 = 99 \text{ cm}$$

If the load carried by a single dowel just below the wheel is P, then, the total load carried by the dowel system is

$$\left(1+\frac{99-30}{99}+\frac{99-60}{99}+\frac{99-90}{99}\right)P = 2.18P$$

Now,

$$2.18P = 4100 \times 0.50$$

Therefore,

$$P = 940.36 \text{ kg}$$

Hence, the load carried by the single dowel just below the wheel is about 940 kg.

Design of tie bar

As the tie bars holds the two slabs together, the following formula is recommended [91] to calculate the area of the steel required as tie bar, per unit joint length.

$$A_s = \frac{b \times f \times W}{f_{sw}} \tag{12.27}$$

where

A_s is the area of the steel per unit length of joint

b is the distance to the other edge of the slab (width of the slab)

f is the coefficient of friction between the concrete slab and the layer below it

W is the weight of the slab per unit area

f_{sw} is the allowable working stress of the steel used as a tie bar.

The length of the tie bar is determined from the bond strength required to sustain a pull equal to the working stress of the steel bar. Thus, the length of the tie bar can be expressed as

$$L = \frac{2 f_{sw} \times A}{f_{sb} \times P} \tag{12.28}$$

where

L is the length of the tie bar

f_{sw} is the allowable working stress of steel

A is the cross-sectional area of the tie bar

P is the perimeter of tie bar

f_{sb} is the permissible bond stress.

12.5.5 Closing Remarks

The PCA [238, 239] method recommends that the warping stresses due to temperature gradient are not critical for the concrete pavement. This recommendation appears to be applicable to roads in India as well. In many parts of India, summers are extreme and long. The warping stresses are likely to be high during summers, but at the same time

the subgrade is very stiff during that time and accordingly the wheel-load stresses are likely to be lower. In concrete, the variation of temperature with depth is not normally linear and the temperature difference between the surface and the midpoint may be more than double the difference between the midpoint and the underside for slab thickness of 254 mm and above [44]. The tensile stresses computed using the Thomlinson's approach are much lower than those computed by the IRC:58–1988 formula [91], which is an adoption of Bradbury's equation [20], where the temperature gradient is assumed to be constant throughout the depth of the slab. Commercial vehicles moving on roads in hot afternoons are usually less in number when the temperature gradient in the concrete slab is high. Keeping there in mind, the temperature warping stresses, therefore, do not appear to be critical [209].

12.6 DRAINAGE CONSIDERATIONS IN PAVEMENT DESIGN

Road engineers recognize that the water entrapped in the base, sub-base or subgrade of a pavement is the predominant single factor which causes premature deterioration of the pavement. Also, the accumulated water on the pavement surface causes inconvenience to the traffic. Water may cause damage to the pavement in many ways including the following [70, 71, 2]:

1. Excess moisture in the subgrade soil causes considerable reduction in its CBR value.
2. Excess moisture in the subgrade causes changes in the volume of the clayey subgrade, which affects the riding quality of the pavement surface.
3. Accumulated water on the surface of carriageway causes hydroplaning, hence, there is a loss of skid resistance.
4. The presence of water in the porous pavement for prolonged durations leads to stripping of bitumen from the aggregates.
5. Mud pumping may occur through cracked portions of the pavement during passage of traffic.
6. Heavy wheel loads moving at high speeds build up significant pulsating pore pressure leading to breakdown of the pavement.
7. Erosion is caused to the shoulders, verges, and embankments, by the flowing water.
8. Chances of pavement being affected by freezing and thawing increase.
9. Ingress of water inside the pavement which is placed on embankment or in cut, increases the chances of stability failure, by increasing the stress and decreasing the shear strength of soil.

Therefore, lack of drainage facility in a pavement and subsequent accumulation of water

entrapped within the pavement may cause a serious damage and fast deterioration of the pavement. Proper drainage considerations, in design and construction, on the other hand, reduce the pavement maintenance cost to a large extent.

There are two possibilities of the entry of water into the pavement, namely the percolation of rainwater run-off, mainly through cracks, joints, shoulder, etc. and the capillary rise from the water-table or the existence of groundwater table at a higher level. The drainage systems required to remove the above two sources of water are called *surface drainage* and *sub-surface drainage* respectively. The following subsection contains discussion on surface and sub-surface drainage systems of the pavement.

12.6.1 Surface Drainage

The gradation of the aggregates and the bituminous mix should be so adjusted that rainwater (or snow water) is not allowed to percolate into the compacted mixtures. This can be done by suitably choosing a bituminous wearing course, which is adequately impermeable and providing requisite cross slopes to the top surface of the pavement, both on carriageway and shoulder to drain off rainwater quickly. Investigations show [140] that bituminous mixes recommended in Indian specifications [213] for various wearing courses, are adequately impermeable to water when newly constructed. The entry of water into the bituminous pavement occurs mainly when the cracks appear on the surface. For a good surface drainage system, transverse slope, longitudinal slope, and longitudinal channels should be provided as discussed below.

Transverse slope

To drain off water from the pavement surface, a suitable cross slope is required to be provided on the carriageway and shoulder. The requirement of a high cross slope for good surface drainage conflicts with that of the drivers' comfort [65, 71]. Also, a large slope may cause erosion mainly along the unsurfaced parts, such as shoulders as the case may be. The cross slopes for different types of road, for Indian conditions, are specified by IRC:73–1980 [79] guidelines.

Longitudinal slope

A minimum longitudinal slope is necessary for the pavement to provide the necessary slope to the longitudinal channels. A minimum longitudinal drainage of 0.3% is adequate for satisfactory drainage as recommended by Indian specifications [70].

Longitudinal channels

For the purpose of carrying water collected from the rainwater run-off, longitudinal channels in the form of surface or sub-surface drains are provided. Longitudinal channels can be open or covered drains. The open drains, being exposed, sometimes get clogged with debris, and hence the effective hydraulic capacity reduces. However, inspection

is easier in open drains compared to closed ones. Closed drains are preferred in urban areas. The sides of the drains should be properly lined, or a suitable turfing should be made to prevent erosion.

Median drainage

Median should be crowned, wherever paved or turfed, for efficient drainage across the pavement [71]. Longitudinal drains and manholes may be provided for proper drainage of median.

Design considerations of surface drainage

Design of surface drainage systems involves two stages first, the calculation of the total discharge that the system requires to drain off, and second, the design of the slope and the dimensions of the necessary drainage system.

The storm water run-off can be determined by the rational method, hydrograph method, computer models, or empirical formulae [71]. The time of concentration, I_c, is the time required for the run-off to flow from the farthest point of the catchment area, to the outlet. The time of concentration can be divided into two parts, namely the entry time and the time of flow [70]. If the drainage point considered in the design is the entry point to the drainage system, then the entry time is equal to the time of concentration. Charts and tables are provided in IRC codes [70,72] to estimate the time of concentration for catchments of different lengths, characteristics, and slopes. Once the time of concentration is determined, the next step is to find out the intensity of rainfall for storm duration equal to the time of concentration. This is done with the help of rainfall maps. In India, the rainfall maps for durations less than one hour are not available [70], so a suitable correction factor is used for the time of concentration less than an hour. The design peak run-off Q is given by [70]

$$Q = 0.028PAI_c \tag{12.29}$$

where

P is the coefficient of run-off, which is the characteristic of the catchment area that determines how much fraction of the storm water remains on the surface

A is the area of the catchment

I_c is the critical rainfall intensity for a selected frequency and duration equal to at least the time of concentration. In case the catchment area is constituted with surfaces of different run-off coefficients, a weighted average is then taken. The capacity of the drains can be designed using the Manning's formula,

$$Q = \frac{1}{n}Av = \frac{1}{n}A(R^{2/3}S^{1/2}) \tag{12.30}$$

where

Q is the discharge

v is the mean velocity

n is the Manning's rugosity coefficient

R is the hydraulic mean radius
S is the gradient of the drainage bed
A is the area of flow cross-section.

12.6.2 Sub-surface Drainage

Even if some water enters the pavement, whatever may be the reason, as mentioned above, water should be drained off quickly, from within the pavement. Thus, a layer of graded aggregates is laid to a suitable thickness so that water percolating into the pavement is drained away fast. This drainage layer is a part of the structural layer of the pavement and a suitably designed sub-base or base course itself can serve as drainage layer. Geotextiles or bitumen treated aggregates can also be used as drainage layers. The gradation of a drainage layer must not contain any fine particles and compaction should be done carefully to avoid the degradation of aggregates. It is found that fine content as little as 5% may drastically lead to fall in the value of permeability [30].

Between the drainage layer and the subgrade, another layer, called the filter layer, is sometimes laid. This layer is also intended for capillary cut-off, and therefore, it prevents further rise of capillary water. The filter material may be granular fine aggregates or suitable geosynthetics, or the filter fabric can also be used instead. The particle size distribution of the filter layer is selected in such a way that the inter-particle gap is large enough to allow the water to flow through it rapidly but small enough to prevent entry of soil so that foundation soil is not washed away.

The filter layer fulfil the sub-surface drainage requirement in low and medium rainfall areas. In high rainfall areas (annual rainfall more than 750 mm), the filter layer as well as the drainage layer are provided together [78]. It is generally recommended that the drainage layer should not be directly laid over the subgrade, as the subgrade particles will tend to be washed away and the drainage layer can get clogged by accumulation of the soil particles. The width of the drainage layer and the filter layer should extend up to the formation width. End plugging with more durable material made be necessary, for both the layers, to prevent particles from getting washed away. Addition of 2% cement or 2% bitumen with the designed drainage or filter layer serves the purpose. Figure 12.20 presents a schematic diagram of drainage and filter layers in a pavement cross-section. Figure 12.21 shows how the provision of longitudinal drains can bring down the existing groundwater level, thereby keeping the pavement unaffected.

Perforated collector tube

If the flow is adequate, perforated pipes are provided longitudinally to collect and carry water from the drainage layer (see Figure 12.22). The pipe must be laid over a stable bed, and should be surrounded by filter material. The holes are provided only in one-half of the circumference of the perforated pipe. The following criteria are used to design the perforation width of the collector pipe [78,30].
For rectangular perforation,

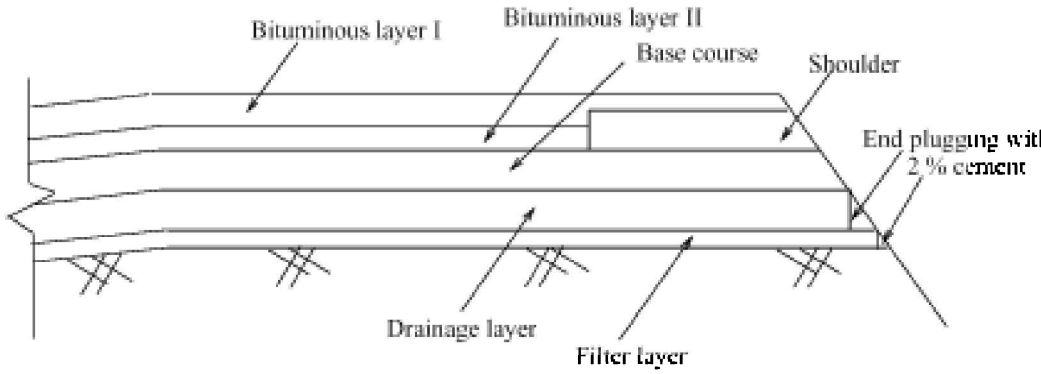


Figure 12.20 Drainage and filter layers in a pavement cross-section.

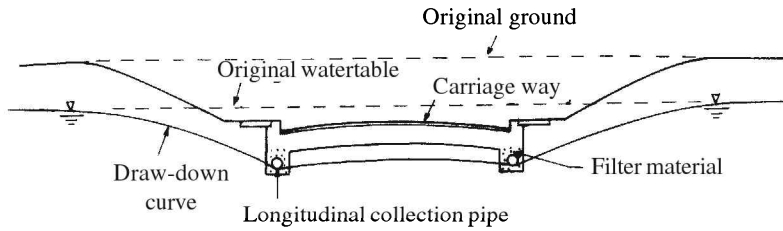


Figure 12.21 Provision of longitudinal drains for existing groundwater table above the road level.

$$\frac{D_{85} \text{ of the filter layer}}{\text{width of the slot}} > 1.2 \tag{12.31}$$

For circular perforation,

$$\frac{D_{85} \text{ of the filter layer}}{\text{diameter of the hole}} > 1.0 \tag{12.32}$$

The slots should be equidistant and the distance between the slots can be calculated from the expected discharge (see Figure 12.22). The location of the outlet should be clearly marked in the field for future maintenance purposes and care needs to be taken such that it does not get clogged. In case the perforated pipes are not used, the drainage material with sufficient permeability should be used as outlets.

Sub-surface drainage network

When a large amount of water flow is expected, drainage layers along with the collector and outlet pipes should be provided. Properly designed longitudinal and transverse drainage network can also be provided for efficient drainage of sub-surface water. Figure 12.23 shows such a sub-surface drainage network.

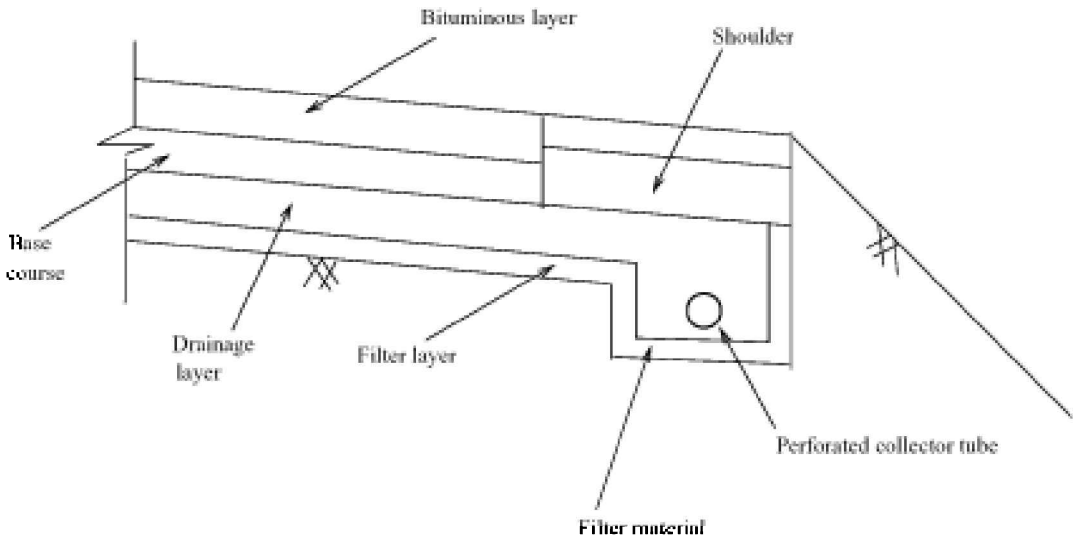


Figure 12.22 Provision of perforated collector tube.

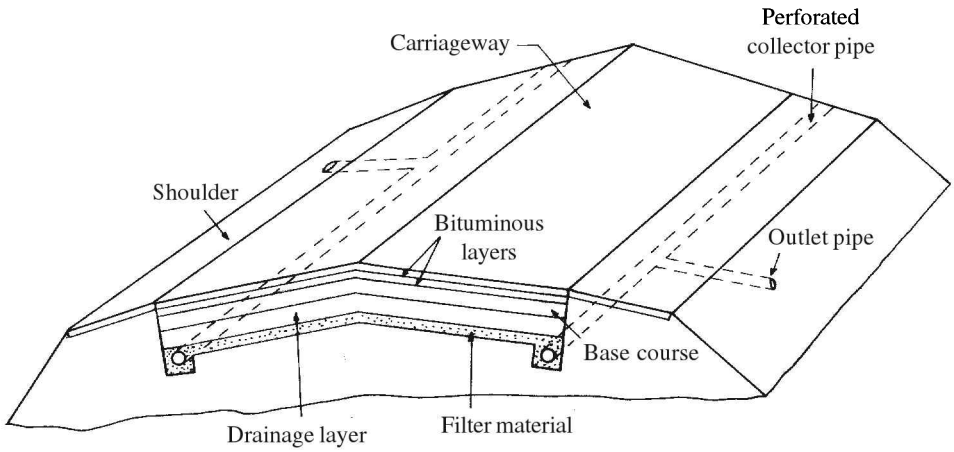


Figure 12.23 Provision of longitudinal and transverse drainage network.

Thickness design of drainage layer

The following example illustrates how the thickness of a drainage layer can be designed.

EXAMPLE 12.10

A two-lane highway has a longitudinal grade of 3.5% and a transverse grade (camber) of 2.5%. The design precipitation rate is 40 mm/h. If the permeability of the drainage layer is 250 m/day, find the thickness of the layer.

Assume that infiltration is 30% of the storm water precipitation, width of the lane is 3.5 m, width of the shoulder is 2.5 m, slope of the embankment is 2(H):1(V) and the drainage layer is placed 500 mm beneath the pavement surface.

Solution

Figure 12.24 illustrates the present problem. AB and EF represent the same camber slope of 2.5% and AE represents the longitudinal slope of 3.5%.

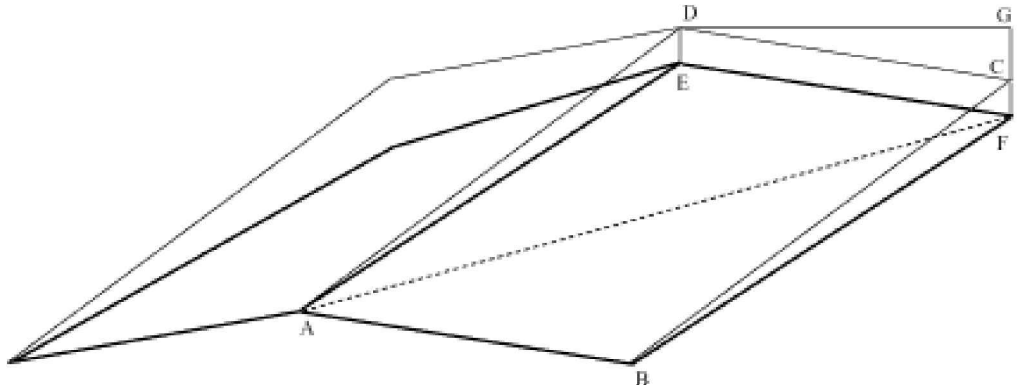


Figure 12.24 A pavement section with transverse and longitudinal slope.

Length of AB = lane width + shoulder width + extra width at the level of drainage layer due to side slope

$$= 3.5 + 2.5 + 2(0.50) = 7 \text{ m}$$

$$\text{Length of AE} = 7 \times \frac{0.035}{0.025} = 9.8 \text{ m}$$

The length of AE is calculated with the help of physical independence of velocity, which means that the time taken for water to flow along AB should be the same as that required to flow along AE. However, the actual flow occurs along AF.

$$\text{Thus, the length of AF} = (\text{AE}^2 + \text{EF}^2)^{1/2} = 12.04 \text{ m}$$

$$\text{Elevation drop at F with respect to A} = \text{CF} + \text{GC} = 9.8(0.035) + 7(0.025) = 0.518 \text{ m}$$

$$\text{Thus, gradient of AF} = \frac{0.518}{12.04} = 0.043$$

$$\text{Now, the run-off, } Q, \text{ for unit width} = \frac{40 \times 0.30 \times 24}{10^3} \times 1 \times 12.04 = 3.467 \text{ m}^3/\text{day}$$

Using the Darcy's equation (refer Eq. (10.7)),

$$3.467 = 250 \times 0.043(h \times 1)$$

Therefore,

$$h = 0.322 \text{ m} \approx 320 \text{ mm}$$

Thus, the required thickness of the drainage layer is about 320 mm.

Gradation design of the filter layer

The following criteria should be satisfied [89] for a good filter layer. These criteria are based only on experience.

$$\frac{D_{15} \text{ of filter layer}}{D_{15} \text{ of subgrade}} \geq 5 \tag{12.33}$$

$$\frac{D_{15} \text{ of filter layer}}{D_{85} \text{ of subgrade}} \leq 5 \tag{12.34}$$

$$\frac{D_{50} \text{ of filter layer}}{D_{50} \text{ of subgrade}} \leq 25 \tag{12.35}$$

where D_x is the particle size corresponding to per cent passing of x. The filter layer should extend to the full width of the pavement. The following example explains how the gradation of a filter layer can be decided.

EXAMPLE 12.11

Following is the particle size distribution of subgrade soil. Design a suitable aggregate gradation for a drainage layer to be laid above the subgrade.

Particle size (mm)	4.75	2.36	1.18	0.60	0.30	0.15	0.08	0.01
% less than or equal to	100	88.4	68.2	49.7	34.4	15.2	8.1	1.0

Solution

The particle size distribution of the subgrade soil is plotted in Figure 12.25.

From Figure 12.25, D_{15}, D_{50}, and D_{85} of the subgrade soil layer are found to be 0.15, 0.60, and 2.10 mm respectively. From Eq. (12.33), the D_{15} of the aggregates in filter layer should be greater than or equal to $0.15 \times 5 = 0.75$ mm. Thus, the point A is marked in Figure 12.25 corresponding to 15% passing.

From Eq. (12.34), the D_{15} of filter particles should be less than $2.10 \times 5 = 10.5$ mm. Thus the point B is marked on the plot.

Similarly, from Eq. (12.35), the D_{50} of the filter aggregates should be less than $0.60 \times 25 = 15$ mm, and the point C is marked on the plot.

Therefore, any gradation which is well within the boundaries, marked as A, B, and C, will satisfy filter criteria. A tentative gradation of the aggregates for filter layer is drawn in Figure 12.25.

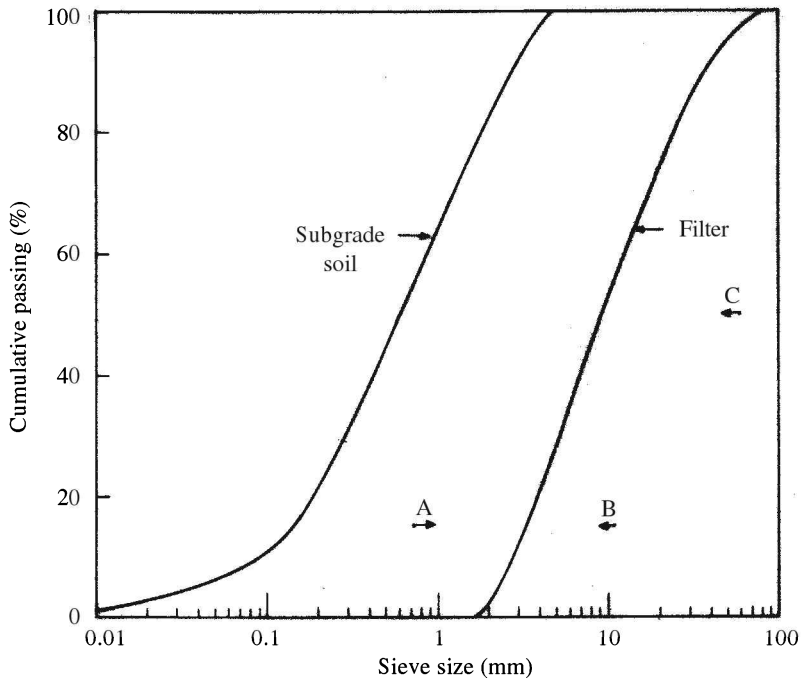


Figure 12.25 Gradation design for a filter layer.

12.6.3 Further Discussion on Drainage Considerations

As a preventive measure for the new roads, the pavements should be constructed at least 0.6 to 1 m from the water level (or high flood level) [89]. When soft material is used in sub-base and there is a chance of materials getting crushed during rolling, making it relatively impermeable and not fulfilling its purpose as drainage layer, it is recommended [89] that 100 to 150 mm of crushed stone base be provided instead.

Curbs, gutters, or turf covers are provided in suitable locations to prevent soil erosion due to water flow. Special drainage provisions are made if the road is (i) in the zones where the level is below the groundwater table, (ii) on hill slopes, or (iii) on embankments. In such situations, drainage blanket, perforated pipes, cross and longitudinal network, and geotextiles are used through proper design. Instead of perforated pipes, aggregate drains can also be used, if properly designed.

12.7 FROST DAMAGE IN PAVEMENT DESIGN

To counter the possibilities of frost damage, two approaches are suggested [106] for use in design of pavements. They are as follows:

1. To provide adequate thickness to the pavement so that frost cannot penetrate and damage the pavement. Experimentally observed design curves are

available between the freezing index and the depth of frost penetration. The *depth of frost penetration* is the lowest level at which freezing is observed to occur. This is also called *frost line*. The thickness of the pavement should be more than the frost penetration depth.

2. To provide adequate thickness to the pavement such that it can sustain the traffic load during the thawing phase, when the strength of subgrade is lower (because of melting down of ice lenses). Design charts have been evolved where reduced subgrade strength during the thawing phase and the corresponding pavement design thicknesses are plotted.

The following are some of the methods suggested for prevention of frost damage in pavements [266]:

(a) *Depth of construction.* Depth of construction more than the frost penetration depth or the depth required at the reduced subgrade strength during the thawing phase, whichever is larger, should be provided.

(b) *Drainage.* Care should be taken to provision a proper drainage system for the pavement. Provisions should also be made for quick disposal of surface and sub-surface water, the details of which have already been covered in the previous section.

(c) *Replacement of frost susceptible soils.* One of the more common methods for minimizing frost damage is to remove and replace frost heaving soil with non-frost-susceptible material. Although this does not necessarily reduce the depth of frost penetration, it improves drainage for the water collected during the thawing phase.

(d) *Chemical treatment.* Chemicals such as salts and sodium silicates, when injected into a frost-susceptible soil, lower the freezing point of the soil moisture and thus ice lens formation is retarded. It is a temporary method as the chemicals slowly tend to leach away.

(e) *Insulation and membrane courses.* Use of insulating material between the frost-susceptible soil and the pavement structure effectively controls the depth of penetration and the rate of heat dissipation. Natural materials like, wet sand, or man-made materials such as, Styrofoam and foamed sulfur may be used for this purpose.

IRC: 37–2001 [89] recommends that in case frost damage is expected, the depth of pavement should not be less than 450 mm, though structural design considerations may recommend a smaller design thickness. In addition, the materials used for construction, should be frost resistant.

12.8 OTHER DESIGN CONCEPTS

This section briefly covers a few other topics related to pavement design. They are:

1. Design of bituminous pavements with cemented base
2. Considerations in stage construction
3. Airport pavement design
4. Reinforced concrete pavement design
5. Full depth bituminous pavement design
6. Considerations in pavement shoulders.

12.8.1 Bituminous Pavement with Cemented Base/Sub-base

Use of conventional aggregates is sometimes a costly proposition when good quality aggregates are not locally available. In such situations, cemented bases and sub-bases can be constructed at lower cost by using lime, fly-ash, clay, Portland cement, low grade marginal aggregates, and so on. This section describes in brief, the design methodology of bituminous pavements with cemented base.

No guideline has been brought out by the Indian Roads Congress for structural design of bituminous pavements with cemented base or sub-base. The available ones [89, 229, 228, 191] for the use of cemented material in pavement structure, may not be applicable for structural design of pavements for high traffic levels.

The following methodology can be adopted for mechanistic design of pavements with cemented base. This methodology is similar to that proposed by Austroads [182], South African code [236, 52], and other researchers [178, 162].

(a) The original value of E is taken as the initial E value of the cemented material. Tensile strain ε_t is computed at the bottom of the cemented base. In the vicinity of the thermal cracks of the layers, the stress value will be about 1.5 times [162] the computed stress for the uncracked section. The fatigue life of the cemented base is then calculated from the fatigue equation of the material. It may be assumed that the flexural fatigue of bituminous surfacing does not govern the pavement performance as long as the cemented layer is intact.

(b) When the life of the cemented material is completely exhausted, it may fully crack, and at this stage it can rather be considered as a granular layer with elastic modulus of only 10% of the original [236]. The fatigue life of the bituminous mixture, estimated from the maximum tensile strain ε_t at the bottom of the bituminous layer, therefore, becomes the criteria for the pavement longevity for this phase and the corresponding life is calculated.

(c) The design may be based on adjusting the layer thicknesses in such a way that the sum of these two fatigue lives becomes equal to the design life. The flow-chart of the proposed design methodology is explained in Figure 12.26.

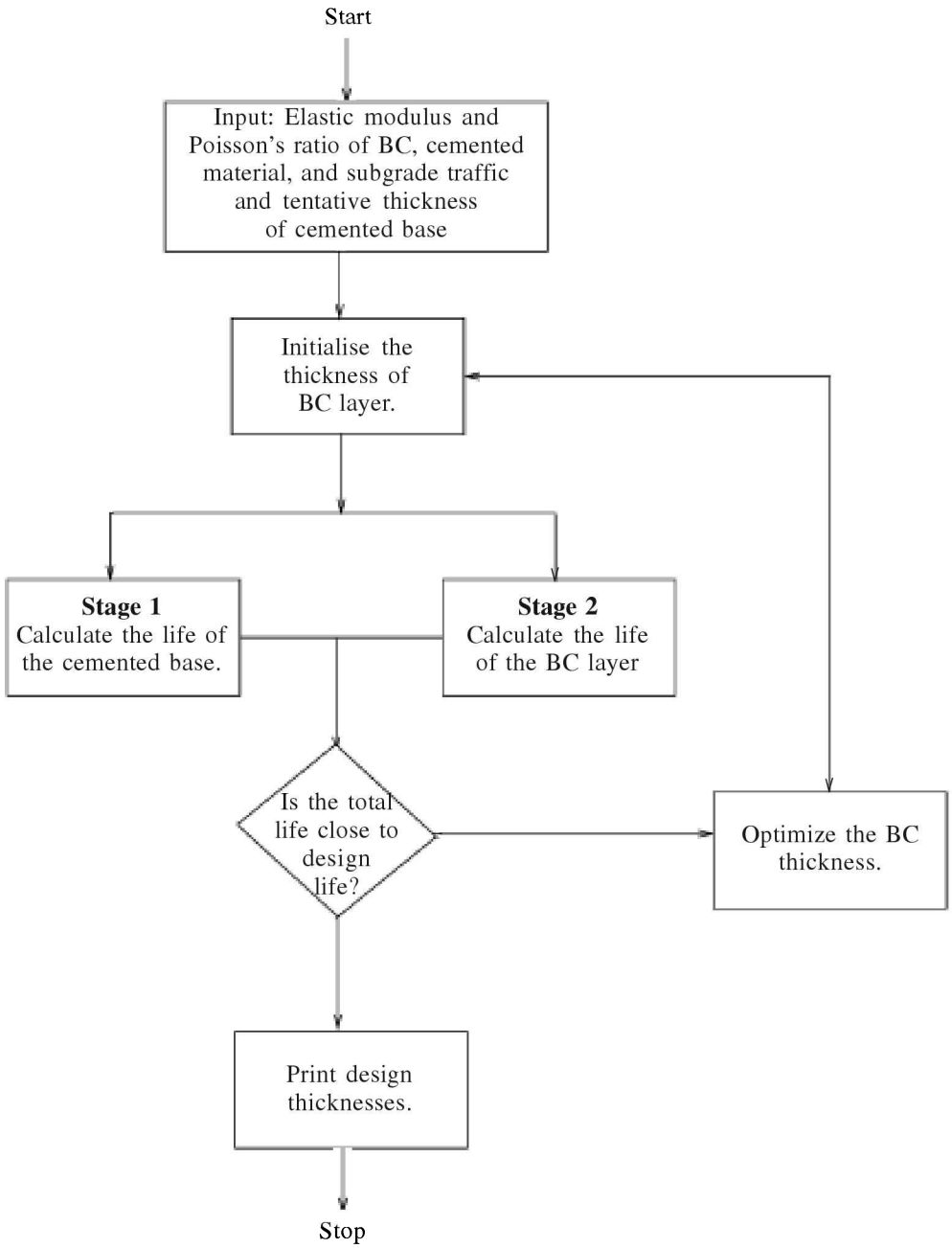


Figure 12.26 Flowchart for design of bituminous pavements with cemented base.

While selecting a cemented base, it is necessary to make provisions for the maintenance associated with shrinkage and thermal cracking. Shrinkage cracks are a natural characteristic of cemented materials. When a cement-treated material tries to shrink, friction is developed between the treated layer and the underlaying material and consequently, internal stresses are induced. The stresses eventually exceed the tensile strength of the treated material (which is very low), and cracks start appearing. Some cracks appear after a few days whereas others may appear up to four months later [236]. The initial cracks may be rehabilitated by sealing [52]. Thus, provisions should be made for keeping the minimum necessary thickness of the cemented sub-base.

12.8.2 Stage Construction

Implementation of road projects may require a large outlay for providing design thicknesses of different layers of bituminous pavement for the whole of the design period. In case, full funding is not available initially, lower thicknesses are provided (compared to what has been designed for the estimated design period) for some chosen layers in the first stage (say, Stage I) and additional thicknesses are added in the second stage (say, Stage II) before serious distress starts appearing. This is known as *stage construction*. This type of strategy is also desirable for situations where the traffic data cannot be correctly estimated for new roads. After Stage I construction, when the pavement is operational, actual traffic count can be made and additional thicknesses to be laid over the existing layers can be estimated precisely. AASHO [1] road test has indicated that the performance of pavement, if constructed in stages, is better than a one-time pavement for the same period of design life. Three possible pavement conditions may arise after the expiry of Stage I, which are:

(a) The pavement may become completely cracked

(b) The pavement may deteriorate as predicted in the design

(c) There may be little distress in the pavement because of low traffic on the road.

Any additional thickness of the bituminous layer required for the Stage II depends on the pavement condition noticed after the expiry of Stage I. It is essential to estimate the in-situ elastic modulus of the existing bituminous layer at this stage for finding out the required additional thickness for Stage II construction.

Stage construction is desirable for better performance of pavements, but studies [46] indicate that this results in higher thickness requirement of bituminous layers for a given design life, than it would have been if laid as one-time construction. From mechanistic point of view, stage construction is equivalent to putting the necessary overlay for extension of the life of the pavement. Overlay design considerations have been discussed in Section 14.5.2.

Figure 12.27 depicts a stage construction chart based on the provisions of IRC:37–2001 [89], for a total design period of 100 msa. The line indicated in the diagram shows,

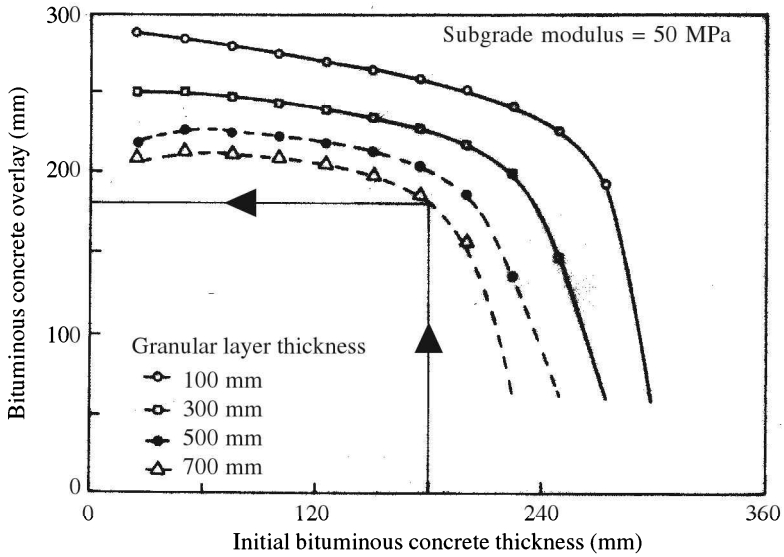


Figure 12.27 A stage construction design chart based on IRC:37–2001.

that if 180 mm Stage I BC layer is put over the 700 mm granular layer, another 180 mm BC layer would be needed after the expiry of Stage I (when the pavement is assumed to have completely failed), such that the total life of the pavement (including both Stage I and Stage II) is 100 msa.

12.8.3 Airport Pavement

In principle, the design methodologies of highway pavement, airport pavement, and runway are similar, if the mechanistic pavement design is chosen. The differences between the highway and the airport pavements are: (i) the repetitions in runway are smaller in number compared to those in highway, (ii) the aircraft wheel configuration is different from that of highway vehicles, and (iii) the pavement design in airports must include the criterion of standing aircraft. On the other hand, except for parking or loading/unloading cases, the highway pavements are designed for running vehicles. For the design of airport pavements, the stresses in the pavement are evaluated due to a given type of aircraft and its wheel configuration and the design thicknesses are checked with the help of the cumulative damage principle. Separate design charts are available for the design of airport pavements and runways as developed by Fedual Aviation Agency (FAA), Asphalt Institute, PCA, etc.

12.8.4 Reinforced Concrete Pavement

Reinforcement in the concrete slabs, as mentioned, is provided only for countering the shrinkage due to temperature and moisture changes. The shrinkage stress is maximum

in the middle and less at the edges or corners. The requirement of steel reinforcement is calculated on the basis of maximum force that can overcome the frictional force between the concrete slab and the layer just below it. The following formula is recommended by the Indian guidelines [91] for calculation of the area of the steel required.

$$A_s = \frac{d \times f \times W}{2 f_{sw}} \tag{12.36}$$

where d is distance between the free transverse joints (for longitudinal steel) or free longitudinal joints (for transverse steel).

Equation (12.36) is similar to Eq. (12.27) except that the area of steel here is half of that obtained from Eq. (12.27). The reinforcement provided in the concrete pavement does not contribute towards the flexural strength, therefore, its position in the concrete slab is not important. Generally, reinforcement is placed 50 mm beneath the surface [91].

12.8.5 Full Depth Bituminous Pavement

Full depth bituminous pavements (FDBP) are the pavements where no granular layer is used. The fatigue and rutting failures are taken care of by the bituminous layer. Tentative thickness of the bituminous layer is assumed, and checked for fatigue and rutting strains. Both the strains should be closely equal to the values allowed. The thickness of the bituminous layer is accordingly adjusted. Figure 12.28 shows a flowchart for the design algorithm of FDBP.

For obvious reasons, the FDBP has a higher construction cost, however, it has some advantages over the conventional construction. The FDBP has no granular layer, and it being almost impermeable, the chances of getting damaged due to water percolation and improper drainage are reduced. The FDBP is less affected by frost. Also, the construction time is less compared to that of the normal pavement. The FDBP is not a common practice in India and as a result, no design specification for FDBP is available.

12.8.6 Pavement Shoulders

Pavement shoulders, though not intended to carry traffic directly, serve a number of purposes. The shoulders are sometimes used as footpath or bikeway. They sometimes carry traffic in special situations, give psychological comfort to the drivers, and accommodate vehicle parking. When a construction activity is underway on the main carriageway, shoulders are used as temporary driving lanes. Asphalt Institute [237] recommends that design equivalent axle load repetitions in the shoulder may be taken as 2% or greater than those of the actual carriageway.

Shoulders should be well-shaped and special attention should be paid to keep them in shape during the maintenance operations. This is necessary for proper and quick

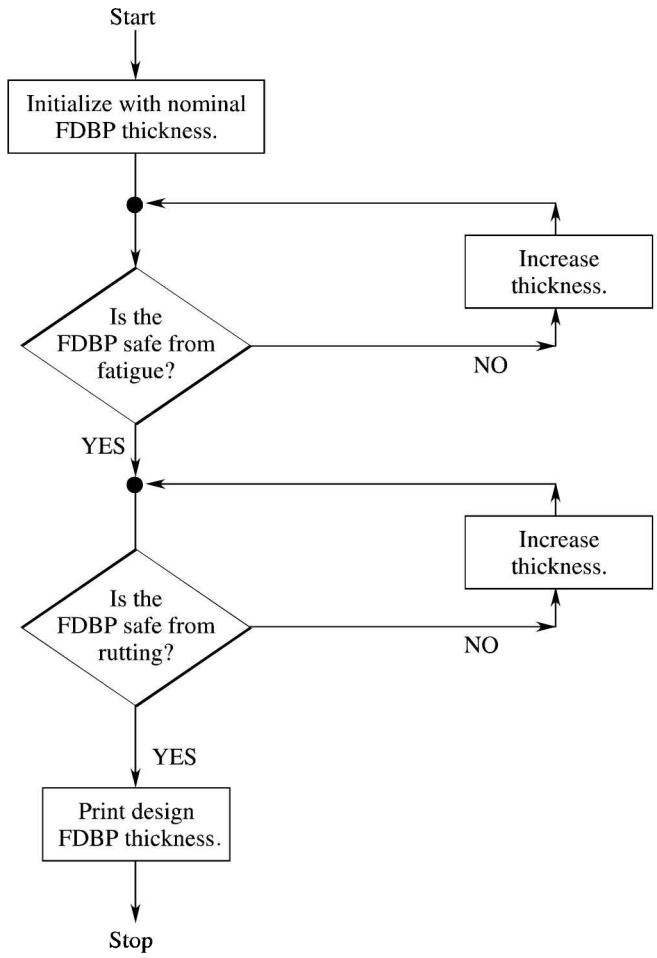


Figure 12.28 Flowchart for the design of full depth bituminous pavement (FDBP).

drainage of the surface run-off. The cross-fall of the shoulder should be at least 0.5% steeper than the slope of the carriageway subject to a minimum of 3% [79]. Paved shoulders prevent water percolation into the subgrade of the pavement, as well as give lateral support to the base and surface courses.

EXERCISES

1. A 100 km road is to be designed as per the CBR method. The CBR test has been conducted at every kilometre and different CBR values are obtained at each point. How do you choose the design CBR value? Justify.
2. As a pavement designer, what are the various vehicle parameters you would be interested in?

3. What is an equivalent single-axle load? How can it be determined?
4. How is the axle load of a moving wheel measured? Where, in pavement design, is this survey data used?
5. What is the difference between Load Safety Factor (LSF) and VDF? Can LSF be less than one? Justify. Can Lane Distribution Factor (LDF) be less than one? Justify.
6. Mention the various steps involved in mechanistic pavement design of bituminous pavements as per IRC:37–2001.
7. Explain the principle of dowel bar design in concrete pavements.
8. Between fatigue and rutting design curves, which slope is steeper and why?
9. Can we apply the Miner's cumulative fatigue damage principle to the design of bituminous pavements as well? Explain.
10. What is Structural Number (SN)?
11. Design thickness of a parking area should be more than that of an in-service highway where vehicles are moving. Justify your argument for or against this statement.
12. How do the corner, edge and interior stresses vary due to (i) temperature and (ii) due to load? Where and when is the most critical situation found?
13. Why is it that, during the day time, the stress is tensile at the bottom of a concrete pavement.
14. In a certain design process, the structural design of a proposed new bituminous pavement was by mistake performed using the concrete pavement design. Comment on the conceptual mistakes, if any, made.

13 Highway Construction

13.1 INTRODUCTION

In this chapter, a brief history of highway construction is first presented together with the merits and shortcomings of the practices of road construction followed in earlier days. Next, the common equipment used in present-day highway construction is introduced and the techniques associated with various stages of highway construction are briefly described before discussing the construction of individual layers. Weather-related factors affecting highway construction, earthwork, and soil stabilization are also discussed. Finally, the chapter introduces the various layer specifications of bituminous and concrete pavements, and discusses the construction process of these layers. A few other relevant topics are also introduced.

The layer specification and construction techniques vary from country to country. In advanced countries, pavement construction is fully mechanized. In India too, the construction of National Highways, Expressway corridors is now being done by sophisticated automatic construction equipment. Mechanized construction results in better uniformity and quality of construction through online control of construction parameters. Discussions on the specifications and construction techniques in the preceding chapters are mainly based on the MORT&H Specifications for Road and Bridge Works [215]. There are other specifications too, laid down by Indian Roads Congress, Border Roads, etc. Also, the technical terms used to identify a layer may be different in different countries, but some of them may basically be the same. For example, slurry seal, surface dressing, and seal coat represent a common practice in various countries, yet sometimes called by different names.

13.2 HISTORY OF ROAD CONSTRUCTION

Roads began as pathways, possibly created by the large animals who pushed aside the surrounding vegetation on their way. Routes had been formed by humans by about 30,000 B.C. Oxen were initially used for travelling longer distances by road and the use

of mules and donkeys came later around 3000 B.C. Horses started being used from about 2000 B.C. Wheel was first developed in Mesopotamia around 5000 B.C. [143].

The oldest road was constructed from Knossues to Leben, through the mountains of Crete, in 2000 B.C. It reportedly had elaborate longitudinal drains, a 200 mm base-course of sandstone in a clay-gypsum mortar, and a 14 m running surface of basalt blocks. The Indian road technology reached an advanced stage by about 1000 B.C. with the wide use of brick-paved roads and sub-surface drainage system. A special feature of Greek and Roman roads was that they were laid well above the ground level and thereby had a good provision for drainage. The roads were straight, cambered, and composed of multiple layers which were well compacted and contained lime as the stabilizing agent. The roads had carefully constructed longitudinal drains on both sides and used different foundations for different types of soil conditions [202]. Lime mortars were used by the Greeks but the Romans added volcanic pozzolana to produce a stronger and more durable mortar. Roman roads were 4–10 m wide with an average width of 5.5 m [143]. Figure 13.1 shows a typical Roman road cross-section. The strong and large size aggregates were put at the bottom and the stone size gradually decreased on moving towards the top. Romans also used different construction techniques for the primary and secondary road system and accordingly defined them [202].

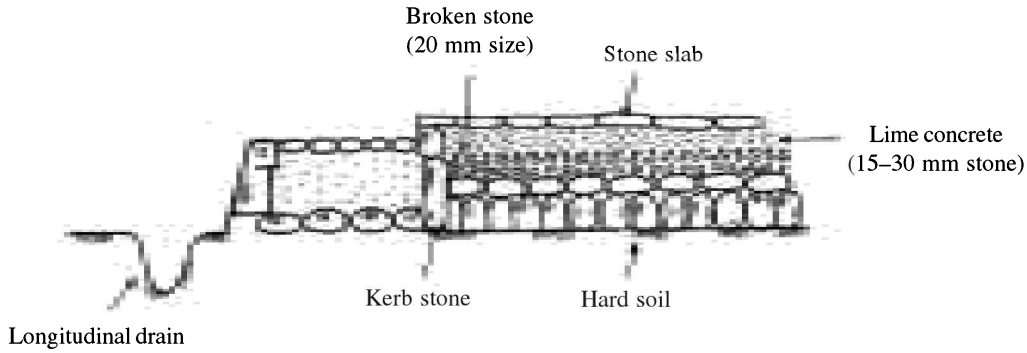


Figure 13.1 A typical Roman road cross-section [143, 202].

13.2.1 Trésaguet Pavement

Trésaguet, a French Engineer, used large pieces of quarried stone at the lower level of the pavement. Smaller pieces of broken stone were then compacted into the space between the larger stones to produce a level surface. The running surface was made with a layer of 25 mm-sized broken stone. The pavement structure was placed in a trench in order to keep the running surface levelled with the surroundings, but this in turn created drainage problems of its own. The random use of large stones was also a problem which resulted in bumpy surface due to differential settlements [143]. A typical pavement cross-section as per Trésaguet design is presented in Figure 13.2.

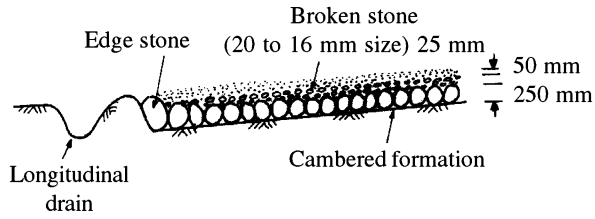


Figure 13.2 A typical cross-section of Trésaguet pavement [143, 202].

13.2.2 Telford Pavement

Telford's construction methodology was somewhat similar to that of Trésaguet except that he kept the formation surface levelled and the upper surface cambered gradually by the use of proper size stones. The stones were placed carefully so as to ensure good aggregate interlock. Telford pavement required more masonry work but it was stronger than the Trésaguet pavement. The running surface consisted of 20 mm gravel of 150 mm thickness. Telford used a raised pavement structure above the ground level wherever possible. He also provided cross-drainage. Drainage principles were known to Romans, but remained forgotten for a long time. Telford rediscovered them. Figure 13.3 shows a typical cross-section of a Telford pavement.

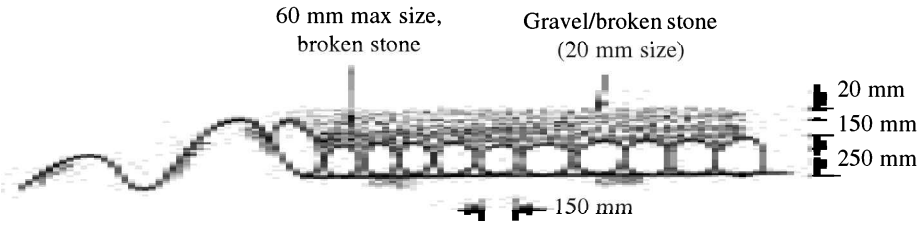


Figure 13.3 A typical cross-section of Telford pavement [143, 202].

13.2.3 McAdam Pavement

As the narrow iron-tyred wheels caused major distress to the pavement surfacing, McAdam looked for alternatives from the then current methods of road construction and maintenance. By observing construction and design of various roads, McAdam realized that 250 mm layers of well-compacted broken angular stone would have the same strength as that of pavements with large foundation stones. McAdam used the surface stone of size smaller than the tyre width, and thereby a good running surface was obtained [143]. Some of the important ideas which evolved out of McAdam methodology are:

(a) The use of angular aggregates helped to develop strong interlocking.

(b) Well-graded material was used which enhanced impermeability and ease in compaction.

(c) It was strongly felt that an impermeable, durable binder was needed for road construction to keep the surface intact.

Figure 13.4 shows a typical cross-section of a McAdam pavement.

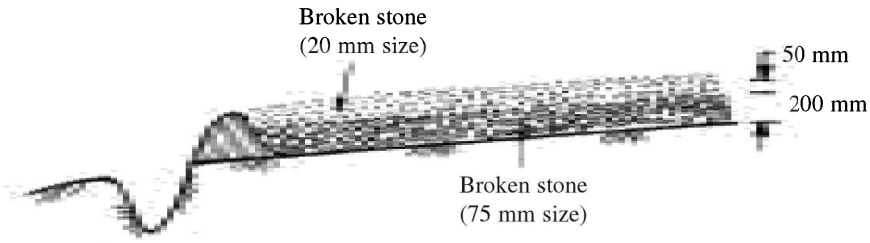


Figure 13.4 A typical cross-section of a McAdam pavement [143, 202].

13.3 EQUIPMENT USED IN HIGHWAY CONSTRUCTION

Various equipment used in highway construction are discussed in this section which include earth moving equipment, aggregate spreader, roller, nuclear gauge, road broom, binder sprayer, and paver finisher. These are discussed in the following.

13.3.1 Earth Moving Equipment

Tractor or crawler is the basic equipment used for earth moving activities, to which various attachments may be fixed for specific purposes. Bulldozer is another equipment where a movable steel blade is attached to the front of a tractor. Bulldozers are used for clearing the way along the construction line and also for moving earth. Another attachment to tractor is 'loader', in which a bucket is used for picking up, transporting, and depositing soil [41]. 'Hoe' is another type of attachment used for digging pits or trenches [41]. Rotary tiller is an equipment meant for field pulverization of subgrade soil. Other items of equipment which serve the same purpose are disc harrows, crow bars, or even ploughs.

13.3.2 Aggregate Spreaders

Aggregate spreaders or gritters are used to ensure uniform spreading of aggregates over the construction surface. Three types of aggregate spreaders [151] are discussed in this section.

Truck-mounted spreaders

In a truck-mounted spreader or a fantail spreader, the aggregate carrier is hinged to the body of the truck. The operator can change its slope and also the gate opening to control

spread of aggregates. Vertical fins are provided so as to ensure uniform spread of aggregates over the pavement width. The uniformity of spread largely depends on the skill of the operator-cum-driver.

Truck-propelled spreaders

In the truck-propelled spreaders, aggregates are moved onto the rotating belt, and through an adjustable gate, aggregates are spread uniformly over the surface. When the truck is empty, the spreader is disconnected and attached to a loaded truck.

Self-propelled spreaders

Self-propelled spreaders consists of two hoppers, attached at the front and rear of a pneumatic-tyred tractor [187]. Aggregates are fed to the hoppers through screw feed and the rate of spread can be controlled through an adjustable gate.

13.3.3 Rollers

Five types of rollers are described here, namely the smooth-wheeled rollers, pneumatic-tyred rollers, sheepfoot rollers, grid rollers, and vibratory rollers.

Smooth-wheeled rollers

Smooth-wheeled rollers or drum rollers can compact all types of soil, except rocky soil. The disadvantages of a smooth-wheeled roller are that (i) the level of compaction is sometimes not satisfactory, because of its large contact area and consequently low pressure and (ii) the weaker aggregates tend to be crushed during compaction.

Pneumatic-tyred rollers

In pneumatic-tyred rollers a number of tyres are placed close and parallel to each other. This is suitable for coarse grained soils (with 4–8% passing IS75 micron sieve). Depending on the type of material, a pneumatic-tyred roller can compact faster and with fewer passes than a smooth-wheeled roller. Figure 13.5 shows a pneumatic-tyred roller.

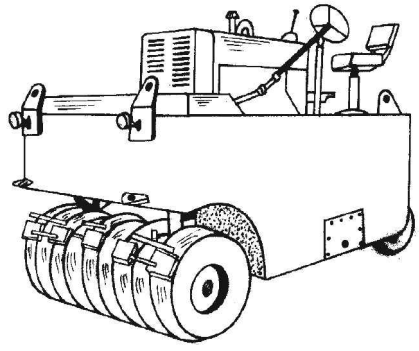


Figure 13.5 Pneumatic-tyred roller.

Sheepfoot rollers

The sheepfoot roller or padfoot roller has some protrusions on its drum which penetrate into the soft soil and compact it by kneading and tamping. The padfoot roller has relatively large footprints. The weight of the roller can be adjusted by changing the amount of water poured into the rolling drum. Successive passes of the sheepfoot roller decrease the depth of penetration of the protrusions and this is known as 'walk-out'. The measure of walk-out gives some idea regarding the level of compaction attained. The sheepfoot roller is suitable for fine-grained soil and not for granular soil. The sheepfoot roller can effectively compact soil over a wide range of moisture contents than a rubber-tyred roller can. Blending of materials is assisted by this compaction method, and it also eases the moisture control [99]. Figure 13.6 shows a sheepfoot roller.

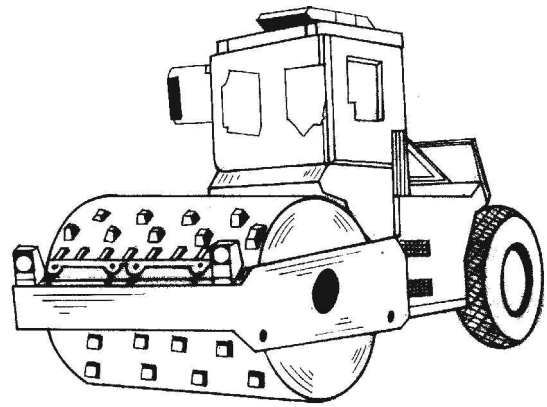


Figure 13.6 Sheepfoot roller.

Grid rollers

In grid rollers, the drum is covered with heavy steel grid. The grid produces high contact pressure which helps to avoid situations, such as plastic wave formation due to shear deformation of soil. Grid rollers are suitable for compaction of granular particles and help in breaking and rearranging the particles. However, grid rollers cannot work effectively on clayey soil as the soil gets clogged within the grid [99].

Vibratory rollers

In these rollers, a vibratory motion along the vertical direction is induced along with the rolling operation. The amplitude of vibration is of the order of 1 or 2 mm. These vibratory drums are isolated by means of shock absorbers from the frame on which the operator sits. This vibratory effect helps in better compaction, specially that of granular materials. The extent of compaction caused by the vibratory roller depends on:

(i) Static weight
(ii) Vibratory mass

(iii) Frequency and amplitude of vibration

(iv) Rolling speed

Figure 13.7 shows the photograph of a vibratory roller under work.

Figure 13.7 A vibratory roller under work.

13.3.4 Nuclear Gauge

Nuclear gauges are used for field density measurement. The scatter of gamma rays (source: cesium isotope) and neutrons (source: beryllium isotope) from the nuclear gauge give a measure of the bulk density of soil and moisture content respectively. Calibration curves are used where scatter level, bulk density, and moisture content are plotted for known values. Nuclear gauges measure the compaction level quickly but adequate precautions must be taken to avoid radiation hazards. They also need frequent calibration.

13.3.5 Road Brooms

Road brooms are used for resurfacing on the existing surface. The road brooms are used to remove dirt, mud, dust, and other deleterious particles from the existing surface, such that the adherence of the surface to be laid with the existing surface, becomes strong. In the rotary road broom, a circular broom suspended from a frame rotates in a horizontal plane. The position of the broom can be adjusted from the control installed in the vehicle on which it is mounted. The bristles of the broom are made of steel, nylon, or fibres and sometimes a high-pressure air jet is used to serve the same purpose.

13.3.6 Sprayers or Binder Distributors

Sprayers or binder distributors are used when a binder (cold or hot) is required to be sprayed uniformly. Sprayers can be of hand-spray type or of vehicle mounted multi-

nozzle type. A sprayer can either be self-propelled or towed. A good sprayer is expected to serve the following purposes [68]:

(a) Uniformity in rate of application of the binder
(b) Adjustbale application rate depending on the requirement of a particular type of construction
(c) The area covered, in which the binder is sprayed, should be in accordance with the area covered by the aggregate spreader

Depending on the rate of application of the binder, a sprayer can be categorized as:

(a) Constant rate of spread distributor
(b) Constant volume distributor
(c) Constant pressure distributor

The components of binder distributors are spray bar, spray jet (two types—slotted jet or whirling jet), binder pump, air compressor, binder tank, and burners. In modern-day sprayers, the vehicle speed, the application rate, the binder distribution along horizontal and vertical directions, and the binder temperature are controlled through a microprocessor.

13.3.7 Paver Finisher

Paver finishers have the arrangement for multipurpose jobs related to pavement construction. The main components of the paver finisher are:

(i) A loading hopper and a suitable distribution mechanism
(ii) A compaction and vibrating arrangement
(iii) A mechanism for the construction of a smooth surface finish, free from surface blemishes [215]. Figure 13.8 depicts a paver finisher in operation.

Figure 13.8 A paver finisher in operation (*Courtesy: Lucknow, NHAI*).

13.4 STAGES OF CONSTRUCTION

This section describes some of the stages of construction activities which are common for various types of constructions, such as mixing, compaction, curing, and so on.

13.4.1 Pulverization

Pulverization is needed in certain cases when some materials (such as soil, lime, cement, or fly-ash) are to be mixed together for stabilization purposes. Standard guidelines specify the particle distribution of these materials and if the particle size of the materials available in the construction site is larger than what is specified, pulverization is recommended. The degree of desired pulverization is confirmed through field tests.

13.4.2 Mixing

Mixing can be done in two ways—either by using the in-place method or by using the stationary plant method. As the names suggest, in the former method, mixing is carried out at the location where laying is to be done and in the latter method, mixing is done separately in the mixing plant and then transported to the site of laying. Manual mixing is more labour intensive, slow, and also the uniformity of mixing is poor. Manual mixing is generally not recommended except for small projects.

13.4.3 Binder Spraying

The binder is spread uniformly by a binder distributor as per the stipulated spraying rate. The binder spray may have double or triple overlap area depending on the requirement. The overlap can be cross-checked by closing some of the nozzles as shown in Figure 13.9 and accordingly the height of the spray bar can be adjusted.

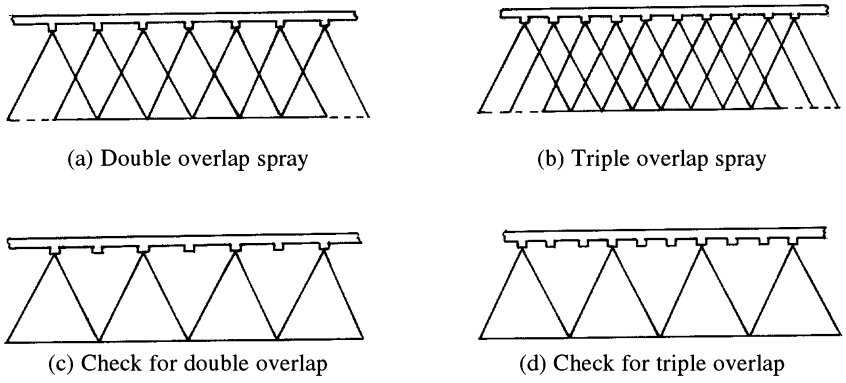


Figure 13.9 Checking of double and triple overlap.

13.4.4 Rolling and Compaction

Proper compaction is needed at every stage of highway construction to achieve the specified density and thus, the strength. Compaction is done by tampers and various types of rollers.

Rolling of a section starts at the lower edge and proceeds towards the upper edge with a specified overlap with the earlier rolling area. Thus, for pavement with normal camber, rolling should start from the outer edge of the pavement towards the central line. On the other hand, for a superelevated pavement stretch, rolling should progress from the inner edge to the outer. The grade and the cross-slope should be properly maintained during rolling. Layer by layer compaction is always recommended for achieving better strength and stability. Rollers for bituminous construction should not move at a speed more than 5 km/h [215] and are not permitted to stand on the part of the pavement which is not fully compacted. Sometimes moistening the wheels with water is preferred to avoid adherence of the particles to the surface of the roller.

The densities of various types of construction are checked by the sand replacement method, balloon method, Hilf rapid method, drive cylinder method, nuclear gauges, and so on. For each type of construction, the achievable density is mentioned by the code of practice, or sometimes the directions are issued by the Engineer-in-Charge. The compaction level of bituminous mixes is given in terms of percentage fraction of the Marshall mould density, and for subgrade soil it is specified in terms of the percentage of the maximum dry density. Thus, for checking compaction of the subgrade soil, determination of field moisture content is also another parameter which needs to be measured in the field or in laboratory.

13.4.5 Curing

For most of the types of highway construction, a curing time is necessary for the setting of a particular layer. For example, for construction with bituminous emulsion, requisite time by which emulsion breaks, adheres to the aggregate surface, and water is evaporated away should be provided. In some other cases, such as stabilization of soil with lime or cement, moisture is not allowed to evaporate. In that case, bituminous coating is sometimes applied at the top to prevent moisture loss, or water is sprinkled externally at regular intervals. For curing of Dry Lean Concrete or Cement Concrete pavement either (i) the liquid curing compounds are spread which prevent evaporation of the water inside the construction, or (ii) the surface is covered by gunny bags/hessain etc. and kept wet, or (iii) the ponding method is adopted, if necessary. For hot mix construction too, some curing time is essential, during which temperature comes down, bitumen gains its strength, and the whole mix becomes hard and strong, and ready for taking traffic loading, or can act as a base for the next layer.

As a general rule, during curing, no construction equipment should be allowed to move over the finished surface as it may cause unwanted depressions on the surface.

13.5 SEASONAL LIMITATIONS OF PAVEMENT CONSTRUCTION

Pavement construction is done in the open, therefore, the quality of construction becomes dependent on weather. Extreme weather is not suitable for pavement construction. For example, pavement construction is not advised to be carried out when the weather is foggy, rainy, windy, or when the temperature in the shade is less than 10°C [215]. For hot mix and cutback construction, the surface should be dry. However, for construction with bitumen emulsion, the surface needs to be slightly damp though not wet.

13.6 EARTHWORK

Earthwork consists of earth cutting and/or earth filling. If any project involves both cutting and filling activities along the project stretch, the lift and haul should be properly optimized such that earth removed from one place could be used most economically for filling some other place, nearby. The following parts of this section discuss some aspects of earthwork.

13.6.1 Cleaning and Grubbing

The first step towards preparation of subgrade [215] involves removal of bushes, roots, grass, weeds, and top organic soil not exceeding 150 mm. This can be done manually, or a crawler/pneumatic-tyred bulldozer of adequate capacity can be used for clearing purposes. Existing roots of trees should be removed by digging up to their bottom. Existing old structures, such as culverts, bridges, fences, and others which fall within the right of way also need to be removed [215].

13.6.2 Excavation for Road and Drain

After the alignment of the road and its side drains is set, excavation of earth is carried out wherever necessary. Excavation is also necessary to collect soil from the borrow pits for filling work. Adequate precautions need to be taken during excavation against soil erosion, water pollution, etc. Figure 13.10 shows an excavation in progress.

13.6.3 Embankment Construction

Miscellaneous backfills such as soil, moorum, gravel, fly-ash [75], and so on, can be used for construction of a highway embankment. For National Highways and Expressways, the maximum dry density should not be less than 1.75 g/cc [89]. Clay with liquid limit exceeding 70 and Plasticity Index (PI) exceeding 45 is considered unsuitable for embankment construction [215]. The top 500 mm of the filling soil should have the

Figure 13.10 Earth excavation in progress (*Courtesy: Mr. Pinaki Roychowdhury).*

acceptable free swelling index. The size of the coarse material in embankment and in subgrade should not exceed 75 mm and 50 mm respectively [215]. Materials for embankment construction may have to be collected from borrow pits, however, use of borrow pits close to the embankment should be discouraged.

13.6.4 Replacement of Weak Soil

If the subgrade soil is weak, the affected layer should be replaced, up to a specified depth, by a better quality of soil with a higher CBR value. "Japan Road Association" recommends [151] that in such a case, care should be taken about the following two aspects:

(a) The remaining part of the weak subgrade should not be disturbed.

(b) The replaced soil should be thoroughly compacted.

13.7 STABILIZATION OF SOIL

Stabilization of soil discusses various methods employed for modifying the properties of soil to improve its engineering performance, both in terms of its strength and durability. Stabilization technique controls the unwanted properties in subgrade soil such as, excessive compressibility, permeability, frost susceptibility, settlement, volume change, and so on. Mixing suitable proportions of coarse and fine-grained material, adding materials having cementing property, adding chemicals, applying electrical current or heat, plugging in voids, consolidation—are few of the various possible

options available for stabilization of soil. The following subsections briefly discuss some of these options.

13.7.1 Mechanical Soil Stabilization

Mechanical soil stabilization is a process which enhances soil strength when soils of different plasticity indices or aggregates of desired gradation, are mixed with the existing material. In this process, the properties of the soil can be improved by

(a) changing the composition of soil by addition or removal of certain constituents and by reducing the void volume of in-situ soil

(b) densification by compaction and subsequent rearrangement of particles.

The principle of mechanical stabilization lies in proportioning the various locally available matter to produce a gradation which is stronger and can be easily compacted. Generally, the Fuller's curve gradation [see Eq. (10.17)] is adopted as a target gradation. Similarly, soils of different plasticity indices are mixed together to achieve the desired plasticity index. The problem of determining the mixing proportions to achieve a desired gradation or plasticity index has already been addressed in Section 10.3.5.

13.7.2 Stabilization with Cementing Additives and Chemicals

When local soil, gravel, moorum, etc. cannot be stabilized effectively by mechanical means, stabilization with cementing additives and chemicals is tried. Soil stabilization with lime, cement, bitumen, and chemicals is described in the rest of the section.

Soil-lime stabilization

Lime treatment is effective for the soil having a plasticity index of 10 or more but the soil must have a fraction of not less than 15% for passing the 425 micron mesh [76]. Generally, clayey soil, black cotton soil, moorum, and alluvial soil can be suitably stabilized with lime. The addition of lime to such soil decreases both the liquid and plastic limits of soil and increases both the CBR value and unconfined compressive strength.

The quantity of lime is obtained by mix design in the laboratory by unconfined compressive testing. The mix design [215] takes into account many aspects simultaneously, such as the type of soil, moisture-density relationship, and purity of lime. Strength enhancement in soil with application of lime is due to base exchange and/or flocculation. The gain in strength is only appreciable when lime is added above lime fixation point, and strength increases as more lime is added. However, at a certain stage, the gain in strength ceases and unreacted free lime in soil is then observed. This means that the optimum lime content is reached for that soil. The tentative values of optimum lime content for soil with different mineralogical compositions are shown in Table 13.1 [76].

Table 13.1 Suggested optimum lime content for soils of different mineralogical compositions

Type of soil	*Suggested optimum lime content*
Kaolinite soil	4%
Illinite soil	8%
Montmorillonitic soil	10%

Optimum lime content can also be determined from the pH value of the resultant mix. Percentage of lime which gives resultant mix pH of 12.4 is identified as the right quantity of lime required for stabilization of the given soil [76].

The strength of the soil-lime mixture is determined by the UCS/CBR method. The samples are prepared at optimum moisture content with lime added according to the value of optimum lime content. If the strength obtained is higher than the desired strength, the final recommendation for lime content can be reduced below the optimum lime content level. The minimum CBR requirement for soil-lime stabilized base is 15–30% depending upon the traffic and type of road while the minimum UCS requirement is 700 kN/m^2 [76].

Lime for stabilization purpose should be either slaked at site or pre-slaked lime transported suitably can be used. Storage of lime should be done carefully to avoid prolonged exposure to the atmosphere. The soil to which lime is to be added should be free from organic or any other deleterious matter. The mixing should be thorough and uniform for effective soil-lime stabilization and could be done either by an in-place mixing method or in a stationary plant [76].

The stages involved in the in-place mixing method are [76]:

(i) Pulverization
(ii) Spreading of lime
(iii) Mixing
(iv) Addition of water
(v) Final grading
(vi) Compaction
(vii) Curing

Pulverization has two stages—scarifying soil layer up to the required depth and breaking it down into small sizes, suitable for lime addition. Rotary tillers, disc harrows, and rotavators, or any other agricultural equipment can also be used. After pulverization is complete, the loose soil is shaped with the help of grader and lime is added by mechanical means. Then, water is added with the help of a water spreader, mixed with soil, and the surface is shaped using a suitable grader. It is then compacted and left for curing.

In the stationary plant method, the stages involved are[76]:

(i) Material collection and pulverization
(ii) Mixing
(iii) Transportation and spreading
(iv) Compaction
(v) Curing

A specialized plant for mixing could be used, however, for a small-scale job, a concrete batch mixer can be used as well. Usually, the stationary plant method is costlier than the in-place mixing method, but better control and more uniformity in mix is achieved through this method compared to the in-place mixing method.

Soil-lime stabilization should not be done when the air temperature in shade is 10°C or less [215]. The compacted thickness of the lime stabilized layer should not be less than 100 mm or more than 200 mm [215]. The time gap between mixing and compaction should not exceed more than 3–4 hours, otherwise the soil-lime mix starts gaining strength and extra compacting effort often becomes necessary [76].

Soil–lime–fly-ash stabilization

Soil–lime–fly-ash mixture can be used as a base or a sub-base in bituminous and concrete pavement construction. The soil suitable for lime–fly-ash stabilization should have PI value between 5 and 20, liquid limit less than 25, and particles smaller than 425 micron should be between 15 and 25% by dry weight of the soil–lime–fly-ash mixture [192]. Lime must have a purity level of 50% or more. Fly-ash is pozzolana with some cementitious property; it reacts with lime in presence of moisture, and forms cementitious compounds [192]. The optimum quantity of lime should give due consideration to the fraction which would react with fly-ash and the rest which would react with clay minerals. The tentative proportions of lime, fly-ash, and soil are suggested as 3:12:85 by weight [192], however, laboratory experiments decide the actual proportions keeping in view the objective of strength maximization. The minimum thickness of soil–lime–fly-ash as base or sub-base is recommended as 150 mm in IRC:88–1984 [192].

Soil–cement stabilization

Lime stabilization does not give good results when soil is gravelly or sandy in nature. Cement stabilization can be tried in that case as the other available option. The strength achieved through cement-stablization is faster and higher than that achieved through lime-stabilization, and cement-stabilization is specially recommended for soil stabilization in waterlogged areas. The quantity of cement is obtained by mix design in the laboratory with the help of unconfined compressive testing and durability tests. The procedure of cement-stabilization is the same as that of the lime-stabilization and involves the following steps [191, 215]:

- *Pulverization of soil.* Soil is pulverized with the help of a scarifier attached to a grader, in case the existing soil is to be used. The soil transported from borrow pits, or from other places, should be spread evenly to achieve the required depth. Sieve analysis at site will confirm whether the required degree of pulverization has been achieved.
- *Mixing requisite quantity of cement.* The quantity of cement needed for stabilization is determined experimentally, from the moisture-density relationship of soil, strength and durability tests of soil samples mixed with varied quantities of cement. At the site, mixing of requisite quantity of cement is done either by in-place mixing or by stationary plant method, depending on the equipment availability, suitability, and the type of project. Depending on the field moisture content, the quantity of water to be added during construction is determined beforehand. Extra water is added to take care of the possible evaporation during construction.
- *Compaction.* Compaction is carried out by the sheepfoot roller, pneumatic-tyred roller, and smooth-wheeled roller at the final stage. It should not be delayed after mixing is completed.
- *Curing.* Minimum seven days of curing time is provided and then subsequent pavement layers are laid to prevent drying [215].

Stabilization with bituminous binder

Bituminous materials are used for soil stabilization to make it waterproof and to bind the soil particles firmly [65]. Stabilization of soil with bituminous materials can be done using bituminous emulsion, low viscosity bitumen, cut-back bitumen, foamed bitumen, and so on. Bitumen, at this consistency, shows a tendency of getting attracted towards the fine particles having higher specific surface. This non-uniform distribution of bitumen dispersion is beneficial in gaining strength through better inter-particle contact of larger particles [143].

When water or steam is added to hot bitumen, foamed bitumen is produced which occupies a volume at least ten times that of the normal bitumen. The surface area of foamed bitumen being large and viscosity being low, it can effectively coat fine particles with large specific surface [143]. Foamed bitumen too, can be very effectively used for stabilizing the desert sand.

The presence of a mineral which shows less affinity to water is desirable for stabilization by bitumen. For stabilization with emulsion bitumen, it is first necessary to mix the soil with water. Sometimes cement is also added, which serves three purposes, namely:

(i) Cement helps emulsion to break

(ii) Cement absorbs excess moisture

(iii) Cement gives added strength

Similar to other stabilization techniques, requisite quantity of bituminous binder is mixed either in a movable plant or at site. The mix is then evenly spread and compacted.

Chemical stabilization

In chemical stabilization various chemicals such as phosphoric acid, calcium chloride, and sodium chloride are mixed with soil in right quantities, which in turn improve the characteristics of the existing soil.

13.7.3 Thermal Stabilization

Thermal stabilization is performed by increasing (or decreasing) the temperature of the soil in such a way that physical, and/or chemical, changes in soil take place which enhance the engineering properties of the soil. Freezing provides only a temporary stabilization. On the other hand, heating the clayey soil above 400°C, its engineering property changes completely and permanently. However, stabilization by heating requires a lot of energy and it is therefore recommended for special situations such as one where soil already contains some fuel in itself [99].

13.7.4 Closing Remarks

Only a few stabilization techniques have been discussed in this section. There are numerous other methods that can be adopted for soil stabilization on a case-to-case basis. When additives are tried on soil, as a general rule, the strength at various proportions of the additives are determined, then keeping in mind the other engineering properties of additive mixed soil and the economy, the optimum proportions are finalized.

The geotextiles show a promising application in soil stabilization. They are also used (i) to prevent crack propagation, (ii) to provide reinforcement in surface dressing, (iii) as drainage layers, and (iv) in erosion control.

13.8 BITUMINOUS PAVEMENT CONSTRUCTION

As already mentioned in Chapter 11 (see Figure 11.1), the layers in a bituminous pavement are named as: subgrade, sub-base, base, binder course, and wearing course. The binder course and the wearing course are together called *bituminous surfacing*. Depending upon the design, sometimes only the base course is provided instead of the base and the sub-base course. Similarly, only the wearing course is sometimes provided instead of both the binder course and the wearing course. The characteristics, specifications, and the construction procedure of these courses are described, in brief, in this section. Other than the above mentioned courses, a brief discussion on various interlayers is also included at the end of the section.

13.8.1 Subgrade

After the vegetation and organic dirt are removed from the earth surface, it requires mild compaction (two passes of 80–100 kN smooth wheeled roller, or equivalent, as per the recommendations). There should not be any soft spot present in the subgrade. Figure 13.11 shows compaction of subgrade as being implemented in the field.

Figure 13.11 Compaction of subgrade on NH-25 (*Courtesy: NHAI and ICT Ltd., Kanpur*).

Soil is spread in uniformly in layers with thickness not exceeding 200 mm and is then compacted. Motor grader is used for maintaining a suitable grade during construction. If the moisture content is low, requisite amount of water is added uniformly with the help of a sprinkler system. If the roadbed material is too wet, it should be dried through aeration and exposure to sun.

13.8.2 Granular Base/Sub-base Course

Construction of granular base or sub-base consists of laying and compacting suitably selected aggregate material over the formed subgrade. It may have a number of layers with different materials, compaction levels, and thicknesses specified by the design. The materials of construction can be natural sand, moorum, crushed stone, crushed concrete, brick metal, crushed slag, and so on, subject to their conformance to specification requirements. As per IRC37: 2001, the sub-base material, for bituminous pavement construction, should have CBR value 20% for traffic up to 2 msa and 30% for traffic more than 2 msa [89]. The following discussion throws light on specification and construction of some types of base or sub-base.

Granular sub-base

The name of this course is Granular Sub-base and is more popularly known as GSB. The aggregates to be chosen should have 10% fines value as 50 kN in soaked condition. The materials as per the specified gradation with requisite water should be spread with the help of a motor grader. Rolling should be done by a 80–100 kN static roller with plain drum, or by a 200–300 kN pneumatic roller whose speed should not exceed 5 kmph. Figure 13.12 shows a photograph of dumping and spreading phase of GSB.

Figure 13.12 Dumping and spread of GSB (*Courtesy: NHAI and ICT Ltd., Kanpur*).

Water Bound Macadam (WBM)

Water Bound Macadam (WBM) is constituted with a compacted layer of clean, crushed aggregates and screening material laid on a properly prepared subgrade, base or sub-base course [215]. Binding material is used wherever necessary and water is added for proper compaction. Broken stones, crushed slag, overburnt bricks, and any other naturally occurring aggregates conforming to the physical requirement, can be used for WBM construction. The screening materials generally consist of the same material as the coarse aggregates. Binder materials may not be necessary in those cases, where crushable type of screening material, such as moorum or gravel, is used.

If WBM is placed over the subgrade, a suitably designed filter layer is laid first. Coarse aggregates are then spread evenly. Care is taken such that no segregation between the coarse and fine particles develops. Rolling can be done with a vibratory roller, or any other roller suitable for the job decided, based on the trial run. Sprinkling of water can be done whenever necessary. Rolling is discontinued for application of the screening materials, which are gradually applied over the surface and broomed. Subsequent dry rolling is preferred so that the screening materials can occupy the space between the aggregates. The job is assumed to be complete when no further screening can be forced into the void space of the coarse aggregates. Water and binder material

(if necessary) are applied and the surface is swept with brooms. After completion, the pavement is allowed to dry overnight, and hungry spots are repaired on the next day. Further construction over WBM can only take place after WBM is completely dry.

Wet Mix Macadam (WMM)

The construction of Wet Mix Macadam (WMM) consists of laying and compacting clean, crushed, and graded aggregates, premixed with water. WMM is prepared in a mixing plant, in which aggregates and water with suitable proportion are mixed together. The optimum moisture content of the mix is determined in the laboratory. The aggregates, immediately after mixing, are laid on the surface. For laying of WMM, lateral confinement is necessary. After the completion of the construction, setting time is given, during which it is desirable that not even construction equipment should pass over the surface. Figure 13.13 shows the WMM work in progress.

Figure 13.13 WMM work in progress at km 67 of NH-25 (*Courtesy: NHAI and ICT Ltd., Kanpur*).

Crushed cement concrete as base/sub-base

Crushed concrete, derived from crushing of rejected or damaged cement concrete can be used for base or sub-base course provided the pieces are strong, durable, clean, and conform to the gradation requirements [215]. The construction procedure is the same as WBM, except that no screening or binding materials are added. If needed, the top layer is treated with penetration grade bitumen.

13.8.3 Cemented Base/Sub-base Course

Various types of cemented materials can be economically used as base/sub-base. The reader may refer to Section 13.7.2 for a discussion on cemented materials, covered while discussing soil stabilization aspects.

13.8.4 Bituminous Sub-base

Bituminous Penetration Macadam

The bituminous penetration macadam consists of one or two layers of construction of crushed stones with alternate application of binder and key aggregates. First, dry, clean, coarse aggregates are laid on the prepared base, and compacted with 80–100 kN smooth steel wheel roller. Specified quantity of binder is then applied, and immediately after the application of binder, key aggregates (aggregates of specified gradation and of smaller size) are spread with a mechanical spreader. The surface is compacted by rolling to make a smooth finished surface. Construction of a surfacing course over the bituminous penetration macadam is completed within 48 hours. In case, there is a delay, the surface is covered by a seal coat.

Built-up Spray Grout

Built-up Spray Grout consists of two-layer composite construction of compacted crushed aggregates with application of bituminous binder after each layer, and the key aggregates placed on the top of the second layer [215]. The technique followed here is the same as other construction techniques described before, except that the sequence of spreading the coarse aggregates, binder, and key aggregates varies for different techniques.

13.8.5 Bituminous Binder Course

The commonly used binder courses include Bituminous Macadam (BM) and Dense graded Bituminous Macadam (DBM). BM has a high level of voids and therefore it is pervious to water. During the summer season, its stiffness becomes too low [60]. Thus, guidelines restrict that it should be used only for the roads whose design life is less than 5 msa [89]. DBM can also be used as a base course, or as a profile corrective course [215].

Aggregates are mixed according to the specifications and the optimum bitumen content (OBC) is found out from the Marshall requirement set in the specification. For some mixes, such as BM, OBC is not determined by Marshall test, rather density value is used.

Before laying, the surface should be thoroughly cleaned and any loose materials should be removed. This can be done using mechanical brooms or high pressure air jets.

The bituminous mixes for pavement construction are produced in the bituminous mixing plants. There are two kinds of plants, namely (i) bituminous batching plants and (ii) continuous plants. In a batching plant, ingredients are separately weighed in batches and then mixed. In a continuous plant, materials are continuously fed in through volume proportions and mixed. A stationary plant manufactures bituminous mixes at the place of its location (away from site), and the mixes are then transported through dump trucks. In a portable plant, bituminous mixes are manufactured near the pavement construction site itself, and after the completion of construction at that location, the plant is shifted to the next location.

The dump trucks used for transportation of mixes should be clean from inside, and may also have an insulation system to prevent heat loss. A thin coating of lubricating oil is sometimes applied to the inner surface to prevent sticking of bituminous mixes.

After laying of bituminous mixes, compaction is immediately done. Vibratory rollers of 80–100 kN dead weight and pneumatic tyred rollers of 120–150 kN with nine wheels are generally specified for this compaction. The finish rolling is done using 60–80 kN smooth wheeled tandem rollers [215]. The level of compaction should be a specified fraction of the Marshall density determined in the laboratory, for individual category of mixes. Various temperature ranges for different stages of construction are recommended by the specification; the corresponding table from the MORT&H specifications [215] is presented in Table 13.2.

Table 13.2 Manufacturing and rolling temperatures of bituminous mixes as per MORT&H specification [215] (4th revision)

Bitumen penetration grade	*Bitumen mixing* (°C)	*Aggregate mixing* (°C)	*Mixed material* (°C)	*Rolling* (°C)	*Laying* (°C)
30/40	160–170	160–175	170 maximum	100 minimum	130 minimum
60/70	150–165	150–170	165 maximum	90 minimum	125 minimum
80/100	140–160	140–165	155 maximum	80 minimum	115 minimum

13.8.6 Bituminous Wearing Course

Wearing courses are provided at the top of the pavement surface to

(a) have adequate skid resistance,

(b) have a waterproof surface,

(c) prevent (to some extent) entry of air to the pavement, which causes aging of the bitumen,

(d) arrest disintegration of particles from the existing road surface, and

(e) reduce hazards caused by limited visibility, poor night visibility, and dusty sorroundings, and so on.

Some of the commonly used wearing courses are Bituminous Concrete (BC), Semi Dense Bituminous Concrete (SDBC), Mix Seal Surfacing, Surface Dressing, Premix Surfacing, and Fog Spray, etc. Various wearing courses are discussed in the forthcoming paragraphs.

The construction procedure for BC and SDBC is the same as that of BM or DBM discussed in the previous section. The Marshall stability requirements for BC and SDBC have already been mentioned in Table 10.6. Figure 13.14 compares the upper and lower limits of BC and SDBC.

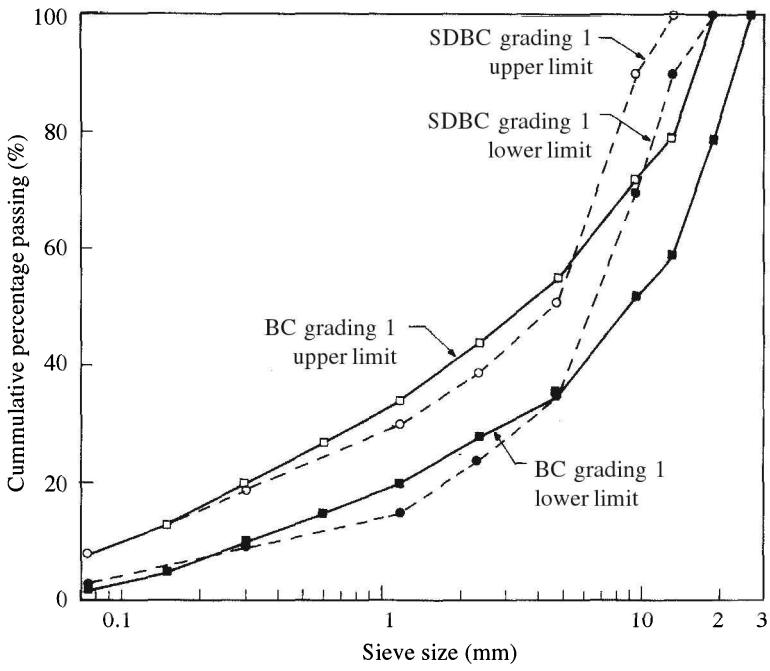


Figure 13.14 Aggregate gradation showing upper and lower limits of BC and SDBC.

Fog spray or Fog seal

In Fog spray, emulsified bitumen is sprayed over the finished surface. Fog spray can be used for sealing of cracks which are less than 3 mm in width. This reduces the tendency of small size aggregates to loosen from the top surface of a newly constructed pavement. Fog spray can be applied by means of a self-propelled or towed bitumen pressure sprayer.

Slurry seal

In slurry seal, bitumen emulsion, fine aggregate, Portland cement filler, and water are applied to the top surface which is then compacted. Slurry seal can be spread to a thickness of 1.5–5 mm.

Minor surface repair, arrest of fretting, and sealing of cracks are done by application of a thin-layer slurry seal. Slurry seal may enhance the riding quality slightly [187], however, for roads with high speed traffic, the chances of hydroplaning increase [235].

The pavement surface should be thoroughly cleaned and broomed before the slurry seal is applied. If emulsion is used as binder, pavement surface may be slightly pre-dampened. For a properly designed slurry seal, no tack coat may be necessary. Bitumen-

aggregate-slurry, ready for use, should be smooth, freeflowing, and a homogeneous mixture; emulsion bitumen should in no way breakdown before its application to the pavement surface.

Surface dressing

Surface dressing may be constituted with single or double layers. Each layer is constructed by spraying bitumen over the prepared base, then spreading aggregates, and finally rolling and compacting. Uncrushed rounded gravel should not be used in the construction of surface dressing. The aggregate to be used should preferably be hydrophobic in nature and of low porosity.

Figure 13.15 shows a cross-section of surface dressing at various stages of construction. After spraying a predetermined quantity of bitumen over the top surface, the single size aggregates, as per specification, are spread over it. It may be assumed that the aggregates which are just spread over the binder, show 50% void (see Figure 13.15(a)). When the aggregates are compacted by a roller, they start reorienting themselves, and the void drops down to 30% (see Figure 13.15(b)). Aggregates are further compacted and reoriented by the passage of traffic, and the void finally reaches 20% (refer Figure 13.15(c)). Aggregate particles, at this stage, lie evenly on their flattest side, which is called the *Average Least Dimension*, ALD. Thus, the height of the surface dressing is equal to the ALD. The quantity of bitumen is so adjusted that it occupies 3/4th of ALD [97, 232]. If the height of binder is more than this optimum height, it may cause bleeding and hydroplaning, and if it is less, it may cause fretting of aggregates. And thus, the optimum height of bitumen as 3/4th of ALD is arrived at. This constitutes the basis of design of surface dressing, and the quantity of the aggregate and binder is determined on this basis.

(a) 50% void immediately after spreading the aggregates

(b) 30% void immediately after rolling the aggregates

(c) 20% void after traffic operation

Figure 13.15 The concept of surface dressing.

For surface dressing with emulsion bitumen as the binder, however, the quantity of emulsion should be so adjusted that when laid it occupies the full height as that of ALD, such that, after evaporation the height of binder comes down to 3/4th of ALD. This has been shown in Figure 13.16.

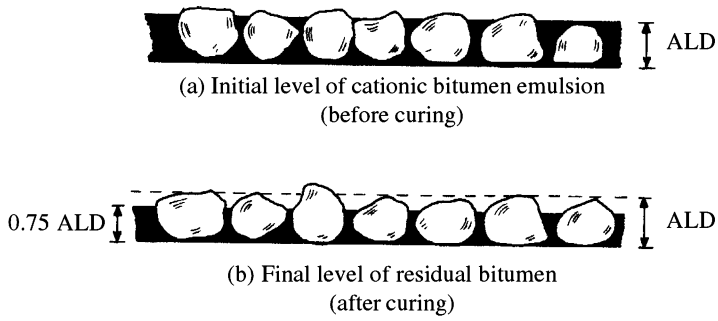


Figure 13.16 Surface dressing with emulsion bitumen as the binder.

Design considerations of surface dressing. The design of surface dressing determines

(i) the type of surface dressing,
(ii) the aggregate type and gradation and binder type, and
(iii) the quantity of binder and aggregates.

In as early as 1935, Hanson [97] proposed a scientific procedure for estimating the quantity of surface dressing. The following example illustrates the design of surface dressing which is based on Hanson's [97, 232] approach.

EXAMPLE 13.1

The aggregate gradation for a single layer surface dressing with nominal size of 13 mm is presented in the following table [215]. The flakiness index of the aggregates is obtained as 20%. Design the quantity of binder (penetration grade) and aggergates to be used for construction.

Sieve size (mm)	19.0	13.2	9.5	6.3	2.36	0.075
Cumulative % passing	100	92.5	17.5	5	1	0.75

Solution

Estimation of quantity of aggregates

The aggregate gradations are plotted in Figure 13.17. The median size of the aggregates corresponding to the 50% passing as found from the graph is 11.5 mm.

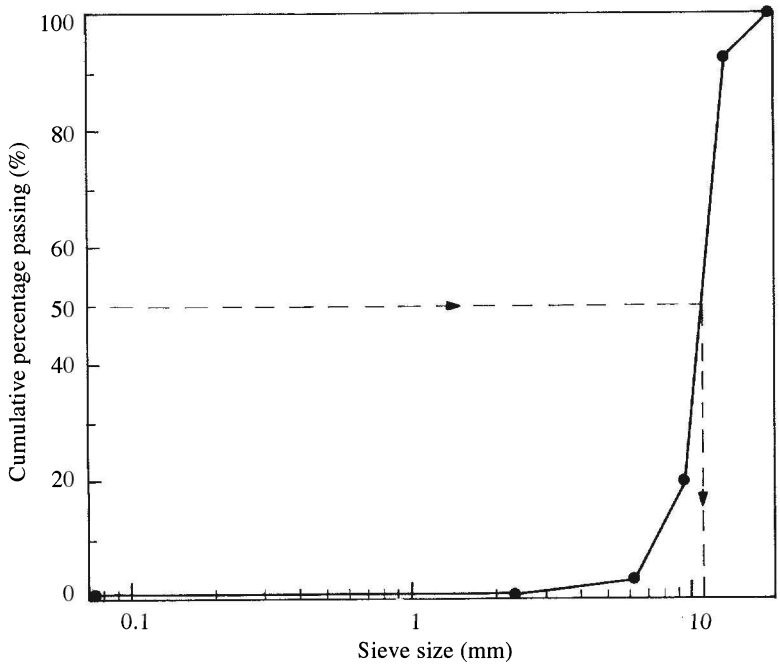


Figure 13.17 The gradation of the aggregates in the surface dressing design problem.

Now for obvious reasons, the median size of the aggregates is not the ALD. This would have been only possible when all the aggregates were of the same size and spherical in shape. Thus, for the present case, ALD is evaluated using the nomograph [97] shown in Figure 13.18. ALD of the aggregates (corresponding to median size 11.5 mm and flakiness index 20%) is found to be 8.6 mm.

As shown in Figure 13.15, 50% initial void after rolling and traffic compaction becomes 20%. Thus, the quantity of aggregates required for laying surface dressing in the 1 m × 1 m area is

$$\frac{80}{50}\text{ALD}$$

Assuming 10% loss due to whip off and 5% loss of aggregates due to handling (i.e. the total loss of 15%), the actual quantity of aggregates required is

$$\frac{80}{50}\times\text{ALD}\times\frac{115}{100} = 1.84\text{ALD} = 0.0158\ \text{m}^3$$

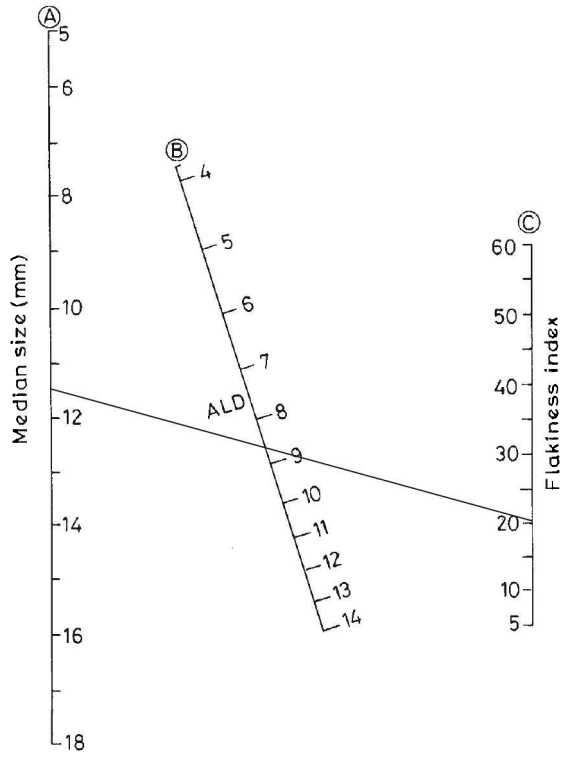


Figure 13.18 Relationship between mean aggregate size, flakiness index, and ALD [97].

This estimation of volume calculation remains unchanged when it is assumed that the void ratio of the loose aggregates carried in trucks is also 50%.

Estimation of quantity of binder

After the final compaction of aggregates due to traffic rolling, the surface dressing layer has 20% voids and binder is required to fill 3/4th (i.e. 75%) of the voids. Therefore, the quantity of binder required per 1 m × 1 m is

$$\frac{75}{100} \times \frac{20}{100} \times \text{ALD} = 0.15\text{ALD} = 0.00129 \text{ m}^3$$

Construction of surface dressing

It is important that the surface be dry and thoroughly cleaned just before applying the binder. If the base to be treated consists of stabilized soil or of porous aggregates, a suitable bituminous primer should be applied uniformly by a mechanical sprayer [232].

Bituminous material is spread uniformly in specified quantities. Immediately after the application of bituminous material, the cover material is spread uniformly by means of a mechanical gritter so as to cover the surface completely in the first coat. While rolling, the

surface is broomed to ensure uniform spreading. Immediately after the application of the cover materials, the entire surface is rolled with a 60–80 kN roadroller. However, traditional steel-wheeled rollers tend to crush the aggregates and if their use cannot be avoided, their weights should be limited to 80 kN [187]. Rolling is continued until the particles are firmly embedded in the bituminous material and form a uniform closed surface. Excessive rolling which results in crushing of the aggregate particles should be avoided [232].

Where straight run bitumen is used for the construction of surface dressing, it can be opened to the traffic on the next day. But, for cutback or emulsion bitumen, sufficient time should be provided till it is completely cured [232].

Premix surfacing

Construction of premix surfacing involves laying and compacting wearing course consisting of small-sized aggregates and binder (penetration grade or emulsion or cutback form) mix to form a layer of thickness 20 mm. Depending on the size specification of aggregates, the premix surfacing can either be open graded or be close graded. The *close graded* premix surfacing is also known as *Mixed Seal Surfacing*. Mixing is done separately, transported to the site, spread over the prepared surface, and compacted by a 80–100 kN roller (or any other equivalent equipment). The rolling operation needs to be completed before the temperature of the mix falls below 100°C [215]. Generally, a seal coat is applied over the premix surfacing and it is specially recommended when the open graded premix surfacing is chosen.

Seal coat

Seal coat is used to seal the voids of the top bituminous construction. Seal coat is of two types [215], namely:

(i) Single application of bituminous binder followed by stone chips cover, known as Type A seal coat

(ii) Application of fine aggregates premixed with the binder, known as Type B seal coat.

Mastic asphalt

In mastic asphalt, well-graded aggregates, filler, and bitumen are mixed in such a proportion that a dense, almost voidless mixture is obtained. The acceptability of mastic asphalt is judged by the hardness number which should be 10–20 at the time of laying [215]. Mastic asphalt is prepared in a mechanically-agitated mastic asphalt cooker. First, the full quantity of preheated filler and bitumen are added at a definite temperature. Then, the fine aggregates and balance half of the binder are added, and the mixture is cooked for one hour. Finally, the coarse aggregates are added and the mixture is cooked for a total period of three hours. Care should be taken to ensure that the temperature

never goes beyond 210°C. Mastic asphalt is laid in hot stage and a uniform thickness is achieved by using wooden floats or machines. The skid resistance of mastic asphalt can be low, therefore, as a preventive measure immediately after the construction when the surface is still hot, hard stone chips of specific dimensions are spread over the surface, and rolled.

13.8.7 Interlayer Coats

This subsection discusses the various interlayer coats of the pavement. The purpose of the interlayer coats is:

(a) To provide adequate bonding between the layers

(b) To achieve a waterproof surface

(c) To hold the loose materials together

(d) To fill up the voids

Figure 13.19 gives an idea regarding the relative position of various coats.

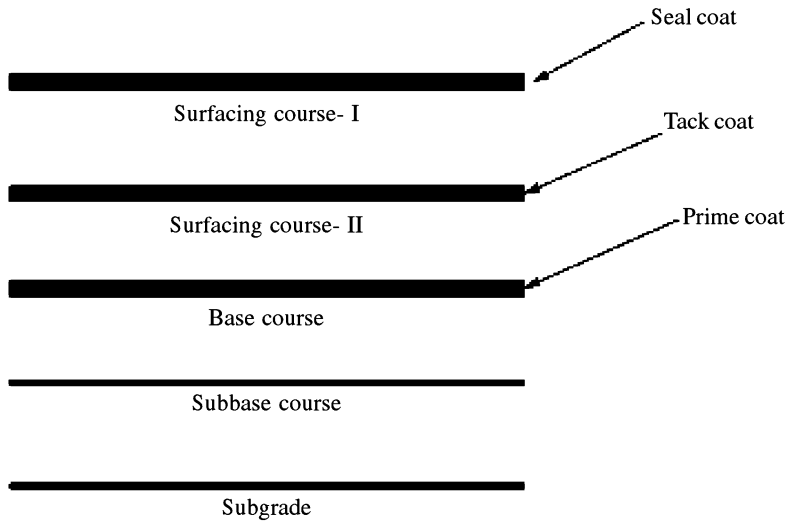


Figure 13.19 Interlayer coats in a bituminous pavement.

Prime coat

Single coat low viscosity bitumen is used as the prime coat where the granular base acts as the absorbent. The type of bituminous primer is chosen, based on the porosity of the granular course. Cutback or emulsion bitumen can be used as the prime coat. Curing time must be provided for

complete absorption of the binder by the base course. Self-propelled or towed bitumen pressure sprayers are best suited for application of the prime coat [215].

Tack coat

Tack coat is a single coat, low viscosity liquid bituminous binder applied to existing road surfaces to ensure good bonding between the surface and the bituminous overlay. The rate of application of tack coat should be controlled so that the film thickness does not become more than that required. Self-propelled or towed bitumen sprayers can be used for applying tack coat. The surface on which the tack coat is to be applied should be free from dust or dirt. For emulsion bitumen, the tack coat should be allowed to cure by evaporation until the time a new layer is laid over it [215].

Stress relief interlayers

The stress relief courses are applied to prevent further propagation of cracks, and are known as Stress Absorbing Membranes, SAM, and Stress Absorbing Membrane Interlayers, SAMI. These layers are composed of elastomeric bitumen-rubber membrane and open-graded aggregates. Alternatively, they may consist of bitumen impregnated with geotextiles. SAMI can be applied to the cracked pavement surface but overlay should be laid within 12 months.

13.8.8 Closing Remarks

Various layers of bituminous pavement construction have been discussed. As mentioned earlier, the specifications may vary in different guidelines. Sometimes due to fund constraints or any other reason, the construction to the full form may not get completed. Thus, incomplete construction in the form of earth road, gravel road, soil stabilized road, or WBM road (their meanings are self-explanatory) may also exist.

13.9 CEMENT CONCRETE PAVEMENT CONSTRUCTION

This section discusses various layers of concrete pavement, namely the construction of subgrade, base/sub-base, and concrete surfacing. The provisions of concrete pavement joints are described at the end.

13.9.1 Subgrade

Where the concrete layer is laid directly over the subgrade, it should be ensured that the subgrade is moist at the time of laying. If the subgrade is dry, water can be sprinkled over the surface before laying the concrete course, however, care should be taken to ensure that no soft

patches or water pools are formed on the surface. Water should be applied not less than 6 hours or more than 20 hours in advance of laying concrete [220]. As an alternative arrangement, concrete can be laid over a waterproof polyethylene sheet, and in that case moistening the subgrade surface becomes redundant. This polyethylene sheet acts as a capillary cut-off layer [220]. The minimum modulus of subgrade reaction, k, obtained from the plate load test should be 5.5 kg/cm^3 [220, 91].

13.9.2 Base/Sub-base

Sub-base for concrete pavement can be chosen from a number of alternatives depending upon the convenience, material availability, and many other factors. Sub-base can be constituted with brick flat soling, WBM, granular aggregates, crushed concrete, slag, stabilized soil, and so on. According to IRC:15 [220], the sub-base can be of three types:

(a) With granular material, composed of brick soling with one layer of sand below it (WBM, well graded granular materials, etc.)

(b) With stabilized soil

(c) With semi-rigid material, such as, lime burnt clay pozzolana concrete, lime fly-ash concrete, lean cement concrete, and so on.

The following is a brief discussion on dry lean cement concrete as sub-base and rolled cement concrete as base.

Dry lean cement concrete sub-base

Dry lean cement concrete (DLC) is sometimes used as a sub-base for concrete construction with a recommended thickness of 100 mm or 150 mm [77]. The maximum aggregate to cement ratio is 15:1. The water content required should be calculated from compaction point of view in the trail stretches [215]. The average compressive strength of DLC cubes after 7 days should not be less than 10 MPa (tested on five samples) and the individual compressive strengths should not be less than 7.5 MPa, after 7 days [215, 77].

Before construction of the DLC sub-base, the prepared subgrade is sprinkled with water to moisten the surface. It may also be rolled with one or two passes of a smooth-wheeled roller, if necessary [215]. The DLC mix prepared in a batch mixer should be transported to the site immediately. The material is laid by a paver, without any segregation. The paving machine should have high amplitude paving bars, such that the sub-base can be subjected to initial compaction. The work should not proceed if the temperature rises above 30°C and in that case ice water may be sprinkled.

The curing of DLC can be done by spraying a liquid curing compound, or by covering the surface with gunny bags. The construction of the cement concrete pavement can only start after 7 days of sub-base construction [215, 77].

Rolled cement concrete base

The maximum aggregate-cement ratio in rolled cement concrete is 15:1. The optimum moisture content is determined through trial mixes by varying the water content 5–7% of the dry weight of the mix. The mix is then laid on the prepared base which is moistened beforehand, if necessary. Full compaction is achieved by proper rolling and using a suitable water content [215].

13.9.3 Concrete Surfacing

The proportion between the cement, aggregate, and water is determined by the standard concrete mix design technique [227]. Mixing is done in a batch mixer for uniform distribution of materials. The spreading of concrete should be done uniformly and with care such that no segregation of materials takes place. A separation membrane, made up of an impermeable plastic/polyethylene sheet (150 microns thick) is preferably laid over the sub-base without any creases in it, on to which concrete is laid. Concrete should be rolled with appropriate equipment such that the formation of honeycombs or voids is avoided. The finishing of the surface is done by a power-driven finishing machine or a vibrating screed. A slip form paving machine can automatically spread, compact, and finish the surface through its feedback sensors. In fixed form pavers, separately powered machines for spreading, compacting and finishing are used. A slip form paver requires guidewires, parallel to the edge of construction and maintained at a fixed height, and is installed on both sides. The alignment of the slip form paver is controlled automatically with respect to the guidewires. Figure 13.20 shows a concrete pavement under construction using a slip form paver.

Texturing of the surface is done by brooming over the laid surface of the pavement along

Figure 13.20 Concrete pavement construction by slip form paver (*Courtesy: Mr. Pinaki Roychowdhury*).

the transverse direction. The broomed surface should be free from any form of irregularities—the irregularities which may be caused due to loosening of larger aggregates nearer to the top surface. Care should be taken so that rapid drying of water from the concrete surface is prevented. A curing compound with high water retentivity should be spread over the finished surface and it should not react chemically with concrete. For final curing, continuous ponding should be done or moistened hessain should be kept over the surface for about a fortnight. As an alternative to ponding, an impervious liquid can be spread over the surface to restrict evaporation of water from the laid concrete, and curing then takes place due to water already present in concrete, which otherwise would have evaporated. Forms are removed from the freshly prepared concrete layer after curing for about fourteen hours. Proper measures need to be taken when concreting is done during monsoon, hot, or cold weather.

After curing is over and before opening the road to traffic, the temporary seal material is removed, and the joints are filled with the recommended joint sealing compound. The pouring of sealing material is carefully monitored so that it does not spill over the pavement surface.

13.9.4 Joints in Cement Concrete Pavement

The principles behind various types of joints in concrete pavement have already been described in Section 11.2.2. The dowel bars and the tie bars, as required in the joints, are provided as per the design recommendations. The transverse and longitudinal joints in sub-base and those in the concrete pavement are placed in a staggered manner.

Joint fillers

Joint fillers are the materials used to fill the joints. The three basic categories of joint materials [2] are:

(i) Liquid sealants
(ii) Performed elastomeric seals
(iii) Cork expansion joint fillers

In the case of liquid fillers, materials such as bitumen, rubber, silicon, and polymers are put inside the joints and allowed to set. Elastomeric seals exert resistive force towards the movement of the joints. Joint fillers should be poured into the groove when it is in the maximum expanded form. Premoulded joint filler is another type of joint filler where a joint filler of specified size is manufactured separately. The height of the joint filler should be less than the thickness of the concrete slab.

Joint sealers

Joint sealers are the compounds which are poured above the joint fillers and prevent water percolation into the joints. Joint sealers can be of hot elastomeric type or cold polysulphide type. Before putting the joint sealer, the joint groove must be cleaned by blasting filtered, oil-free air through it [215]. A suitable bituminous primer is applied and dried before the application of joint sealers [220]. For hot application of joint sealants, they are heated by

a thermostatically controlled heater and poured into the groove. Heating is not necessary for cold joint sealers.

13.10 RELATED TOPICS

Emulsified bituminous mix, precoated aggregates, recycling of bituminous pavement, construction of shoulder, are the four different topics briefly discussed in this section.

13.10.1 Emulsified Bituminous Mix

Emulsified Bituminous Mixes (EBMs) are gaining wide acceptance among road engineers because of their ecological performance and economic advantages. In EBMs, bituminous emulsions and aggregates are mixed cold using in-place, or portable plant method.

Shortly after placing and compaction of EBMs, water starts evaporating, causing breaking of emulsions that brings about damage by weather and development of resistance to traffic. On curing, EBMs develop a strength comparable to a bituminous concrete (BC) as shown in Figure 13.21 [257].

Advantages and disadvantages of EBMs

EBMs have certain distinct advantages over the hot mix laying technology for bituminous road construction. These can be enumerated as follows [129]:

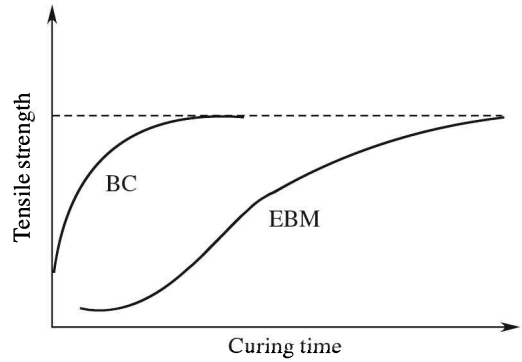


Figure 13.21 Strength of BC versus EBMs.

(i) Heating of the binder and the aggregates is avoided which minimizes hazards from fuming, fire, and environmental pollution in the nearby area.

(ii) Ease in construction as compared to hot mix application.

(iii) EBMs can be applied on wet surfaces and construction can be taken up even in rainy seasons.

(iv) Marginal saving in bitumen consumption and significant conservation of energy is observed in EBMs.

(v) Potholes can be repaired during rainy season. Inherent anti-stripping agents are present in EBMs.

(vi) Oxidative hardening of bitumen due to heating in hot mix paving can be avoided.

The disadvantages of EBMs [129] can be listed as follows:

(i) Slightly higher cost of construction compared to hot mix bituminous construction.

(ii) Setting of the mix takes a longer time. Traffic cannot be opened on a newly constructed surfacing soon after its construction.

(iii) Bitumen remains softer for long time in hot climate.

13.10.2 Precoating of Aggregates

Precoated aggregates are the aggregates which are already coated with a thin film of bitumen before application to actual construction. For effective precoating, the aggregates must be free from dust and moisture, and the precoating thickness should be small. Generally, precoating is done by applying a fine spray of bituminous binder over the moving stream of aggregates [187]. After precoating is completed, sufficient time should be given for the coat to dry up.

13.10.3 Recycling of Bituminous Pavement

In the recycling process, existing bitumen and aggregates are used in the new construction, thus conserving some materials.

First, tests are conducted on the samples taken from the existing pavement to evaluate the reusability of bitumen and aggregates. The aggregates and the bitumen are separated out. Aggregates are tested for their gradation and the recovered bitumen is tested for its properties, and based on the results, their suitability for use as recycled material is determined. The optimum quantity of existing aggregates and bitumen, to be mixed with fresh aggregates and bitumen, is determined from the mix design through trial mixes. Sometimes, rejuvenators are suggested to be used for restoration of properties of the existing binder. The rejuvenators are made up of a soft binder of appropriate penetration. The mix design procedure for the recycled bituminous mixture,

suggested in the Asphalt Institute manual [163] is recommended by the Indian specifications [215].

The recycling process can be classified as following:

(a) Recycled in-situ (hot or cold mix process)

(b) Recycled in plant (hot or cold mix process)

In the first method, the existing surface is heated and scarified, but not removed [215]. A layer of freshly prepared bituminous mix is then spread over the surface and compacted immediately. In the second method, the bituminous layer is removed partly or wholly and remixed in the plant with fresh aggregates, bitumen, and rejuvenator. Removal of the existing bituminous surface can be done by two processes, that is, either by removal at ambient temperature or by the hot removal process [215]. In the removal process at ambient temperature, scarifiers, grid rollers or rippers are used to remove the bituminous material. In the hot removal process, as the name suggests, the pavement surface is heated by suitable means and a milling drum is used to collect the soft bituminous material. During this heating process, the temperature of the surface of the road should not exceed 200°C for more than 5 minutes. The depth of scarification should be such that the bottom level is parallel to the existing finished road surface [215].

13.10.4 Shoulder Construction

Shoulders on either side of the pavement may be constructed with selected earth or a granular material. In treated shoulders, some stabilizing material is added to prevent erosion loss and ingress of water. Paved shoulders consist of sub-base, base, and surfacing. In a paved shoulder, the construction of a layer of the shoulder and that of the corresponding pavement layer should proceed simultaneously [215]. Then only, the construction of the next layer is taken up. If the layers are different in composition (in pavement and shoulder), first the pavement is compacted and then the shoulder.

EXERCISES

1. What technological lessons do you derive from McAdam pavement?
2. Discuss the various types of earth moving equipment.
3. What are the various types of rollers used for highway construction? Mention their specific uses.
4. What are the conditions for curing?
5. How can the field density after compaction be checked?

Highway Maintenance

14.1 INTRODUCTION

An in-service pavement requires maintenance. The maintenance may be routine in nature, or may be a major reconstruction.

A pavement is designed against an assumed design period. After the expiry of the design period, the pavement is likely to fail structurally and, therefore, it would require a major renewal to extend its life further. Even within the service life of a pavement, the top wearing course is likely to be subjected to considerable distress due to the movement of vehicles on it. Thus, the wearing course needs some routine maintenance for smooth movement of vehicles over the pavement.

The present chapter introduces the reader to the maintenance issues of a pavement. The chapter is divided into five sections, apart from the introduction, of which the first section discusses the various forms of distresses of pavement, their possible origin, and their quantification in terms of distress indices. The next two consecutive sections deal with the functional and structural evaluations of pavement—their techniques, related equipment, and analysis procedures. Remedial measures to extend the longevity of pavement are discussed in the fourth section while the last section briefly mentions the need for evolving maintenance strategies subjected to possible fund constraints.

14.2 DISTRESSES IN PAVEMENTS

In most of the cases, the distresses in pavement are measured as the distress per unit area of the pavement. For example, one may specify a pavement as 25% cracked and 30% corrugated. In some cases, the distresses are subjective, such as corner cracks, or are expressed in some other units, such as rutting which is expressed as depression measured by a 3 m straight edge, or aggregate polish which is expressed in terms of skid resistance, and so on. According to the "code of practice for maintenance of bituminous surfaces of highways", IRC:82 [39], the defects of bituminous surfacing can be grouped

under four categories, as follows:

1. *Surface defects.* For example, fatty surfaces, hungry surfaces, smooth surfaces, streaking, and so on.
2. *Cracks.* For example, alligator cracks, longitudinal cracks, hairline cracks, shrinkage cracks, edge cracks, reflection cracks, and so on.
3. *Deformations.* For example, rutting, corrugation, shoving, shallow depressions, settlement, heaving, and so on.
4. *Disintegration.* For example, stripping, loss of aggregates, ravelling, potholes, and so on.

Subsequent subsections present a brief discussion on some of the major forms of distresses of both bituminous and concrete pavements.

14.2.1 Alligator Cracking or Fatigue Cracking

Bituminous pavement surfaces can exhibit distress due to flexural fatigue as a result of repetitive applications of vehicular loads. The cracks on the bituminous surface allow the surface water to percolate into the base and subgrade of the pavements which further accelerates the deterioration process. The fatigue cracks are like hexagons, joined together one after another, and hence, also known as *alligator cracking*. The following photograph (Figure 14.1) shows a typical fatigue cracking.

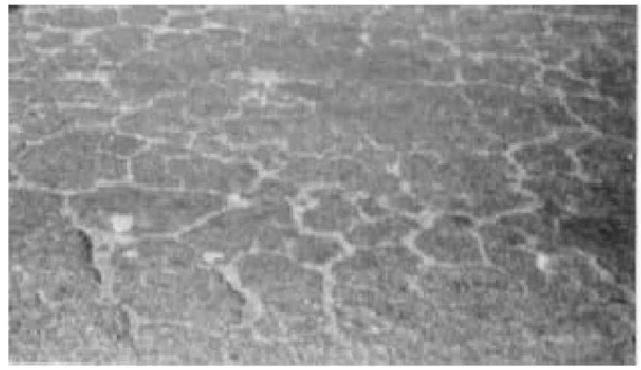


Figure 14.1 A typical fatigue cracking.

14.2.2 Block Cracking

Block cracks are the approximate rectangular cracks formed on the surface of the bituminous pavement. These characterize temperature shrinkage cracks, which originate due to daily variations of temperature, and essentially are of non-traffic origin. The traffic loading increases the severity of block cracks. As the cracks start appearing on the pavement surface due to hardening of the bitumen used, application of a new coat

of suitable bituminous layer, or recycling bitumen, can serve as a remedy to prevent block cracking [2].

14.2.3 Corner Break and Spall

Corner break and spall are the cracks developed in the concrete pavement at the corners of the concrete slabs. Spalling is not generally extended through the whole slab thickness [2]. These failures are due to the combined effect of mud pumping, heavy repetitive loading, poor load transfer across the joints, and thermal curling.

14.2.4 Corrugation

Corrugation is the plastic deformation of the top bituminous surface of the pavement along the horizontal direction. Its manifestation is in the form of undulations or ripple formations on the top surface of the pavement. Corrugation occurs due to lack of stability of asphalt mixes in warm weather. It is mostly observed where vehicles exert a greater horizontal force to start or stop, such as in the intersection legs where brakes are applied. If the corrugated surface is thin, it can be scarified and material can be relaid. The elevated spots are cut with a mechanical blade, with or without heating. The surface can be rolled afterwards.

14.2.5 Depression

Depression, as the name suggests, is the localized area where the pavement surface sinks a little with reference to the finished surface. Depression in a pavement occurs due to differential settlement of inadequately compacted subgrade (or other layers) due to traffic loading. Water accumulates on the depressed zone after rainfall, which percolates and causes further damage to the pavement in that area. Depression can also be due to inappropriate mix design or settlement of the lower pavement layers. Depressions can be removed by filling the depressed part with premix aggregates, followed by adequate compacting.

14.2.6 Fatty Surface or Bleeding

This is a surface defect associated with bituminous pavements only. It is the accumulation of bitumen on the surface of the pavement which occurs at high temperatures during the daytime. Bitumen at a high temperature, softens and occupies the available void space in the aggregates. If the space is inadequate, bitumen expands out onto the surface and forms a sticky, shiny surface over the pavement, called *bleeding* or *fatty surface*. It is an irreversible process, that is, the bitumen does not go back to the void space during the winter season. Proper mix design, which includes the selection of

appropriate grade of bitumen, and provision of requisite void space can control this bleeding phenomenon. Loss of cover aggregates, heavy prime or tack coat, non-uniform application of binder can be the other possible reasons of bleeding, therefore, these should be properly designed and controlled during construction. If the bleeding is uniform and without surface irregularities, small size, clean, angular sand, or small aggregates can be used over the surface. This is called *sand blotting* or *sand blinding* [39]. If the bled surface has irregularities, it is advisable to remove the affected portion, and relay it with a properly designed mix.

14.2.7 Hairline Crack on Bituminous Pavement Surface

Hairline cracks are small and fine cracks over the surface of the bituminous pavement. These cracks develop due to insufficient bitumen content, excessive filler at the surface, or improper compaction (i.e. over-compaction, compaction when the base is unstable, or compaction at a high temperature).

14.2.8 Hungry Surface

Hungry surface is a situation just reverse of the fatty surface. If the bitumen distribution rate is lower than the designed value, small cracks develop on the surface, and loss of aggregates may start taking place from the surface due to traffic. On the other hand, a hungry surface may also develop if the aggregates have a strong absorption affinity towards bitumen. Fog seal or slurry seal can be used as requisite measures to take care of the hungry surface situation.

14.2.9 Lane/Shoulder Drop-off or Heave

Shoulder drop-off is a situation when the shoulder elevation becomes lower than the level of the pavement lane. This occurs due to the following reasons:

(a) Gradual consolidation of the shoulder
(b) Erosion of shoulder materials due to rain or weather

Heaving of shoulder may occur due to frost heaving of the shoulder soil.

14.2.10 Loss of Aggregates

Loss of aggregates occurs subsequent to stripping or ravelling. The possible reasons for the loss of aggregates are:

(a) Improper mix design
(b) Improper design of surface dressing
(c) Inadequate rolling

(d) Traffic allowed to flow before proper setting of the binder
(e) Surface has become hungry due to absorption of bitumen unaccounted in mix design.

The treatment needed to prevent loss of aggregates depends on specific reasons. A layer of slurry seal, or fog seal, or relaying of surface dressing, or complete replacement of the disintegrated layer can be used as some of the possible solutions.

14.2.11 Map Cracking in Concrete Pavements

Map cracking refers to the small map-like cracks which are superficially located over the top surface of the concrete pavement. Such cracks are caused due to improper finish of the top surface, or due to the reinforcement bars being too close to the surface [2].

14.2.12 Patch

It is the repair work done on the existing potholes, depressions, or the corrugated pavement surface. Generally, patch work is a visually distinguishable feature of the pavement surface.

14.2.13 Polished Aggregate or Smooth Surface

Smooth surface or polished aggregate, as the name suggests, is a situation which arises due to repetitive passage of traffic on the aggregates of road, whose polished stone value (or the abrasive strength) is less. The skid resistance of the pavement therefore decreases, and this requires replacement of the top course with fresh angular aggregates, having a higher abrasive resistance.

14.2.14 Potholes

Potholes are bowl-shaped holes, caused by localized disintegration of materials, of varying sizes on the surface of the bituminous pavement, sometimes extending to the base course [39]. Due to variation in a large number of parameters involved, during highway construction, it may not be possible to maintain the same level of homogeneity althrough. The localized disintegration starts occurring from those places, which are the weakest spots on the pavement stretch. Potholes may occur due to a number of causes, such as:

(a) Inadequate construction quality control
(b) Ingress of water and subsequent damage
(c) Ravelling

Potholes are repaired by patchwork; a good bond is necessary between the existing pavement and the patchwork.

14.2.15 Pumping or Mud Pumping

Pumping is a failure generally observed in concrete pavements. When traffic moves on the cracked surface, or over the concrete joints, accumulated water along with subgrade soil (or sub-base particles) ejects out. This phenomenon is called *pumping* or *mud pumping* and it is more prominently observed in respect of the cases where the concrete pavement is put directly on the subgrade layer. The following situations lead to the occurrence of the pumping phenomenon [266].

(a) Material under the concrete slab saturated with water

(b) Frequent passes of heavy wheel loads

(c) Material under the concrete pavement is erodable in nature with low permeability.

The mechanism of mud pumping can be explained as follows:

(a) A void is first formed below the concrete slab. This can happen either due to post-construction plastic deformation of soil or due to warping of the concrete slab.

(b) Water accumulates in the void. Water may come from the surface infiltration or from other groundwater sources.

(c) Due to repetitive application of heavy vehicles, soil suspension in water is formed which is ejected out through cracks/joints at each pass of the heavy vehicle. If the material below the concrete slab is granular, the pumping may not occur because quick drainage takes place through this material.

(d) Gradual removal of the particulate material from below the pavement makes the voids bigger, and the pavement may fail due to corner break, and faulting at joints.

Figure 14.2 explains the mechanism of mud pumping.

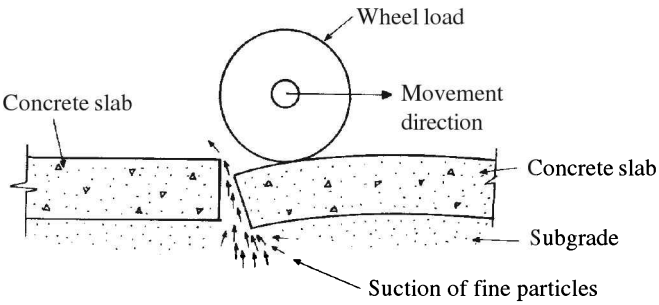


Figure 14.2 The mechanism of mud pumping.

14.2.16 Reflection Cracking

When a pavement (concrete or bituminous) is overlaid with a bituminous layer, sometimes, the same pattern of cracks as was in the existing pavement surface propagates upwards and comes up to the top surface of the new overlay. This is called *reflection cracking* because it appears, as if the cracks on the existing surface have been reflected onto the top overlaid surface. Reflection cracks occur due to the relative movement of the existing cracks of the original pavement. If the original cracks in the pavement are controlled, reflection cracks are automatically checked. To prevent reflection cracks, stress relief layers, geotextiles, or overlay reinforcement are provided as interlayers between the existing pavement and the overlay. Stress relief layers are the open graded aggregate specification which do not allow the cracks to propagate upwards. Geotextiles or overlay reinforcement, on the other hand, bear the tension themselves, and do not allow the cracks to propagate further.

14.2.17 Ravelling

Ravelling is the gradual wearing of the top surface, mainly due to weathering of bitumen. The binder becomes hard due to weather action, looses its binding property, and the aggregate particles are dislodged from the pavement surface, as the traffic moves over it. This form of pavement distress is termed ravelling. If the extent of ravelling is not severe, it can be rectified with one coat of slurry seal, or fog seal. Otherwise, a renewal coat may be necessary.

14.2.18 Rutting

As already discussed in Section 12.3.6, accumulation of permanent deformation along the maximum travelled wheel path is called *rutting*. The extent of rutting depends on the traffic repetitions the pavement has undergone, properties of the materials used in construction of the pavement, densification achieved during construction, average temperature of the pavement surface, and so on. If rutting is due to compaction of the layers, it can be rectified by applying a profile corrective course. A cross-section of the profile correcting course is shown in Figure 14.3.

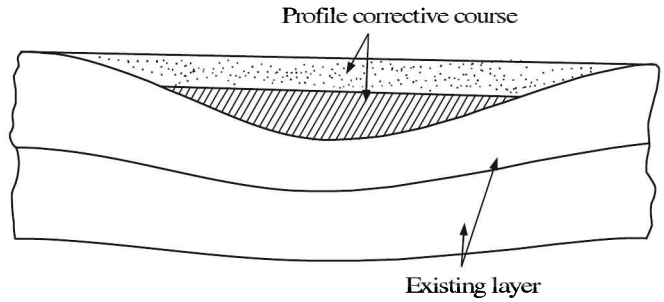


Figure 14.3 A cross-section of the profile corrective course [215].

14.2.19 Slippage

Slippage is a relative movement between the top wearing course and the layer below it along the horizontal direction. Slippage occurs when a horizontal thrust is applied by the vehicles, specially by the braking vehicles. It occurs in the absence of adequate bonding between the layers, that is, when the tack coat or the prime coat is either missing or inadequate. Slippage is associated with the crescent-shaped cracks on the pavement surface. For repair work, the wearing course of the affected area may be removed, and relaid ensuring proper interlayer bonding [39].

14.2.20 Streaking

Streaking is the appearance of alternate lean and heavy lines of bitumen, along the longitudinal or transverse directions. This is the result of non-uniform application of binder during construction [39].

14.2.21 Stripping

Stripping is a phenomenon in which the segregation of bitumen and aggregates takes place in the presence of moisture. The loss of bonding between the aggregates and bitumen causes loss of aggregates, further infiltration of water, loss of strength, and subsequent failure of the pavement. The stripping phenomenon is caused due to the following [39]:

(a) Use of hydrophilic aggregates
(b) Improper mix design with excess of fines
(c) Continuous exposure to moisture or accumulated water
(d) Opening the road to traffic before proper setting of the binder
(e) Aging of the binder

As a preventive measure, the stripping potential of the aggregate–binder should be checked before laying the pavement. If necessary, an anti-stripping agent can be used during the mixing process. The areas affected by stripping need re-laying with the fresh mix.

14.2.22 Swell and Blow Up

Swelling and blow up of the pavement occur due to expansion of the subgrade soil. The expansion could be either due to expansive nature of soil used in subgrade, or due to frost action. Such a failure is called swell and blow up in bituminous and concrete pavements respectively. Blow up also occurs due to infiltration of certain materials into the joints of the concrete pavement, which expand during the summer season, causing enough pressure to cause blow up of the concrete pavements [2].

14.3 FUNCTIONAL EVALUATION OF PAVEMENT

Various types and forms of pavement distresses have been discussed in the previous section. In a pavement, in fact, a number of distresses may occur simultaneously, because many of the distresses are interrelated, and the occurrence of one may as well initiate the other. Individual assessment and quantification of the distresses may not therefore be very useful. Rather, there is a need to assess the functional condition of the pavement as a whole. Table 14.1 illustrates the recommendations as per the Indian specification for classifying pavement condition based on visual assessment [74]. Two terms for the functional assessment of pavement, were developed from the AASHO [1] test—Present Serviceability Rating (PSR) and the Present Serviceability Index (PSI).

Table 14.1 Criteria for classification of pavement sections [74]

Classification	*Pavement condition*
Good	No cracking, rutting less than 10 mm
Fair	No cracking, or cracking confined to a single crack in the wheel track with rutting between 10 mm and 20 mm
Poor	Extensive cracking and/or rutting greater than 20 mm and cracking exceeding 20%

As a part of the functional pavement evaluation in AASHO [1] road test, people were asked to drive on the pavement stretch with a vehicle of their choice, and they were asked to rate the pavement surface in a scale ranging from 0 to 5. Later, PSI was developed which statistically correlated the physical measurements on pavement conditions to the subjective judgement of human rating (i.e. PSR). Thus, PSI is an empirical equation containing terms such as, cracked area, patched area, rut depth, and slope variance. PSI, as the functional index of pavement condition, has some deficiencies. For example, it was developed from the evaluation of a panel of experts in the AASHO test, and therefore, may not hold good in the present context. Also, the kind of profilometers that were used in the test are not in vogue [266] today.

Likewise, various other indices have been evolved to quantify pavement distress as a whole. Automatic equipment has been developed which can be driven over the road to acquire continuous data of functional condition of the pavement. The acquired data is analyzed in the laboratory to extract the desired information. In this section, two basic functional surface characteristics of pavement, namely pavement roughness and skid resistance of pavement have been discussed.

14.3.1 Pavement Roughness

The objective of roughness measurement is to obtain a single or a number of parameters characterizing the level of roughness of a given stretch. A road profile is a two-

dimensional slice of the road surface taken along any imaginary longitudinal straight line; and the profile measurement is a series of numbers representing elevations relative to some reference level. The problem lies in reducing these huge data-points to a representative index called *roughness* [234].

To obtain roughness information from a measured profile, two basic requirements are [200]:

1. The profiler must be capable of sensing the relevant information present in the true profile of the road.

2. A suitable algorithm must be able to process the measured values to extract the desired information as the summary roughness index.

A *profiler* is an instrument used to produce a series of numbers to represent a profile. Following contains a brief discussion on profilers.

Various types of profilers

A profiler works by combining the following three ingredients:

(i) A reference elevation

(ii) A height relative to the reference

(iii) A longitudinal distance

Some types of profilers are [234]:

(i) *Beam static profilometer.* It uses a precise linear vertical distance transducer to measure the profile. The beam is manually moved along the road to record the profiles.

(ii) *Dipstick auto-read road profiler.* It contains a precision inclinometer that measures the difference in height between the two supports. The device is 'walked' along the line being profiled. The longitudinal distance is determined by multiplying the number of measurements made.

(iii) *Inertial profilers.* A linear potentiometer, a laser, or an ultrasonic sensor measures the relative displacement between the road surface and any inertial reference. The motion of vehicle frame is measured by double integration of the signal from accelerometers. The longitudinal distance of the instrument is usually obtained from the vehicle speedometer. The inertial profiler needs certain speed, that is, it must be in moving state in order to function. The speed is maintained constant for ease of data processing. Profilers based on this principle can be called as *response-type road roughness meters* [234].

Various roughness indices

Some of the indices used to quantify road roughness are:

(i) *International Roughness Index* (IRI). It is calculated from a measured single longitudinal road profile. First, the profile is smoothened with a moving

average of a given base length, to remove local irregularities. Then, the response of a quarter car model, in the form of vertical vibration is added which on dividing by the length of the profile yields, IRI [201].

A quarter car consists of a sprung mass and an unsprung mass, with spring and dashpot configuration as shown in Figure 14.4. Ratios between masses, spring constants and damping coefficients are fixed for a standard quarter car. The simulation model generates vibration response caused to the moving quarter car (at a fixed speed of 80 kmph) by the roughness of road profile. The analysis of the response of a model vehicle (quarter car, in this case) due to road roughness is, however, beyond the scope of this book.

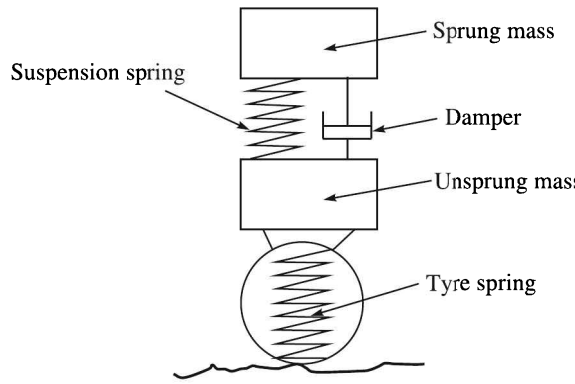


Figure 14.4 Quarter car model.

(ii) *Mean Panel Rating* (MPR). The concept of Mean Panel Rating (MPR) evolved out of AASHO road test. It is the average of ratings given by a panel of pavement experts while driving over a given road stretch. These ratings are processed statistically to yield a single rating, for the panel as a whole, which is called *Mean Panel Rating* (MPR). Thus, MPR gives an idea about the average degree of discomfort of riding due to roughness over a given stretch of road. Panel ratings depend strongly on the instructions given to the members of the panel to define as to which physical property or quality is to be judged. Thus, MPR is a subjective judgement of road roughness.

(iii) *Profile Index* (PI). This index is calculated in a similar fashion as the quarter car simulation used in IRI computation. However, the ratios between the masses, spring constants, damping coefficients are chosen different in this case. The root mean square value of the response profile of the quarter car normalized to the scale between 5 (perfectly smooth) to 0 (maximum possible roughness), is referred as PI [200].

(iv) *Root Mean Square Vertical Acceleration* (RMSVA). The rate of change of slope of the measured profile is the spatial vertical acceleration. The root mean square of this spatial acceleration is the RMSVA.

(v) *Waveband Indices.* A road profile is assumed to be comprised of short, medium, and long wavelengths. The Power Spectral Density (PSD) distribution plotted against the wave number (the number of waves in unit length) gives some quantitative idea of the roughness level. The nature of this plot remains the same for various roads and the area enclosed is observed to be linearly proportional to IRI.

Closing remarks

A number of profilers are used to measure roughness and also a number of indices have been proposed, most of them are apparently uncorrelated to each other. Roughness information can be derived from the true profile of the road, as well as from the vibration response of the vehicle plying on it. It may be argued that the study on the vehicle vibration response could act as a better roughness index compared to that on the true profile of the surface, because it is the vehicle vibration and the related discomfort which a road user is more concerned with. Different vehicles would show different vibration responses, and that is why a quarter car model (with fixed ratios of mass, damping coefficient and spring constants) is chosen as a standard vehicle for the roughness study. For this reason, IRI, as the roughness index, is gaining acceptance in most of the countries.

14.3.2 Skid Resistance

The *skid resistance* is the retarding force generated due to interaction between the pavement and locked tyre when the vehicle is moving. *Skid number* is defined as 100 times the frictional coefficient between the wheel tyre and the pavement surface.

$$\therefore \quad \text{Skid number} = 100 \times \text{coefficient of friction} \tag{14.1}$$

Factors affecting skid resistance

The following are the factors affecting the skid resistance of pavement surface:

(i) *Aggregate quality.* Aggregate polish reduces the skid resistance. Hard aggregates, which are fine grained, sometimes show tendency to get polished quickly, compared to softer and coarse-grained aggregates [187]. Thus, a compromise needs to be made between the durability and the desired skid resistance.

(ii) *Binder.* Binders which are soft or temperature susceptible may cause bleeding, thereby reducing the skid resistance.

(iii) *Climate.* For similar reasons, climate affects the skid resistance. In hot climate, as bitumen softens the chances of bleeding increase. Hence, the skid resistance decreases.

(iv) *Surface drainage.* If the surface drainage system is not proper, water accumulates on the surface and causes loss of skid resistance. In such a

situation, tyre tread and the surface texture may not be sufficient to drive away the water and the tyre may start slipping on the water film. This phenomenon is called *hydroplaning*.

Measurement of skid resistance

Skid resistance can be measured by a portable skid tester, known as the British Pendulum Tester. Figure 14.5 shows the photograph of a portable skid tester developed by Transport Road Research Laboratory (TRRL), UK. The test uses a pendulum, as a spring loaded rubber slider. The pendulum is released from horizontal position, and it slides over the specimen whose skid resistance is to be measured. The scale attached to the pendulum measures the energy lost, and the friction coefficient of the object is estimated by the following formula [111]:

$$f = \frac{W \times X \times Z}{P \times D \times p} \times 100 \tag{14.2}$$

where

f is the coefficient of friction (expressed as percentage)
W is the weight of the swing arm
X is the distance of effective centre of gravity of the swinging arm from the centre of oscillation
Z is the vertical distance of the edge of the scale below the zero of the scale
P is the load on the slider
D is the sliding distance
p is the length of the arm of the pointer.

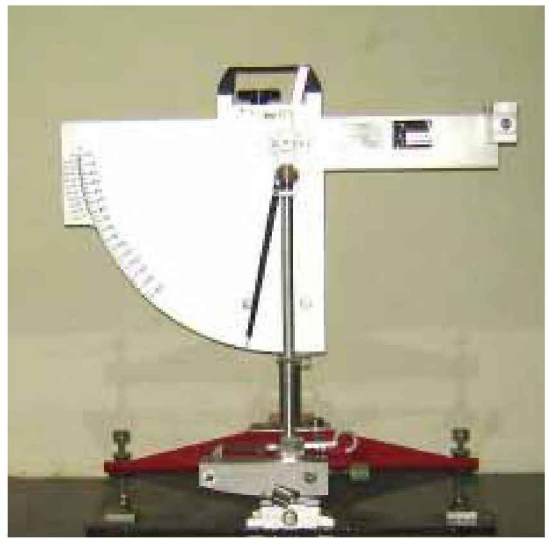


Figure 14.5 A portable skid tester.

The pendulum skid tester can measure the skid resistance for small areas only, and it takes time for each measurement, then it may be difficult to use for network level evaluation of skid resistance. Also, it requires regular calibration. Other devices are also available, such as locked wheel trailer and yaw mode trailer, which can measure the skid resistance of a long stretch of road, mounted on a vehicle moving at normal traffic speed. In a locked wheel trailer, a two-wheel trailer whose wheels are locked, is pulled in the forward direction. The locking force is measured and the skid number is obtained. In the yaw mode equipment, the locked wheels of the trailer are turned at a specific angle to simulate the effect of turning. Mu-meter is a skid testing equipment based on this principle. Skid resistance can also be measured by the stopping sight distance method, where the wheels of a running vehicle are locked, and the braking distance is measured.

EXAMPLE 14.1

A force of 25 kN is required to pull two locked wheels at a speed of 40 kmph, loaded with 50 kN of weight. Calculate the skid number.

Solution

$$\text{Skid number} = 100 \times \frac{25}{50} = 50, \quad \text{at a speed of 40 kmph}$$

Improvement of skid resistance

For bituminous pavement, the skid resistance of the surface can be improved by putting a thin layer of small aggregates, sand, and binder. Slurry seal, surface dressing, fog seal, sand blinding, etc. can also improve the skid resistance of the surface. Excess bitumen, if any, which comes out due to bleeding can be taken out by heating and cutting. Grooving on a bituminous pavement surface can also be done, which is effective only in cold climate. *Grooving* is a technique by which shallow, narrow channels are made on the surface of the pavement by means of a narrow rotating diamond saw-blade [266]. For the concrete pavement, the skid resistance of the surface can be improved by acid etching, or grooving. Grooving on concrete pavements is more effective than that on bituminous pavements, and transverse grooving is preferred to longitudinal grooving.

14.4 STRUCTURAL EVALUATION OF PAVEMENT

The structural evaluation of pavement can be broadly classified into two major categories, namely destructive evaluation and non-destructive evaluation (NDT) of the pavement. In non-destructive evaluation, the structural strength of the pavement is evaluated without causing any damage to the pavement or disruption of traffic. Destructive and non-destructive evaluation of pavement are briefly discussed in this section.

(a) In destructive evaluation, samples are retrieved from the pavement and analyzed in the laboratory. The pits dug in the pavement give a measure of the thickness of various layers of the existing pavement, which is turn gives an idea about the field compacted thickness compared to the thickness which was originally laid.

Bitumen extraction is generally employed to check the bitumen content and aggregate gradation used in a pavement construction. The sample taken out (by core cutter or from pits) from the in-service pavement, is broken into pieces, and put into a centrifuge bitumen extractor, where bitumen is dissolved in a solvent (tricholoroethylene, benzene, methylene chloride, and so on) and is separated out from the mix by the action of centrifugal force. The quantity of bitumen is measured after the solvent is evaporated, and this gives an idea about the quantity of bitumen used in the actual construction. The aggregate proportions are also checked by sieve analysis. Corrections are made for the amount of fines which goes out along with the dissolved bitumen during the extraction process. Also, necessary corrections are made for the water content, if present, in the mix. For concrete pavements, the beam samples taken are tested for their flexural strength and crushing strength [230].

The physical properties of bitumen and aggregates are tested if required, such as estimating the suitability of recycling.

(b) A number of NDT devices have been developed for the structural evaluation of pavement. The NDT equipment is used to determine the (i) in-situ moduli of pavement layers, (ii) load transfer efficiency at joints in the concrete pavements, and (iii) location and extent of void in a pavement structure. The NDT equipment used for pavement evaluation is broadly classified into four major categories on the basis of their type of loading, as briefly discussed below:

- **Static creep deflection method.** In this type of equipment, a static load is applied to the pavement and the deflection is measured. Benkelman beam test is such an example. A multiprobe Benkelman beam measures the static deflection at a number of points. California travelling deflectograph and LaCroix deflectograph are some examples of equipment under this category. Pavement evaluation by Benkelman Beam (BB) is covered in Section 14.4.1.

  The problem with the static creep kind of equipment is that (i) the static (or slow moving) deflection response measured with this equipment is different from the deflection response of the moving wheels in the in-service pavement, and (ii) the fixed reference with respect to which deflection is measured, may also be a part of the deflection bowl, hence may give erroneous readings.

- **Steady state deflection devices.** The NDT devices that fall in this category, measure the deflection response of the pavement to a low frequency oscillatory load. The Road Rater and Dynaflect are the two such devices. The basic operating principle of these devices is to impart a vibratory loading by means of some eccentric loading mechanism and to measure the deflection caused to the pavement at a series of points through velocity sensors. The fixed point referencing problem, as in the static creep method, is taken care by this equipment with the use of inertial reference (velocity sensors). However, the steady state loading applied to the pavement does not correspond to the actual form of loading applied by the vehicles [266]. Figure 14.6 presents a schematic diagram of a Road Rater.

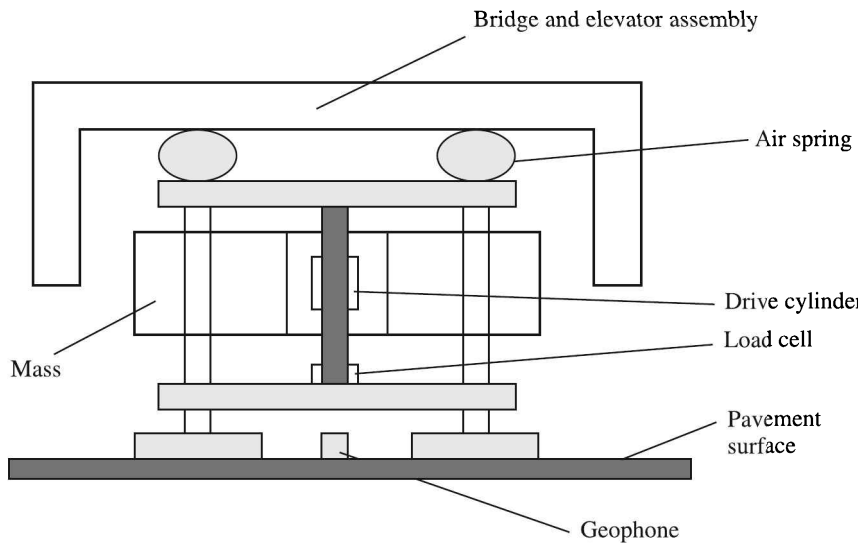


Figure 14.6 A schematic diagram of a Road Rater.

- **Wave propagation devices.** The vibrations propagate through layered media at various speeds. Individual waves have different reflectivity characteristics. The sensors, like, geophone or accelerometers, placed at a distance, sense the arrival of various waves, and the elastic moduli of the respective layers can be estimated therefrom. Among the various available analysis methods, the Spectral Analysis of Surface Waves (SASW) is popularly used for pavement evaluation.
- **Impulsive loading devices.** The impulsive loading type NDT devices apply an impulsive load to the pavement and record the resulting pavement deflections at several radial distances from the load application point, during a short loading time. The duration and the impulsive nature of loading closely simulate the nature of loading imparted to the pavement by the

vehicles. The examples of impulsive loading devices are Dynatest, Phoenix, and KUAB FWDs. These devices generate impulsive load through rapid deceleration of the falling mass. The deflections caused to the pavement are measured with the help of velocity sensors, that is, geophones. The peak deflections at each measurement location constitute the deflection basin. In a multidepth deflectometer, deflections at various depths are measured by installing sensors at various depths.

14.4.1 Benkelman Beam

Benkelman beam was devised by A.C. Benkelman as a deflection measurement test of bituminous pavement for the WASHO road test in 1953 [74]. The Benkelman Beam Deflection (BBD) technique is a popular test all over the world for estimating the required overlay thickness. The popularity is possibly because of its simplicity and low cost. The permissible maximum allowable Benkelman Beam deflection for satisfactory performance of a road stretch depends upon the traffic, material of construction, and the environmental factors. This forms the basis of the BBD study. Benkelman deflection more than the allowable deflection suggests that the pavement may require an overlay.

In India, the earlier guidelines [73] on strengthening by overlay using the BBD method, have been revised, and the present guidelines [74] have evolved from a broader perspective of experience gained through research and practice [61] in India and in other countries.

Principle of BBD study

A conceptual working of a Benkelman beam is depicted in Figure 14.7. A′B′ represents the position of a Benkelman beam when the probe A′ is placed between the dual wheel of a loaded truck. The point A′ touches the maximum deflected point of the deflected bowl. When the truck moves forward by a given distance (from P′ to P), the deflection bowl also moves forward, and the probe point A′ comes back to a point position A. This

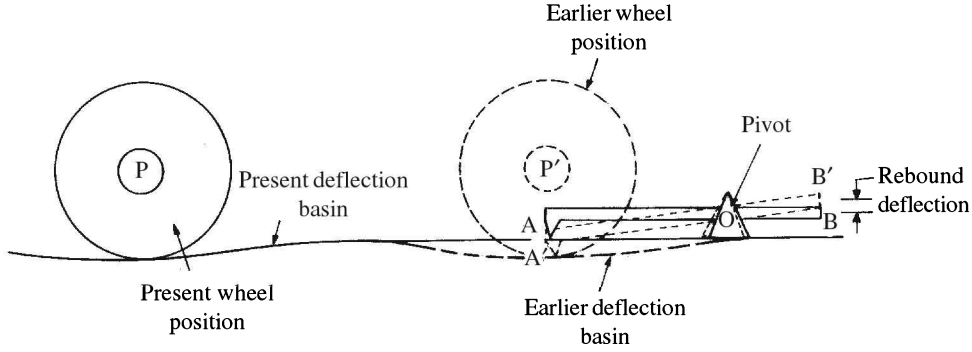


Figure 14.7 Conceptual working of a Benkelman beam.

deflection is called the *rebound deflection*, and is used for the estimation of overlay thickness. There may be some residual deflection at the present position of the truck wheel (at P) and therefore, the truck is further moved forward to measure the residual deflection. As mentioned earlier, if the deflection bowl has a large spread, the pivot (O) may itself fall within the deflection bowl, giving erroneous results. This error, to some extent, is taken care by incorporating corrections in the observed reading. This aspect has been explained in the example problem. In dynamic equipment (steady state or impulse type), however, the velocity of movement of deflection bowl is measured, which on integration over a time period gives the maximum deflection of the deflection bowl, without any error of this kind. In Benkelman beam, the length of AO is double that of OB; the dial gauge being placed at B, and the rebound deflection of the pavement is twice the reading obtained by the dial gauge.

The reader may note that the diagram of the Benkelman beam shown in Figure 14.7 is of conceptual nature only. Figure 14.8 shows the diagram of an actual Benkelman beam drawn to scale. It may be noted that in this diagram the lengths of the arms are 2.44 m and 1.22 m, thereby maintaining a 2:1 ratio.

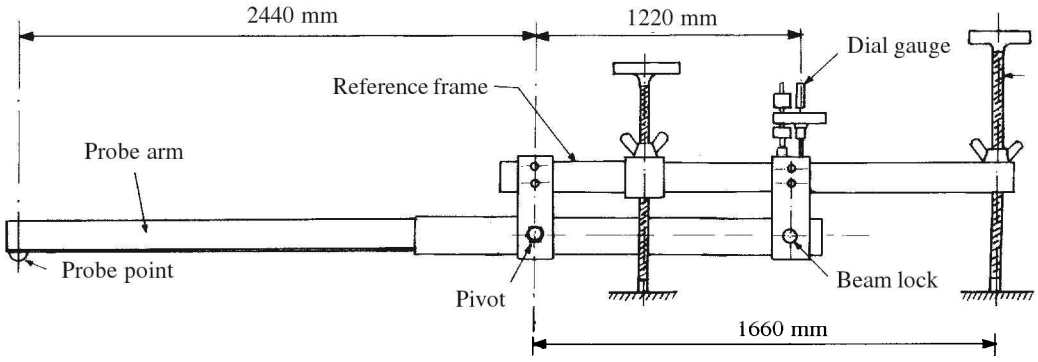


Figure 14.8 Diagram of a Benkelman beam [73].

Deflection measurement by Benkelman beam method

To collect the BBD data of a road section, under the Indian guidelines IRC:81-1997, at least 10 equidistant points in each lane (the interval between the points should not be more than 50 m) are marked on the pavement along the outer wheel path [74]. Further instructions for selecting the points depending on the type of road are also available in the guidelines.

Indian guidelines have adopted the CGRA (Canadian Good Roads Association) method of BBD evaluation of pavement, which is described as follows. A standard loaded truck, with rear axle weighing 8100 kg and fitted with dual tyre each having a tyre pressureof 5.6 kg/cm^2, is used in the BBD study. The dual wheel of the truck is centred above the selected point. The probe of the Benkelman beam is placed between the two wheels. The lock of the Benkelman beam is removed and the beam is checked

for its free movement. The truck driver is asked to slowly move 2.7 m from the selected point and stop. The dial gauge reading for the corresponding deflection is noted when the recovery of the pavement is less than or equal to 0.025 mm/minute; this reading is called the *intermediate reading*. The truck is moved forward by another 9 m, and the final reading is taken. Pavement temperature is also recorded every hour by inserting a thermometer in the standard hole filled with glycerol. The difference between the final and the initial dial readings and also the difference between the intermediate and initial readings, are both calculated. If the difference of values lies within 0.025 mm, then the actual pavement deflection is twice the final differential reading. If it is not so, then, a term *apparent pavement deflection* is defined as twice the final differential reading. The true pavement deflection is equal to the apparent pavement deflection plus 2.91 times twice the difference between the final and intermediate dial readings [74]. The design of overlay making use of the BBD readings has been described in the next section.

14.4.2 Falling Weight Deflectometer

In the Falling Weight Deflectometer (FWD) test, an impulsive load with a short loading time is applied on to the road surface by means of a weight falling on a set of springs (see Figure 14.9). With proper choice of the drop weight, spring constant, and the falling height, a representative impulsive load simulating a real traffic load can be obtained.

If a truck is assumed to move with a speed of 60 kmph, and its radius of tyre imprint

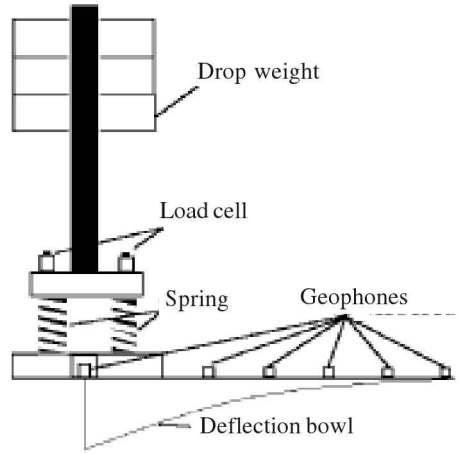


Figure 14.9 Schematic diagram of FWD.

is assumed to be 15 cm, then the time of contact is calculated as $(2 \times 150)/(60 \times 10^6/3600) = 0.018$ s, which is approximately of the order of 0.02 s. Thus, the loading duration of FWD is so designed that it approximately equals 0.02 s.

Principle of FWD study

In FWD study, the deflection of the pavement surface is measured at a number of points at different distances situated radially outwards from the centre of the falling weight. The generated response is usually measured by velocity transducers (geophones) and after the velocity time response is integrated, the values of instantaneous pavement deflection at a number of points are obtained. The test is repeated several times at a particular location and the results are averaged to reduce random errors. If required, the test may also be done with different loads to evaluate the stress dependence of the layer modulus [210]. Figure 14.10 shows a photograph of the FWD test being carried out in the field.

Figure 14.10 FWD testing in progress: National Highway 1.

Back-calculation of layer moduli from the FWD deflection profile

In the FWD test, six or seven discrete surface deflection readings represent the deflection basin. If the behaviour of the pavement under impact loading is assumed to be elastic, the pavement response can be described by knowing only the elastic modulus E and the Poisson's ratio μ of each layer. Some standard μ values may be assumed from the literature, since these have little effect on the stress analysis [162]. Each layer is thus, represented by only one unknown, that is, the elastic modulus of the layer. The purpose of the FWD study is to find out the in-situ elastic modulus of the layers, when the deflection basin is known from FWD testing. The process of estimating unknown elastic moduli from known deflection basin is known as *back-calculation*. Therefore, the minimum number of surface deflection readings needed in back-calculation process must be at least equal to the number of layers to avoid non-unique solution [37]. Because of the rounding off and truncation errors introduced during back-calculation, it may not be possible to reproduce exactly the original layer moduli from a basin generated by a linear elastic solution. Also, for the deviation of material behaviour from the linear elastic

model no solution may exist which matches the measured basin perfectly [37]. The division of a pavement structure into many layers may produce a non-unique solution whereas assuming fewer layers may not be able to reach a solution which matches the measured deflections. Some researchers have reported that there is no unique solution to the set of moduli that would produce exactly a given deflection basin [96]. The thicknesses of the different layers also form an important input to the back-calculation, otherwise a realistic match may not be achieved [210]. Thicknesses may be measured accurately by coring, boring, ground penetration radar, and seismic tests [258] treated as unknown parameters.

As mentioned, the basic philosophy of back-calculation is that when the computed surface deflections match the measured deflections, the resulting layer moduli are considered to be the most appropriate material moduli for the pavement structure [37]. The process is initiated by assuming 'seed values' for elastic moduli of the pavement layers and comparing the resulting deflections (through a pavement analysis, i.e. forward calculation routine) with the measured ones. Adjustments of the elastic moduli are made until the difference between the two deflection profiles is within a given tolerance. Algorithms for convergence should be carefully adopted, otherwise convergence may not even be achieved, or it may take an unnecessarily long time to arrive at a reliable result [145]. There are several back-calculation algorithms (such as, equivalent half-space method, regression method, database search method, optimization method, and so on) suggested by a number of researchers. The methods are inevitably complex and not unique. Figure 14.11 presents a simple back-calculation scheme.

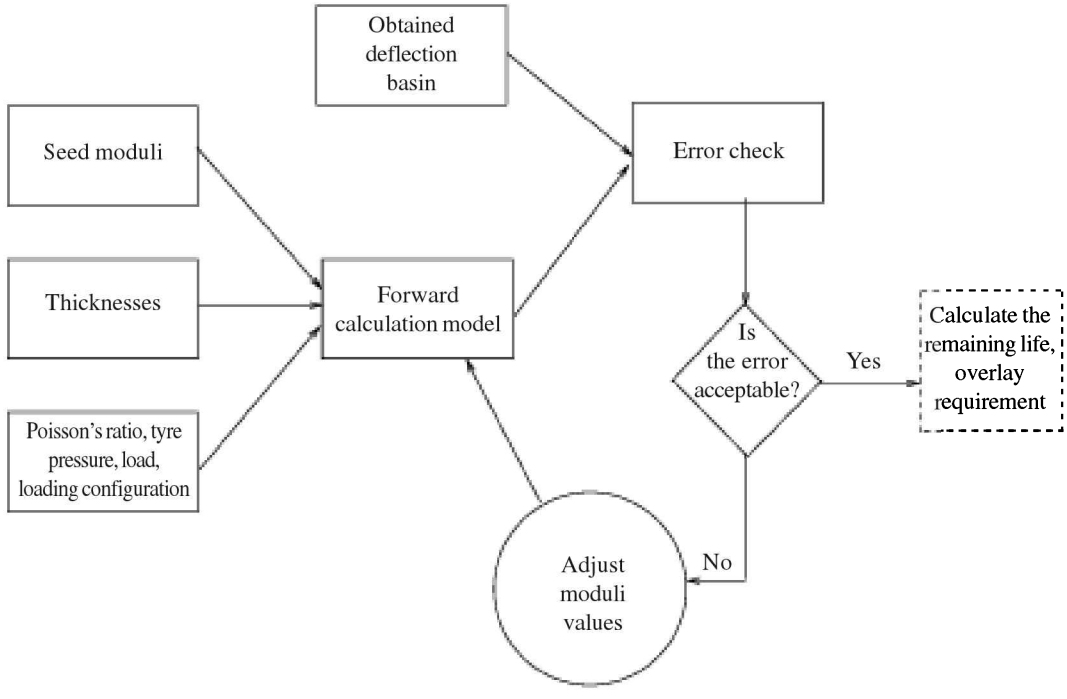


Figure 14.11 A simple back-calculation scheme.

Evaluation of load transfer efficiency of a joint

The load transfer efficiency of a joint can be determined by FWD (or any other impact devices). The location of fall of weight is so adjusted that it is close to the joints of the pavement slabs. The deflections measured in the two slabs close to the joint give the value of load transfer efficiency of the joint. Figure 14.12 shows a diagram of two idealistic extreme situations where the efficiency of the joint is 0% and 100%, respectively. If due to the application of load close to the joint, both the adjacent slabs deflect by the same amount, the joint efficiency is 100%, and similarly, if the other slab (which is not loaded) does not deflect at all, its joint efficiency is 0%.

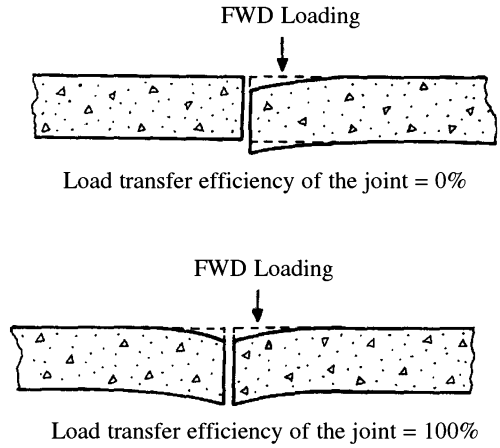


Figure 14.12 Joint efficiency test by FWD.

14.5 PAVEMENT MAINTENANCE

A distressed pavement requires maintenance. Maintenance measures constitute fresh investment on the existing roads. There are two considerations which are of importance in this regard:

1. The maintenance expenditure can be reduced through proper planning, design, construction, and quality control. If the causes of possible distresses are removed, or judiciously taken care of during design, the expenditure due to maintenance measures on in-service roads reduces. For example, if the drainage provisions are designed properly, or, overloading beyond the legal limit is strictly prevented, the premature pavement distresses can be avoided.
2. It is advisable to implement the necessary maintenance measures at an early stage when the distresses have just started showing up. It is seen that proper pavement maintenance measures at the early onset of distresses, can obviate major maintenance expenditure in future. This is because, in general, the rate of deterioration increases with time.

This section briefly introduces various pavement maintenance measures under two categories, namely maintenance measures other than overlay, and maintenance with overlay.

14.5.1 Pavement Maintenance Measures Other than Overlay

The pavement maintenance measures other than overlay are the minor maintenance or repair works which are performed on the pavement. These works do not enhance the structural strength of the pavement, but can improve the functional standards and check the rate of deterioration. These maintenance measures can be of routine type or periodic in nature. A brief explanation of surface repair and drainage maintenance measures is given below.

Surface repairs

Surface repairs are effective when discrete damaged patches (say, potholes, local depressions) exist on a pavement surface which need immediate repair. For surface repairs, if needed, the existing bituminous layers of the specified area are carefully scarified without causing any disturbance to the other layers. Tack coat is applied to ensure good adherence. Granular layer and bituminous layer, as the situation demands, are laid and compacted (see Figure 14.3).

Table 14.2 briefly presents the various repair works recommended against various forms of distresses. Some of them have already been discussed while introducing the various forms of distresses. The reader may note that the table is neither exhaustive nor does it express the only available solutions. The actual repair technique needs to be evolved on a case to cases basis. Also, Table 14.2 presents only the repair techniques, not the preventive measures, which have already been covered in Section 14.2.

Table 14.2 Possible surface maintenance measures for some pavement distresses

Type of distress	*Maintenance measures*
Block cracking	Application of new bituminous coat recycling
Bleeding	Sand blotting/sand blinding
Corrugation	Scarification of elevated part by mechanical blades and rolling
Depression	Application of profile corrective course
Fatty surfaces	Application of hot, dry, small aggregates, and rolling
Hungry surface	Application of fog seal, slurry seal
Loss of aggregates	Application of seal coat, fog coat, or surface dressing
Polished stone	Surface dressing or other suitable form of wearing coat
Pothole	Patching and partial reconstruction
Ravelling	Seal coat, fog coat, or laying of renewal coat
Rutting failure	Milling of protruded portion, profile corrective course recycling
Slippage	Replacement of top wearing coat with proper tack coat
Stripping	Replacement of affected layer with fresh mix
Swell and blow up	Milling of protruded portion, construction of drainage facility

Drainage maintenance

The drainage system provided in the road needs routine (or periodic) attention to check their proper functioning. The camber and the shoulder slopes need to be maintained properly for satisfactory functioning of surface drainage. Depressions, potholes, and rutting should be repaired immediately with premix aggregates to check the accumulation of water and subsequent damage to the pavement. Clogging of open longitudinal drains due to debris accumulation needs regular checking. The sub-surface drainage network should also be inspected regularly for clogging.

14.5.2 Pavement Maintenance with Overlay

The overlay is the extra thickness provided on the pavement surface which strengthens the pavement structurally, and thereby enhances its longevity. The overlay design comprises the determination of thickness and the type of material to be laid over the existing pavement surface so as to extend its longevity by a given period. Earlier (prior to 1960), the overlay design used to be based on judgment and experience [266]. There are various overlay design methodologies in vogue now and among which at least three basic approaches may be identified as follows:

(a) Effective thickness approach
(b) Deflection approach
(c) Mechanistic approach

The principle of effective thickness approach has already been covered with the stage construction considerations in Section 12.8.2. The overlay design by the BB method is based on deflection approach, and that by FWD is based on mechanistic approach. These two approaches are now discussed in the subsequent paragraphs of this section.

Selection of homogeneous sections

When the BBD survey data is collected over a long stretch of road, there is a need to subdivide the stretch into a number of (possibly unequal) parts, where the deflection records are somewhat the same in their order of magnitude. There is no specific methodology suggested in Indian guidelines [89, 74] for this segmentation, except doing it by visual observation of kilometre-wise plotted data. However, it is suggested that the minimum length of the section should be at least one kilometre [74], otherwise it becomes inconvenient from the construction point of view if the overlay thickness recommendation changes even for a fractional length of a kilometre.

A simple method suggested by ASSHTO [2] can be adopted for such a situation. According to this method, the cumulative data points are plotted on a kilometre scale, as shown in Figure 14.13. The best fit straight line is drawn through all the data points. Wherever the data points change their location from one side of the best fit line to the

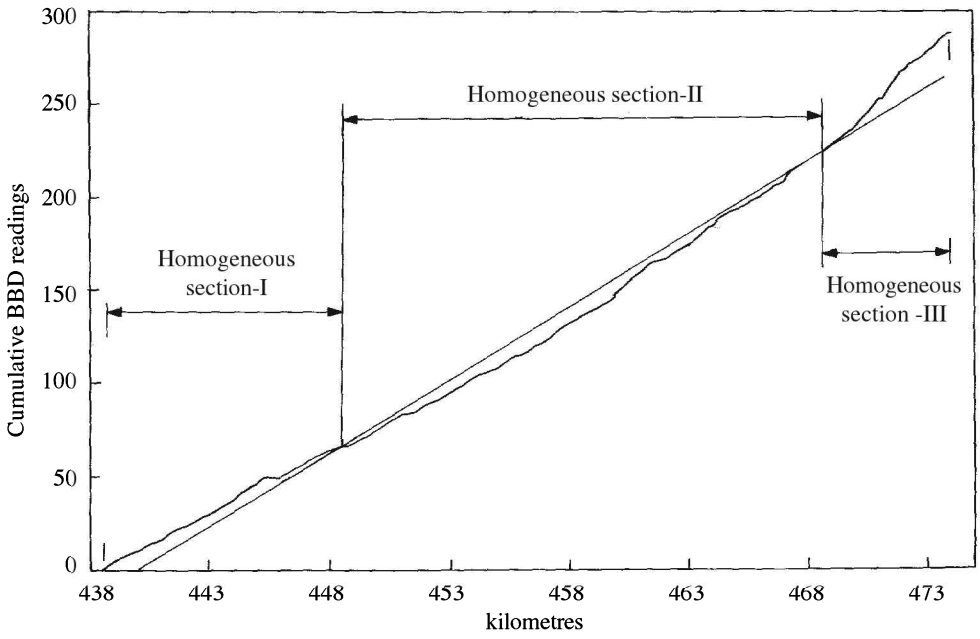


Figure 14.13 Determination of statistically homogeneous section.

other, it can be marked as the start of another homogeneous section[1]. Thus, the BBD data points observed from km 438 to km 474 of a particular road stretch have been delineated into three parts by this method as shown in Figure 14.13. These three stretches can further be referred for overlay design, individually.

BBD method

The following example illustrates the overlay design method as per the IRC:81–1997 [74] guidelines.

EXAMPLE 14.2

The following are the BBD, field moisture content, and temperature readings at equidistant points obtained along a stretch of a major road. If the pavement is to sustain further 20 msa of traffic repetitions, design an overlay thickness for the stretch. The average annual rainfall of the area is found to be 1200 mm, and the soil is of clayey nature, with average plasticity index 12.

---

[1]In fact, the requirement of delineation of data points into homogeneous stretches may arise, not only for BBD survey, but also, in various parameters associated with pavement design, such as, roughness data, CBR data, plate load test data, and so on. A good reliability of the overall design, subjected to a given fund constraint, can be achieved depending on how successfully the delineation has been done.

Solution

The actual pavement deflections are calculated according to the IRC:81–1997 recommendations. Temperature and moisture corrections (data given in Table 14.3) are applied and the final corrected deflections are found out as shown in Table 14.4.

Table 14.3 Data to find temperature and moisture corrections

Sr. no.	*Pavement temperature* (°C)	*Moisture content* (%)	*Dial gauge reading (mm)*		
			Initial	*Intermediate*	*Final*
1	35	10	0.00	0.54	0.56
2	35	11	0.00	0.54	0.54
3	36	10	0.00	0.53	0.53
4	36	10.5	0.00	0.50	0.51
5	36.5	10.5	0.00	0.48	0.49
6	35	10.5	0.00	0.46	0.49
7	34	11	0.00	0.50	0.51
8	34	10	0.00	0.57	0.57
9	34	10	0.00	0.54	0.56
10	35	10	0.00	0.53	0.54

Table 14.4 Corrected deflections

Sr. no.	*Actual deflection*	*Correction factors*		*Corrected deflections*
		Temperature#	*Moisture*†	
1	1.12	0	1.23	1.337
2	1.08	0	1.19	1.285
3	1.06	– 0.01	1.23	1.292@
4	1.02	– 0.01	1.20	1.212
5	0.98	– 0.015	1.20	1.158
6	1.15*	0	1.20	1.380
7	1.02	+0.01	1.19	1.225
8	1.14	+0.01	1.23	1.415
9	1.12	+0.01	1.23	1.389
10	1.08	0	1.23	1.328

Note:

*This value is calculated as $2(0.49) + 2.91 \times 2 \times (0.49 - 0.46)$, and, has already been discussed in Section 14.4.1.

#Temperature correction is +0.01°C if it is in the higher side of 35°C.

†Moisture correction factors are obtained from Figure 14.14; there are six such curves available in IRC:81–1997 [74] depending upon the type of soil and rainfall.

@sample calculation: $(1.06 - 0.01)1.23 = 1.2915$; deflection values are first corrected for temperature, then multiplied by the moisture correction factor.

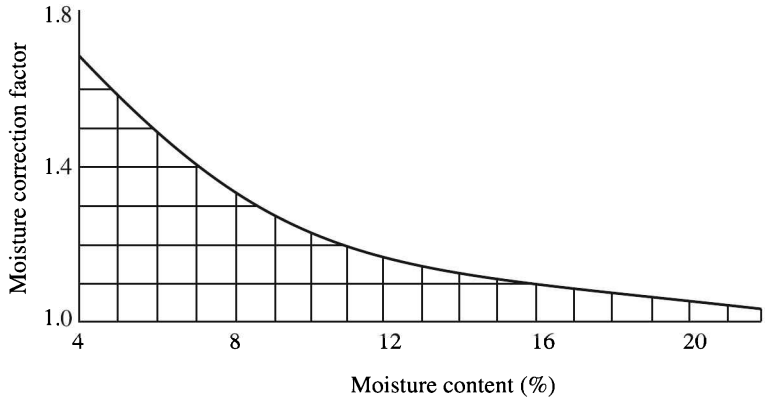


Figure 14.14 Moisture correction factor for clayey subgrade with low plasticity (PI < 15) for low rainfall areas (annual rainfall ≤ 1300 mm) [74].

Mean deflection is obtained as 1.302 mm, and the standard deviation as 0.0839 mm. Therefore, the characteristic deflection is

$$1.302 + 2(0.0839) = 1.469 \text{ mm}$$

The multiplicative factor 2 in standard deviation is chosen as it is a major road. From the overlay design chart (see Figure 14.15), the required overlay thickness is found to be

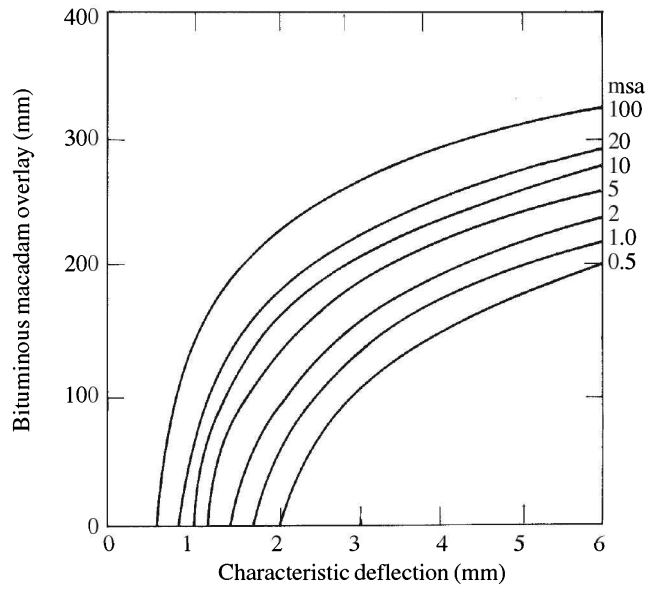


Figure 14.15 Overlay thickness design chart as per Benkelman Beam method [74].

146 mm of BM. This thickness can be converted to equivalent thickness of other layers by the empirical relationship, or by equivalency (see Eq. (12.18)).

FWD method

The overlay design by the FWD method is a three-stage process involving:

(i) FWD study
(ii) Back-calculation of layer moduli
(iii) Estimation of overlay thickness

The first two stages have been discussed in Section 14.4.2. Now, after the layer moduli are obtained, the mechanistic pavement design principles are applied to find out the necessary overlay thickness. For example, as shown in Figure 14.16, the existing pavement structure is of three layers, and it is analyzed as a four-layered structure when the overlay is put for enhancement of its longevity by a given 'msa' level. The overlay requirements in terms of extra bituminous concrete thickness are determined for various traffic levels, and the existing granular layer thickness is evaluated from fatigue and rutting considerations.

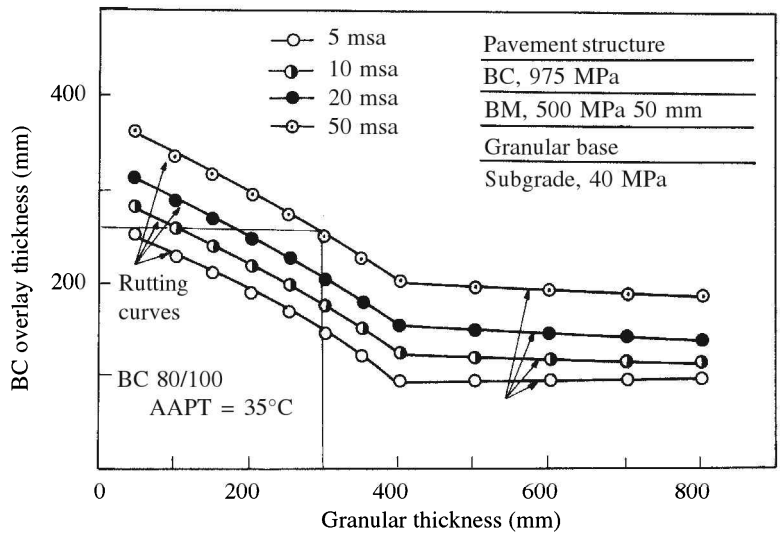


Figure 14.16 An overlay design chart obtained from FWD study.

Closing remarks

- It is not always the maximum deflection but curvature, too, is another important parameter required for determination of the overlay thickness. Pavements which have the same design life with different pavement compositions and subgrade CBR, yield different deflections under the standard axle load. The

maximum BB deflection value, therefore, cannot always be the right criterion for overlay design. The deflected profile of the road should also be taken into account for overlay design.

Thus, multiprobe Benkelman Beam equipment has been evolved, in which deflection at various points is also measured to derive some information about the curvature of the deflection bowl. For example, in Austroads [182] method, a parameter $D_0 - D_{200}$ is used for estimating the overlay thickness, where D_0 and D_{200} are the maximum deflection and the deflection at 200 mm respectively, radially outwards from the point of maximum deflection.

The BBD method is popular because of its low cost and easy test procedure, but the static loading nature and the difficulty in getting a fixed reference point for deflection measurement are the shortcomings associated with this method.

- In a separate study, the overlay thicknesses derived from the BBD method and the mechanistic method are compared [47] for various points of some selected stretches in India, as shown in Figure 14.17. It is interesting to note that though the basic approaches of the BBD and the mechanistic method of overlay design are different, the final overlay recommendations are comparable to each other.

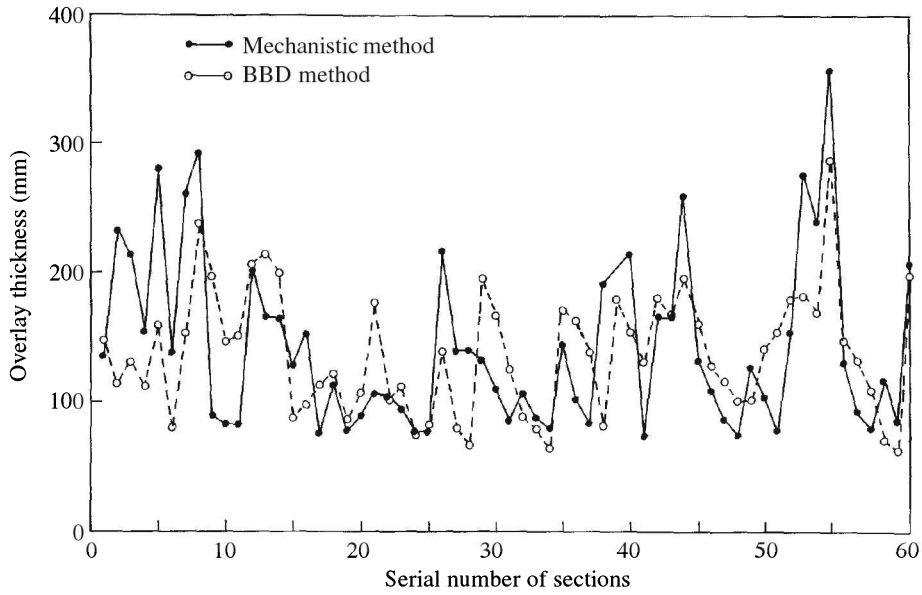


Figure 14.17 Comparison of overlay thickness obtained from BBD [74] and mechanistic pavement design method [47].

- Overlay is discussed here as the structural rehabilitation procedure. In overlay, extra thickness is laid on the existing pavement to extend its longevity.

However, overlay construction increases the height of the pavement. This problem is particularly acute in city streets where road level keeps on rising causing inconvenience to the roadside establishments. The best solution in this case is to recycle the materials of the existing pavement and use them for overlay construction. Recycling is a better rehabilitation method than putting new overlay on the existing surface, as it conserves aggregates, binder and energy, preserves the environment and road geometrics [2]. A brief discussion on bituminous pavement recycling has already been presented in Section 13.10.3.

14.6 MAINTENANCE MANAGEMENT

Figure 14.18 shows a schematic diagram of variation of pavement condition (structural, or functional, or combined) with respect to time. If no rehabilitation measures are taken, the pavement gradually deteriorates and fails at a certain stage, as shown in the figure. Rehabilitation improves the condition of pavement, extends its life, and thus, prevents its failure after the expiry of initial design period. These can be alternative rehabilitation measures (in terms of their extent and frequency), as shown in the figure. Depending upon the frequency and type of rehabilitation, pavement continues to serve satisfactorily for an extended period. The designer has to judiciously recommend suitable rehabilitation measures, chosen from various possible alternatives, such that the fund utilization is optimal and the condition of pavement at any given point of time remains satisfactory. This may be referred as optimal maintenance requirement of a particular stretch.

Similarly, the maintenance requirements of various individual pavement sections of a given road network can be different. Due to fund constraints, it may not be possible to implement

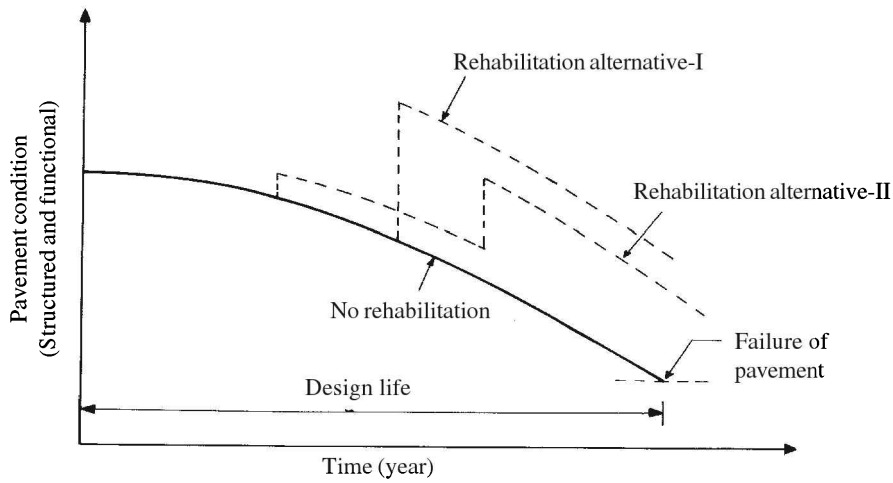


Figure 14.18 Schematic diagram showing pavement condition trends for alternative rehabilitation measures.

the necessary maintenance measures for all the pavement sections of the road network. Also, while the rehabilitation measures are being implemented in some of the sections in the network,

other sections may experience further deterioration. These aspects are covered under pavement maintenance management which therefore includes:

(1) Quantification of structural and functional distresses of the pavement through appropriate indices.
(2) Prioritization of maintenance needs for various pavement stretches of the network.
(3) Forecasting future performance, in terms of the selected indices, with and without maintenance.
(4) Allocation of funds for maintenance measures in different pavement stretches in the network.
(5) Scheduling of maintenance activities in various stretches of the network.

EXERCISES

1. What is reflective cracking? How do you check its propagation? Are the alligator cracks basically fatigue cracks by origin?
2. What is difference between pavement rutting and depression? Do corrugation and ravelling mean the same distress? Explain.
3. What are the causes of pavement bleeding?
4. What are the various reasons for pavement cracking?
5. What is meant by functional failure of a pavement?
6. What are IRI and profiler? Is IRI profiler dependent?
7. What is the difference between the vibratory type and the falling weight deflectometer type machines used in evaluation of pavement?
8. What are the sensors used in FWD? What do they sense?
9. What are the corrections applied to the Benkelman Beam Deflection (BBD) reading? What are the data necessary for these corrections?
10. How do you design overlay thickness from Benkelman Beam deflection reading? Mention only the steps.
11. Is it possible to derive (i.e. back-calculate) layer elastic modulus by BBD testing? Explain.
12. FWD test is used to back-calculate the elastic moduli of pavement layers. How does that help in pavement overlay design? Explain.
13. In a field trip to a particular road section, during August 2002, the following things were observed:
    (a) There were large and small depressions on the road.
    (b) There were many places where alligator cracks were seen.
    (c) Deflections discernible to the naked eye were observed when heavy vehicles passed over the depressed areas.

If you were asked to investigate the possible reasons for failure of the road, how would you proceed? Write the steps only.

PART V

TRANSPORT ECONOMICS

15 Highway Economics and Finance

15.1 INTRODUCTION

Highway construction is a major civil engineering activity which involves large sums of investment. Construction of a six-lane expressway may cost Rs 5–10 crore per kilometre. A well-designed rural road with 3.75 m carriageway, shoulder, and cross-drainage works may cost Rs 6–25 lakh, depending on the construction conditions [51]. Proper economic evaluation is, therefore, an important aspect of highway projects. Study of highway economics and finance mainly involves the understanding of various cost components of highway projects, the prioritization of various alternative highway projects, decision on the scheme of investment in a project at its various stages, funding sources, and policies for road projects.

This chapter is divided into seven sections. The second section contains a general discussion on the scenario of Indian roads. Some parameters used in economic analysis are defined in the third section while the various cost and benefit components of transportation infrastructure have been discussed in the fourth and the fifth section. The sixth section discusses the economic evaluation of alternative highway projects, based on which the implementation decision is taken. Finally, a brief presentation on highway financing has been given in the seventh section.

15.2 INDIAN ROADS AND PRESENT SCENARIO

Transportation forms a vital part of the infrastructure for economic and social development of a country. Roads meet the bulk of the surface transportation requirement and are important in any country's transportation system. Though India has 60.8 km of road length per 100 sq. km, compared to the USA where it is 66.5 km/100 sq. km, the total road length is over 1.8 million km (1996), out of which only about 45% is surfaced [12].

The present total road length in India including National Highways, State Highways, Major District Roads and Rural Roads is 33 lakh kilometers (2001) [190]. Of

this, the National Highways cover a length of 58,077 km (2001) [190] which is less than 2% of the total road length (2001) [188, 190] but carry 40% of the total traffic.

At the time of independence, India's road network was rudimentary and poor even for the small traffic plying on it. In the last few decades, the length of road network has increased from 4 lakh km in 1951 to 33 lakh km in 2000 making India the third largest road network in the world [190, 94]. During this period, the number of trucks and vehicles has increased 10 to 40 times with consequent increase in the share of passenger (26 to 80%) and goods (11 to 60%) traffic carried by road. Goods traffic carried by road has increased 35 folds and passenger traffic 40 fold in forty years from 1951 to 2001.

At the same time, the number of automobiles has grown from 0.3 million in 1951 to 48.5 in 2001 [171, 184]. The freight traffic has also increased from 6 billion ton-km in 1951 to 1200 billion ton-km in 2000. The total estimated losses in India due to increased Vehicle Operating Cost (VOC) caused by poor road surface are about Rs 15,000 crore per year (1996) [12].

On the other hand, the newly conceived 'Golden Quadrilateral' project, being developed as per international standards is expected to make annual savings of approximately Rs 8000 crore [190]. The Golden Quadrilateral project involves four-laning of a 6000 km road length linking Delhi-Kolkata-Chennai-Mumbai, while the North-South and East-West corridors involve a distance of 7000 km road length linking Srinagar-Kanyakumari and Silchar-Porbandar respectively [190]. Also, 17,712 km of roads, earlier state or district highways, were added to the National Highway by the year 2000 [139]. In order to provide road connectivity to the villages, the Government of India launched (December, 2000) the "Pradhan Mantri Gram Sadak Yojana" (PMGSY), where all unconnected villages having a population of more than 1000 persons will be provided with all weather roads by the year 2003. Villages which have population more than 500 persons will have improved connectivity by the end of the year 2007. The cost of this time-bound project is estimated to be around Rs 60,000 crore.

The huge costs involved in highway projects have various socio-economic implications. Any proposed project is expected to have various alternatives in terms of its way of implementation. The total expenditure involved, and that required at various stages of project execution would be different for different proposals. A judicious selection of the best possible strategy among various alternatives requires a thorough understanding of economic analysis of highway projects and of various sources and policies of highway financing.

15.3 SOME PARAMETERS USED IN ECONOMIC ANALYSIS

This section introduces the reader to various parameters and terminologies that are commonly used in economic analysis of highway projects.

15.3.1 Time Horizon or Analysis Period

The investment for highway construction, maintenance, and its benefits are spread over a time span, called the *time horizon* of economic assessment. This is generally selected as twenty to thirty years for a highway project, depending on the policy or type of road.

15.3.2 Interest Rate

Money earns its interest intrinsically. Interest rate is the return obtained after the end of the year as percentage of the capital invested at the beginning of the year. It can either be at a simple rate or be compound rate. It can be described as an extra cost chargeable to the highway project and is payable to the source which generated the investment. The interest rate is used to calculate the amount of money that would accrue at a future date while the discount rate is used to calculate the present equivalent amount of money, of the amount which will be actually invested in future [212].

15.3.3 Inflation

The construction of a major highway project takes a number of years, and meanwhile the cost of material, labour, and equipment, undergoes price escalation as a result of inflation. At the same time, due to inflation, the Vehicle Operating Cost (VOC) increases, thereby reducing the benefit. Thus while deciding the benefit-cost aspects, the effect of inflation also needs to be considered [212] in all the cost and benefit components.

15.3.4 Salvage Value

Salvage value, S, is the worth of the structure at the end of the analysis period. This value is carried over to the next analysis period. There can be different criteria of calculating the salvage value. If, after the expiry of the first analysis period, it is assumed that the pavement materials would be recycled, then the costs of existing pavement materials (to be used for recycling) are considered in computation of salvage value.

Alternatively, if the pavement life is extended further by overlaying, in the next analysis period, the salvage value can be calculated as [237]

$$S = \left(1 - \frac{Y}{X}\right) O_{n_m} \tag{15.1}$$

where Y is the number of years between the last overlay (which is done in the year n_m) and the analysis period for which O_{n_m} was the cost incurred, and X is the number of years the pavement is originally expected to serve. This is based on the assumption that the service life of the last resurfacing overshoots the analysis period, and accordingly a proportionate salvage value is estimated.

15.3.5 Present Worth

Present worth is the total cost of the project, when investments in various years (during the analysis period) are brought back to the equivalent worth of the present year. The present worth can be expressed in the form of the following equation

$$\text{Present worth} = C + \left[\sum_{k=n_1}^{n_m} O_k \times \frac{1}{(1+r)^k}\right] - S \times \frac{1}{(1+r)^{n_m}} \tag{15.2}$$

where

C is the cost of construction
n_1 is the first year in which major maintenance (say, overlay) is done
n_m is last year within the analysis period in which a maintenance job is carried out
O_k is the cost of maintenance in the kth year
S is the salvage value
r is the discount rate.

15.3.6 Capital Recovery Factor

The concept of capital recovery factor is used only when the recurring investments made at different periods of time are brought back to the equivalent investment made at the beginning of the project. This can be understood in the following way. Take the example of the middle term of Eq. (15.2), using the following simplifications:

(a) Maintenance is done periodically in each year.
(b) The maintenance expenditure is always the same, say x.
(c) The total maintenance expenditure calculated taking the first year as the base year is equivalent to a one-time expenditure y. Then,

$$y = \sum_{k=1}^{N} x \times \frac{1}{(1+r)^k} \tag{15.3}$$

where N is the analysis period. Or

$$y = x \times \frac{(1+r)^N - 1}{r(1+r)^N} \tag{15.4}$$

The term $r(1 + r)^N/(1 + r)^N - 1$ is called the Capital Recovery Factor.

EXAMPLE 15.1

As a routine maintenance work, a sum of Rs 20,00,000 each year is to be spent on a particular stretch of a highway during the third year, fifth year, and the seventh year.

Calculate the total present worth of these expenditures, if the annual discount rate is 12% (compound).

Solution

The present worth of the maintenance investment is

$$20,00,000 \times \left[\frac{1}{1.12^3} + \frac{1}{1.12^5} + \frac{1}{1.12^7}\right] = \text{Rs } 34,63,080$$

EXAMPLE 15.2

The construction expenditure of a new highway was estimated to be Rs 200 crore. It was decided that this money be raised from loans. Calculate the instalment to be paid each year, such that the loan is repaid in 15 years? Assume the compound rate of interest as 10%.

Solution

Using Eq. (15.4),

$$200 = x \times \frac{(1.10)^{15} - 1}{0.1(1.10)^{15}}$$

Therefore,

$$x = \text{Rs } 26.294 \text{ crore}$$

15.4 COST COMPONENTS IN TRANSPORTATION SYSTEM

This section discusses various costs associated with the transportation system. An improvement in the transportation system may reduce the cost of vehicle operation on highways, cost of traffic congestion, travel time, and so on, but these savings are achieved as a return benefit of construction of new highways, maintenance of roads, construction of new flyovers, and so on, which again involves huge investment costs. Thus, the cost of transportation can be thought of as consisting of costs incurred by two parties, namely the agency cost and the user cost.

15.4.1 Agency Cost

The *agency cost* is the cost incurred by the agency (government or private) for construction and maintenance of a highway facility.

The construction cost involves the following elements:

(a) Surveying, planning, and design
(b) Acquisition of land
(c) Construction of the highway
(d) Supervision, quality control, and administration
(e) Installation of traffic control devices and other facilities.

The maintenance cost involves the cost of planning and implementation of various maintenance measures in respect of in-service roads during various time phases. The maintenance costs comprise those due to (i) periodic repair, (ii) major rehabilitation, (iii) operational requirements, and (iv) supervision and installation.

15.4.2 User Cost

The various components of cost incurred by the road user are discussed here. Unlike the case of determination of cost of construction, the costs incurred by the users are sometimes difficult to evaluate. In some cases, the user cost is not measurable, such as the pain and grief caused by accidents.

Vehicle operating cost (VOC)

The cost of owning and operating a vehicle on the road is called the Vehicle Operating Cost (VOC). It has two components, namely the variable cost and the fixed cost.

(a) Variable costs may be subdivided into two categories:
   - *Distance related components*, such as (i) fuel consumption, (ii) spare parts, (iii) tyre wear, (iv) lubricants, (v) maintenance labour cost, and so on.
   - *Time related components*, such as (i) depreciation, (ii) value of the passengers' time, (iii) wages of the crew, and so on.

(b) Fixed costs include the capital cost, registration fees, insurance, road permit charges, road and other taxes.

Traffic congestion and restraint

Any renovation on existing pavements causes disruption of traffic in terms of congestion, reduction of speed, and even complete restraint on movements. This delay in travel time is a function of road geometrics, traffic volume, time, and duration of the road renovation work. The effect of congestion is taken into account by applying suitable correction factors to VOC.

- A correction factor, called the *congestion factor* [249], which is a function of the type of road and the volume capacity ratio, is used as a multiplier of the distance related cost component of VOC.

- The ratio of the free speed and speed during congestion is used to multiply the time related cost component of VOC.

The congestion occurs during the peak hours only. Therefore, information about the hourly traffic volume is needed to calculate the congestion factor accurately.

Accident

Roads which have poor functional characteristics, are prone to have more number of accidents. The cost of accident is difficult to measure. An accident may involve loss of human lives, injury to human lives, cost of hospitalization, and damage to the property and vehicles.

Depending on the severity and degree of damage caused by the accident, some values are recommended in the road user cost data [249].

Cost of travel time

The travel time is reduced by better quality of roads, flyovers, and exclusive expressways. The cost of travel time is different for different users. A rich person may be willing to pay higher fare for a fast and comfortable mode of transportation. Thus, a person's earnings can be used as an index to determine the value of time for that individual. However, this approach can only give the time value during the work trip. The time value of non-working hours can be calculated by studying the preferences of passengers for choosing a particular mode of transport [172]. Table 15.1 gives some values of the travel time evaluated based on road user cost data [249]. As is obvious, the cost of travel time for work trip is higher than that of non-work trip.

Table 15.1 Value of travel time of passengers as per 1990 road user cost data [212]

Type of passenger	*Work trip* (Rs/h)	*Non-work trip* (Rs/h)
Bus passengers on trunk routes	27	3
Bus passengers in secondary routes	10	2
Car/taxi/two-wheeler passengers	30	4

15.5 BENEFIT COMPONENT IN TRANSPORTATION SYSTEMS

Various cost components of the transportation system have been discussed in Section 15.4. The objective of a good transportation system is to provide an efficient, quick, and safe means of transportation to its users. The various possible forms of benefits that accrue to users can be summarized as follows [212]:

1. *Road user benefits* such as:
   (i) Savings in VOC
   (ii) Reduction in travel times
   (iii) Reduction in accidents
   (iv) Savings in maintenance costs.
2. *Social benefits* comprise benefits due to improvement in administration, health, education, agriculture, industry, trade, environmental standards, and so on.

It is difficult to estimate the benefit component of a transportation system, and sometimes it is just not possible to quantify benefits for some components. For example, a good transportation system may reduce the noise and air pollution levels, or may enhance the aesthetics, which are all difficult to quantify. However, the direct benefits, for example, the reduction in the VOC due to a proposed new transportation facility can be quantified.

The benefits of a highway project are received by three categories of traffic, namely the normal traffic, the diverted traffic, and the generated or induced traffic [212]. The normal traffic is the traffic which was originally plying on the existing facility. After the construction of the new highway facility, some reduction in VOC is expected, which is enjoyed by the existing traffic. Similarly, some traffic is diverted from their earlier route to the new facility and the new facility may also generate a new traffic stream.

15.6 ECONOMIC EVALUATION OF HIGHWAY PROJECTS

As mentioned earlier, the cost as well as the benefit are spread over the time horizon. It is difficult to compare the cost and benefit components, unless they are brought to equivalent values at a particular base year. Various methods are available for economic evaluation of highway projects which enable an analyst to compare relative benefits which can be derived from various alternative projects. The project which is most beneficial in all respects, is finally recommended for funding and execution. This section briefly discusses various methods of economic evaluation of highway projects.

15.6.1 Cost-Benefit Ratio Method

In cost-benefit analysis, the costs and the benefits of individual highway projects are calculated, bringing all the expenditures to the base year for comparison purposes. Various alternative highway projects are thus compared before taking the decision.

15.6.2 Net Present Value Method

In the net present value method, the cost and the benefits of the individual years are discounted to the present value. Thus, the Net Present Value, NPV, of the base year can

be written as

$$\text{NPV} = \sum_{i=0}^{n} \left(\frac{B_i - C_i}{(1+r)^n} \right) \tag{15.5}$$

where

B_i is the benefit of the ith year

C_i is the cost of the ith year

n is the number of years.

15.6.3 Internal Rate of Return Method (IRR)

Internal rate of return is that discount rate, for which the NPV value is zero. This can be obtained by setting the value of NPV in Eq. (15.5) to zero, and solving (by trial and error) for the value of r. If the rate of return thus obtained is more than the market interest, then the project is adjudged to be acceptable.

15.6.4 Comparison of Various Methods

The cost-benefit model is simple to use, but sometimes when the cost-benefit ratios of two alternatives are close to each other, it becomes difficult to interpret, and choose the best option. Some components, whether to be treated as benefit or cost (i.e. whether they will go to the numerator or denominator), sometimes appear confusing, as savings in cost is actually a benefit. In the NPV method, some discount rate is assumed, and various alternative projects are compared based on the NPVs. If a different discount rate is assumed, the order of choice among the alternatives may change. However, the IRR method takes care of this difficulty, because it finds out the rate of discount by itself. Thus, the IRR method is the most preferred economic analysis tool.

15.7 TRANSPORTATION FINANCING

Development of a road network, which is one of the main concerns of the developing countries, often suffers from fund constraints. An approximate estimation of the expenditure in the road sector in India, has been carried out for the coming ten years, which amounts to Rs 25,000 crore for expressways, Rs 120,000 crore for national highways, and Rs 70,000 crore for state highways [94]. The funding sources of this huge investment can be tapped from the government sector, international bank loans, private sectors, and so on.

For example, the recent Golden Quadrilateral and North-South and East-West corridors will involve an approximate expenditure of Rs 55,000 crore (2001). The

fundings of these projects have tentatively been arranged from various organizations, such as approximately Rs 20,000 crore from cess (from petrol and diesel), Rs 20,000 crore from the World Bank and the Asian Development Bank loans, Rs 10,000 crore from market borrowings, and Rs 4000 crore from the private sector [190, 137]. On an average, the National Highway Authority of India expects to get loans of Rs 1000 crore from Asian Development Bank and Rs 2000 crore from the World Bank every year for the National Highways Development Project [190]. Similarly, for the PMGSY project, it has been decided that the finance will come from 50% of diesel cess, market borrowings, and external funding agencies.

The government funding may come from various sources, such as budgetary allocation, special road development bonds, fund out of cess on diesel or petrol (Central Road Fund), and so on. Road tolls provide an ongoing revenue source which is not tied to the annual government budgetary process. Toll revenue can be utilized for raising the debt finance or supporting the maintenance activities on that particular road stretch, and needs to be operated by government or a private agency. The difficulty with the toll financing for debt realization is that the 'plough back' time-period is long, and fluctuative. However, if at least the maintenance requirement of a particular stretch can be completely financed from toll, it can stay away from the competition with the maintenance requirements of the other roads in the network. *Shadow tolling* is another concept, where government pays the investor for each vehicle that enters the stretch.

Funding for transportation projects can also be attracted from private organizations. Some models of degree of involvement of private organizations in a transportation project financing include:

(i) Completely owned, financed and operated by a private body
(ii) Build, operate, and transfer (BOT) approach
(iii) Build, transfer, and operate (BTO) approach
(iv) Finance, build, and lease approach

Several variations of the BOT approach exist, such as, build, own, and operate (BOO); build, own, operate, and sell (BOOS); build, own, operate, and transfer (BOOT); and so on [137]. The government offers certain incentives towards investment in building transportation facilities in order to encourage private sector participation. For example, government has declared road sector as an industry to facilitate commercial borrowing and to permit floating of bonds in the market; investors are sometimes allowed to develop real estate as a part of the highway project; tax is exempted up to a certain level; concession period is available; and direct foreign investment is sometimes permitted [137].

Various ways of financing transportation projects have been discussed above. Transportaion projects require large outlays of investment, therefore, all the funding possiblities should be explored for successful implementaion of a conceived project.

EXERCISES

1. Why is the study of road economics and finance important?
2. Differentiate between the interest rate and the discount rate.
3. Define salvage value of a highway project. How can it be determined?
4. Explain the concept of present worth used in economic analysis of highway projects.
5. Derive the expression for capital recovery factor taking an example of pavement maintenance scheme. How do you take into account the inflation rate in the expression for capital recovery factors?
6. What are the various cost components of highway projects? Discuss.
7. How can the cost of travel time be measured?
8. Explain the various methods of economic evaluation of highway projects. Also discuss their relative merits and demerits.
9. What are possible sources of financing of a transportation project?
10. What are the constraints of private funding in highway projects?
11. What is shadow tolling?

Bibliography

[1] *AASHO Road Test*, Highway Research Board, Report No. 9, 61-E, Washington D.C., 1962.

[2] *AASHTO Guide for Design of Pavement Structures*, American Association of State Highway Officials, Washington, D.C., 1993.

[3] *A Policy on Geometric Design of Highways and Streets*, American Association of State Highway and Transportation Officials, Washington, D.C., 1984.

[4] Aberg, B., "Void Sizes in Granular Soils," *Journal of Geotechnical Engineering*, ASCE, Vol. 122, No. 3, March, 1996, pp. 236–239.

[5] Agg, H.J., *Direction Sign Overload*, Project Report 77, Transport Research Laboratory, Crowthorne, 1994.

[6] Ashford, N., Stanton, H.P.M., and Moore, C.A., *Airport Operations*, 2nd ed., McGraw-Hill, New York, 1997.

[7] *Asphalt Technology News*, National Center for Asphalt Technology, Auburn University, Vol. 2, No. 2, Fall 2000.

[8] Baaj, M.H., and Mahmassani, H., "An AI-Based Approach for Transit Route System Planning and Design," *Journal of Advanced Transportation*, Vol. 25, 1991, pp. 187–210.

[9] Baaj, M.H., and Mahmassani, H., "TRUST: A LISP Program for the Analysis of Transit Route Configurations," *Transportation Research Record 1283*, 1992, pp. 125–135.

[10] Banister, D., *Transport Planning: In the UK, USA, and Europe*, Spon, London, 1994.

[11] Bell, M.G.H., Bonsall, P.W., and O'Flaherty, C.A. "Driver Information Systems," Chapter 28 in O'Flaherty, C.A. (Ed.) *Transport Planning and Traffic Engineering*, Arnold, London, 1997.

[12] Bhatia, H.S., "India's Road Development—Challenges and Options," *Seminar on Perspective Planning for Road Development in India*, 7-8th October, 1996, The Indian Roads Congress, New Delhi.

[13] Bhattacharya, P.G., and Pandey, B.B., "Flexural Fatigue Strength of Lime-Laterite Soil," *Transportation Research Record*, No. 1986, 1089, National Research Council, TRB, Washington, D.C., pp. 86–92.

[14] Bhattacharya, P.G., and Pandey, B.B., "Fatigue Test Set-up for Lime-Soil Mixtures," *Journal of Civil Engineering*, Institution of Engineers (India), Vol. 68, Part C13, Nov. 1987, pp. 130–135.

[15] Blight, G.E., "Permanent Deformation in Asphaltic Materials," *Transportation Engineering Journal*, ASCE, Vol. 100, TE2, 1974, pp. 263–276.

[16] Bookbinder, J.H., and Desilets, A., "Transfer Optimization in a Transit Network," *Transportation Science*, Vol. 26, 1992, pp. 106–118.

[17] Boussinesq, V.J., *Application des potentiels a letude de lequilibre at du mouvement des solids elastiques avec les notes etendues sur divers points de physique*, Mathematique et danalyse, Gauthier-Villais, Paris, 1885.

[18] Boyce, I.R., Brown, S.F., and Pell, P.S., "The Resilient Behaviour of a Granular Material under Repeated Loading," *Australian Road Research Board Proceedings*, Vol. 8, 1976, pp. 1–12.

[19] Brackstone, M., and McDonald, M., "Car-following: A Historical Review," *Transportation Research,* Part F, Vol. 2, 1999, pp. 181–196.

[20] Bradbury, R.D., *Reinforced Concrete Pavements*, Wire Reinforcement Institute, Washington, D.C., 1938.

[21] Brown, S.F., "Achievements and Challenges in Asphalt Pavement Engineering," Keynote Address, *8th International Conference on Asphalt Pavements*, Seattle, 1997, pp. 1–23.

[22] Brown, S.F., and Pell, P.S., "A Fundamental Structural Design Procedure for Flexible Pavements," *Proceeding of the 3rd International Conference of Structural Design of Asphalt Pavements*, Vol. I, 1972, pp. 369–381.

[23] Brown, S.F., "Simplified Fundamental Design Procedure for Bituminous Pavements," *The Highway Engineer*, Vol. XXI, 209 (8–9), 1974, pp. 14–23.

[24] Brunton, J.M., Brown, S.F., and Pell, P.S., "Development to the Nottingham Analytical Design Method for Asphalt Pavements," *Proceeding of the 6th International Conference on Structural Design of Asphalt Pavements*, Vol. I, 1987, pp. 366–377.

[25] Burmister, D.M., "The General Theory of Stresses and Displacements in Layered System," *Journal of Applied Physics*, Vol. 16, 1945, pp. 126–127, 296–302.

[26] Burmister, D.M., "The Theory of Stresses and Displacements in Layered Systems and Applications to the Design of Airport Runways," *Proceedings of Highway Research Board*, HRB, Vol. No. 23, 1943, Washington D.C., pp.126–144.

[27] Byrne, B.F., and Vuchic, V., "Public Transportation Line Positions and Headways for Minimum Cost," *Proceedings of the Fifth International Symposium on Traffic Flow and Transportation*, 1972, pp. 347–360.

[28] Byrne, B.F., "Public Transportation Line Positions and Headways for Minimum User and System Cost in a Radial Case," *Transportation Research*, Vol. 9, 1975, pp. 97–102.

[29] Ceder, A., and Wilson, N.H.M., "Bus Network Design," *Transportation Research,* Part B, Vol. 20, 1986, pp. 331–344.

[30] Cedergren, H.R., *Drainage of Highway and Airfield Pavements*, John Wiley & Sons, 1974.

[31] Chakroborty, P., Deb, K., and Subrahmanyam, P.S., "Optimal Scheduling of Urban Transit System using Genetic Algorithms," *ASCE Journal of Transportation Engineering*, Vol. 121, 1995, pp. 544–553.

[32] Chakroborty, P., and Dwivedi, T., "Optimal Route Network Design for Transit Systems using Genetic Algorithms," *Engineering Optimization*, March 2002.

[33] Chakroborty, P., and Kikuchi, S., "Evaluation of the General Motors based Car-following Models and a Proposed Fuzzy Inference Model," *Transportation Research,* Part C, Vol. 7, 1999, pp. 209–235.

[34] Chakroborty, P., and Kikuchi, S., "A Method to Calibrate the Fuzzy Rule Based Inference System," *Transportation Research,* Part C, January 2003 issue.

[35] Chakroborty, P., Kikuchi, S., and Luszcz, M., "Lengths of Left-Turn Lanes at Unsignalized Intersections," *Transportation Research Record 1500*, 1995, pp. 193–201.

[36] Chandler, R.E., Herman, R., and Montroll, E.W., "Traffic Dynamics: Studies in Car-following," *Operations Research,* Vol. 6, 1958, pp. 165–184.

[37] Chou, Y.J., and Lytton, R.L., "Accuracy and Consistency of Backcalculated Pavement Layer Moduli," *Transportation Research Record*, No. 1293, TRB, National Research Council, Washington, D.C., 1991, pp. 72–85.

[38] Claessen, A.I.M., Edwards, J.M., Sommer, P., and Uge, P., "Asphalt Pavement Design: The Shell Method," *Proceedings of 4th International Conference of Structural Design of Asphalt Pavements*, Vol. 1, 1977, pp. 35–70.

[39] *Code of Practice for Maintenance of Bituminous Surfaces of Highways*, IRC:82–1982, The Indian Roads Congress, New Delhi, 1982.

[40] *Code of Practice for Road Signs*, IRC: 67–1977, The Indian Roads Congress, New Delhi, 1977.

[41] Coduto, D.P., *Geotechnical Engineering, Principles and Practices*, Prentice Hall, New Jersey, USA, 1998.

[42] Computer Program DAMA, *Pavement Structural Analysis Using Multi-Layered Elastic Theory*, User's manual, Asphalt Institute, October, 1983.

[43] Cooper, K.E., Brown, S.F., and Pooley, G.R., "The Design of Aggregate Gradings for Asphalt Base Courses," *Symposium on Asphalt Mix Design*, at the Association of Asphalt Paving Technologists, San Antonio, Texas, 11–13th February, 1985.

[44] Croney, D., and Croney, P., *The Design and Performance of Road Pavements*, 2nd ed., McGraw-Hill International Series in Civil Engineering, 1992.

[45] Das, A., and Pandey, B.B., "Economical Design of Bituminous Pavements with Two Grades of Bitumen in the Surfacing," *Proceeding of Seminar on Road Financing, Design, Construction and Operation of Highways in 21st Century*, The Indian Roads Congress, New Delhi, 2000, pp. II-35-II-42.

[46] Das, A., and Pandey, B.B., "Analytical Design of Flexible Pavement for Planned Stage Construction," *Indian Highways,* Indian Roads Congress, New Delhi, Vol. 27, No. 4, 1999, pp. 5–11.

[47] Das, A., Reddy, K.S., and Pandey, B.B., "Analytical Overlay Design and Validation: An Indian Case Study," *Proceedings of the 5th International Conference on the Bearing Capacity of Roads and Airfields*, Trondheium, Norway, Vol-III, 1998, pp. 1787–1796.

[48] Davis, R.O., and Selvadurai, A.P.S., *Elasticity and Geomechanics*, 1st ed., Cambridge University Press, 1996.

[49] Deb, K., and Chakroborty, P., "Time Scheduling of Transit Systems with Transfer Considerations using Genetic Algorithms," *Journal of Evolutionary Computation*, Vol. 6, 1998, pp. 1–24.

[50] Dorman, G.M., "The extension to practice of a fundamental procedure for design of flexible pavements," *Proceedings of 1st International Conference of Structural Design of Asphalt Pavements*, Ann. Arbor, Michigan, 1962, pp. 785–793.

[51] *Draft Manual, Design and Specifications of Rural Roads, Pradhan Mantri Gram Sadak Yojana (PMGSY)*, Ministry of Rural Development, Govt. of India, Design and Specification Committee, 2001.

[52] *Draft TRH4: 1996, Structural Design of Flexible Pavements for Interurban and Rural Roads*, 1996, Draft TRH4, TRANSPORTEK, CSIR, Pretoria, South Africa, 1996.

[53] Drew, D.R., *Traffic Flow Theory and Control*, McGraw-Hill, New York, 1968.

[54] *Dimensions and Weights of Road Design Vehicles*, IRC:3–1983, The Indian Roads Congress, New Delhi, 1983.

[55] Dubois, D., Bel G., and Libre, M., "A Set of Methods in Transportation Network Synthesis and Analysis," *Journal of Operational Research Society*, Vol. 30, 1979, pp. 797–808.

[56] Duncan, J.M., Monismith, C.L., and Wilson, E.L., "Finite Element Analysis of Pavements," *Highway Research Record*, Vol. 228, Highway Reseach Board, 1968, Washington D.C., pp. 18–33.

[57] Easa, S.M., and Can, E.K., "Stochastic Priority Model for Aggregate Blending," *Journal of Construction Engineering and Management*, ASCE, Vol. 111, No. 4, 1985, pp. 358–373.

[58] Easa, S.M., and Can, E.K., "Optimization Model for Aggregate Blending," *Journal of Construction Engineering and Management*, ASCE, Vol. 111, No. 3, 1985, pp. 216–230.

[59] Edie, L.C., "Following and Steady-State Theory for Non-congested Traffic," *Operations Research*, Vol. 9, 1961, pp. 66–76.

[60] Final Report, Research Scheme R-56, *Development of Computer Program and Design Charts for Analytical Design of Flexible Pavement*, submitted to the MOST, New Delhi, by Civil Engg. Dept., IIT Kharagpur, March, 1999.

[61] Final Report, *Development of Methods such as Benkelman Beam Deflection Method for Evaluation of Structural Capacity of Existing Flexible Pavements and also for Estimation and Design of Overlays for Strengthening of any Weak Pavement*, Research Scheme R-6, submitted to Ministry of Surface Transport, New Delhi, by Central Road Research Institute, New Delhi, 1995.

[62] Final Report, Research Scheme R-24, *Determination of Elastic Modulus and Fatigue Characteristics of Asphaltic Mixes in Laboratory*, Transportation Engg. Section, Dept. of Civil Engg., IIT, Kharagpur, 1989.

[63] Finn, F., Saraf, C., Kulkarni, R., Nair, K., Smith, W., and Abdullah, A., "The Use of Distress Prediction Subsystems for the Design of Pavement Structures," *4th International Conference of Structural Design of Asphalt Pavements*, 1977, Vol. I, pp. 3–38.

[64] Fuller, W.B., and Thompson, S.E., "The Laws of Proportioning Concrete," *Transactions of ASCE*, ASCE, Vol. 59, 1907, pp. 67–143.

[65] Garber, N.J., and Hoel, L.A., *Traffic and Highway Engineering*, West Publishing Company, St. Paul, 1988.

[66] Gazis, D.C., Herman, R., and Potts, R., "Car-following Theory of Steady State Traffic Flow," *Operations Research*, Vol. 7, 1959, pp. 499–595.

[67] Gazis, D.C., Herman, R., and Rothery, R.W., "Non-linear Follow-the-leader Models of Traffic Flow," *Operations Research*, Vol. 9, 1961, pp. 545–567.

[68] *General Guidelines about the Equipment for Bituminous Surface Dressing*, IRC:SP 34, The Indian Roads Congress, New Delhi, 1989.

[69] Ghosh, P., "Study on Mix Design on Indian Specification," M.Tech. thesis, Dept. of Civil Engg., IIT Kanpur, May, 2002.

[70] *Guidelines on Road Drainage*, IRC SP42–1994, The Indian Roads Congress, New Delhi, 1994.

[71] *Guidelines on Urban Drainage*, IRC SP50–1999, The Indian Roads Congress, New Delhi, 1999.

[72] *Guidelines for the Design of Small Bridges and Culverts*, IRC:SP 13–1973, The Indian Roads Congress, New Delhi, 1973.

[73] *Guidelines for Strengthening of Flexible Road Pavements Using Benkelman Beam Deflection Technique*, IRC:81–1981. The Indian Roads Congress, New Delhi, 1981.

[74] *Guidelines for Strengthening of Flexible Road Pavements Using Benkelman Beam Deflection Technique*, 1st revision, IRC:81–1997, The Indian Roads Congress, New Delhi, 1997.

[75] *Guidelines for Use of Flyash in Road Embankments*, IRC:SP 58–2001, The Indian Roads Congress, New Delhi, 2001.

[76] *Guidelines for the Use of Soil-Lime Mixes in Road Construction*, IRC:51–1992, 1st Revision, Indian Roads Congress, New Delhi, 1992.

[77] *Guidelines for the Use of Dry Lean Concrete as Sub-base for Rigid Pavement*, IRC:SP-49, The Indian Roads Congress, New Delhi, 1998.

[78] *Guidelines for Internal Drainage of Pavements,* unpublised, The Indian Roads Congress, New Delhi, 2002.

[79] *Geometric Design Standards for Rural (Non-Urban) Highways*, IRC:73–1980, The Indian Roads Congress, New Delhi, 1980.

[80] *Geometric Design Standards for Urban Roads in Plains*, IRC:86–1983, The Indian Roads Congress, New Delhi, 1983.

[81] Greenberg, H., "An Analysis of Traffic Flow," *Operations Research*, Vol. 7, 1959, pp. 78–85.

[82] Greenshields, B.D., "A Study in Highway Capacity." *Proceedings of Highway Research Board*, Vol. 14, 1935, p. 458.

[83] *Guidelines for Capacity of Roads in Rural Areas*, IRC:64–1990, The Indian Roads Congress, New Delhi, 1990.

[84] *Guidelines for Capacity of Urban Roads in Plain Areas*, IRC:106–1990, The Indian Roads Congress, New Delhi, 1990.

[85] *Guidelines for Design of Horizontal Curves for Highways and Design Tables*, IRC:38–1988. The Indian Roads Congress, New Delhi, 1988.

[86] *Guidelines for the Design of At-grade Intersections in Rural and Urban Areas*, IRC Special Publication 41, The Indian Roads Congress, New Delhi, 1994.

[87] *Guidelines for the Design of Flexible Pavements*, IRC:37–1970, The Indian Roads Congress, New Delhi, 1970.

[88] *Guidelines for the Design of Flexible Pavements*, IRC:37–1984, The Indian Roads Congress, New Delhi, 1984.

[89] *Guidelines for the Design of Flexible Pavements*, IRC:37–2001, The Indian Roads Congress, New Delhi, 2001.

[90] *Guidelines for the Design of Interchanges in Urban Areas*, IRC:92–1985, The Indian Roads Congress, New Delhi, 1985.

[91] *Guidelines for the Design of Rigid Pavements for Highways*, IRC:58–1988, The Indian Roads Congress, New Delhi, Reprint 1991.

[92] *Guidelines for the Design of Rigid Pavements for Highways*, unpublished IRC: 58–2000, The Indian Roads Congress, New Delhi, 2000.

[93] *Guidelines on Design and Installation of Road Traffic Signals*, IRC:93–1985, The Indian Roads Congress, New Delhi, 1985.

[94] Gupta, D.P., "Road Development in India: Strategies for Implementation,'' *Proceeding of Seminar on Road Financing, Design, Construction and Operation of Highways in 21st Century*, The Indian Roads Congress, New Delhi, 2000, pp. II-27-II-34.

[95] Gujarati, D.N., *Basic Econometrics*, 2nd. ed., McGraw-Hill, New York, 1988.

[96] Hall, K.T., and Mohseni, A., "Backcalculation of Asphalt Concrete—Overlaid Portland Cement Concrete Pavement Layer Moduli,'' *Transportation Research Record*, No. 1293, TRB, National Research Council, Washington, D.C., 1991, pp. 112–123.

[97] Hanson, F.M., "Bituminous Surface Treatment of Rural Highway,'' *Proceedings of New Zealand Society of Civil Engineers*, Vol. 21, 1935, pp. 89–179.

[98] Harmelink, M.D., "Volume Warrants for Left-turn Storage Lanes at Unsignalized Grade Intersections.'' *Highway Research Record 211*, 1967, pp. 1–18.

[99] Hausmann, M.R., *Engineering Principles of Ground Modification*, McGraw-Hill International Editions, Civil Engineering Series, 1990.

[100] Herman, R., Montroll, E.W., and Potts, R.B., "Car-following Theory of Steady State Flow,'' *Operations Research*, Vol. 7, 1959, pp. 86–106.

[101] Herman, R., and Potts, R.B., "Single Lane Traffic Theory and Experiment," *Proceedings of the Symposium on the Theory of Traffic Flow*, 1959, pp. 120–146.

[102] Heukelom, W., "An Improved Method of Characterizing Asphaltic Bitumens with the Aid of Their Mechanical Properties," *Proceedings of the Annual Meeting of the Association of Asphalt Paving Technologists*, Houstan, 1973, pp.1–35.

[103] *Highway Capacity Manual*, Special Report 209, Transportation Research Board, National Research Council, Washington, D.C., 1998.

[104] *Highway Capacity Manual*, Special Report 209, Transportation Research Board, National Research Council, Washington, D.C., 1985.

[105] Holroyd, E.M., "The Optimal Bus Service: A Theoretical Model for a Large Uniform Urban Area," *Proceedings of the Third International Symposium on the Theory of Traffic Flow*, 1967, pp. 308–328.

[106] Horonjeff, R., and Mckelvey, F.X., Planning and Design of Airports, McGraw-Hill Book Company, New York, 1983.

[107] http://www.lsu.umich.edu/hales/countdown, last accessed October, 2001.

[108] http://www.ce.wsu.edu/CEabout/CEfaculty/images/website.ppt, last visited 27th January, 2002.

[109] *Indian Standard Methods for Test for Soils,* Part-XVI, *Laboratory Determination of CBR*, IS:2720(XVI), 3rd reprint, The Indian Standards Institution, New Delhi, December, 1974.

[110] *Indian Standard Method of Load Test on Soils*, IS:1888–1982, 2nd revision, The Bureau of Indian Standards, New Delhi, 1982.

[111] *Indian Standard Methods of Test For Aggregates for Concrete,* Part-IV, *Mechanical Properties*, IS:2386(IV), The Bureau of Indian Standards, New Delhi, 10th Reprint, 1997.

[112] *Indian Standard Methods of Test For Aggregates for Concrete,* Part-V, *Soundness*, IS:2386(V)–1963, The Bureau of Indian Standards, New Delhi, 1997.

[113] *Indian Standard Methods of Test For Aggregates for Concrete,* Part-I, *Particle Size and Shape*, IS:2386(I), The Bureau of Indian Standards, New Delhi, 12th reprint, 1999.

[114] *Indian Standard Methods of Test For Aggregates for Concrete,* Part-III, *Specific Gravity, Density, Voids, Absorption and Bulking*, IS:2386(III), The Bureau of Indian Standards, New Delhi, 8th reprint, 1997.

[115] *Indian Standard Method of Test for Determination of Stripping Value of Road Aggregates*, IS:6341, The Bureau of Indian Standards, New Delhi, 4th reprint, 1991.

[116] *Indian Standard Specification for Cutback Bitumen*, IS:217–1988, 2nd revision, The Bureau of Indian Standards, New Delhi, 1989.

[117] *Indian Standard Bitumen Emulsion for Roads (Cationic Type)—Specification*, IS:8887–1995, The Bureau of Indian Standards, New Delhi, 1997.

[118] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Viscosity,* Part I, *Industrial Viscosity*, IS:1206–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[119] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Penetration*, IS:1203–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[120] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Softening Point*, IS:1205–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[121] *Indian Standard Methods for Testing Tar and Bituminous Materials: Float Test*, IS:1210–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[122] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Ductility*, IS:1208–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[123] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Specific Gravity*, IS:1202–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[124] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Loss on Heating*, IS:1212–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[125] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Matter Insoluble in Benzene*, IS:1214–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[126] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Water Content (Dean and Stark Method)*, IS:1211–1978, 1st revision, The Bureau of Indian Standards, New Delhi, 1978.

[127] *Indian Standard Methods for Testing Tar and Bituminous Materials: Determination of Flash and Fire Point*, IS:1209–1978, 1st Revision, The Bureau of Indian Standards, New Delhi, 1978.

[128] *Introduction to Asphalt*, MS-5, 8th ed., Asphalt Institute, reprint, 2001.

[129] Irya, I.R., Jain, P., and Pundhir, N.K.S., "Specification, Design and Construction Methodology for Use of Semi-Dense Mixes Repair of Bituminous Roads Using Cationic Bitumen Emulsions," *Seminar on Bituminous Roads: Design and Construction Aspects*, New Delhi, 25-26th August, 1994.

[130] "ITE Professional Policies," *1986 ITE Membership Directory*, Institute of Transportation Engineers, Washington, D.C., 1986.

[131] Jumikis, A.R., *Theoretical Soil Mechanics*, 1st ed., Van Nostrand Reinhold, Canada, 1969.

[132] Kanafani, A., *Transportation Demand Analysis*, McGraw-Hill, New York, 1983.

[133] Kasahara, A., Himeno, K., Kawamura, K., and Nakagawa, S., "Performance of Asphalt Pavements at BIBI New Test Road in Japan Related to Their Bearing Capacity," *Proceeding of the 7th International Conference of Structural Design of Asphalt Pavements*, 1992, Vol. I, pp. 106–123.

[134] Kikuchi, S., and Chakroborty, P., "Analysis of Left-turn-lane Warrants at Unsignalized T-Intersections on Two-Lane Roadways," *Transportation Research Record 1327*, 1991, pp. 80–88.

[135] Kikuchi, S., and Chakroborty, P., "Car-Following Model based on Fuzzy Inference System," *Transportation Research Record 1365*, 1992, pp. 82–91.

[136] Kikuchi, S., Chakroborty, P., and Vukadinovic, K., "Lengths of Left-Turn Lanes at Signalized Intersections," *Transportation Research Record 1385*, 1993, pp. 162–171.

[137] Krishnan, R.R., "Need, Problems and Success on Certain Road Privatization," *Proceeding of Seminar on Road Financing, Design, Construction and Operation of Highways in 21st Century*, The Indian Roads Congress, New Delhi, 2000, pp. I-181-I-202.

[138] Kumar, M.J.R., and Pandey, B.B., "Structural Properties of Cement Treated Laterite Soils," *Highway Research Board Bulletin*, Vol. 47, Indian Roads Congress, New Delhi, 1992, pp. 33–42.

[139] Kumar, P., "An Update on Development of National Highways," *Souvenir of Seminar on Road Financing, Design, Construction and Operation of Highways in 21st Century*, The Indian Roads Congress, New Delhi, 2000.

[140] Kumar, V., "Permeability of Bituminous Mixes," unpublished B.Tech. thesis, Dept. of Civil Engg., IIT Kanpur, April, 2002.

[141] Lambe, T.W., and Whitman, R.V., *Soil Mechanics*, John Wiley & Sons, 2000.

[142] Lampkin, W., and Saalmans, P.D., "The Design of Routes, Service Frequencies, and Schedules for a Municipal Bus Undertaking: A Case Study. *Operation Research Quarterly*, Vol. 18, 1967, pp. 375–397.

[143] Lay, M.G., *Handbook of Road Technology*, 3rd ed., Vol. 1, Gordon and Breach Science Publications, 1998.

[144] *Left-Turn Treatments at Intersections: A Synthesis of Highway Practice*, NCHRP Synthesis 225, Transportation Research Board, National Research Council, Washington, D.C., 1996.

[145] Lenngren, C.A., "Relating Deflection Data to Pavement Strain," *Transportation Research Record*, No. 1293, TRB, National Research Council, Washington, D.C., 1991, pp. 103–111.

[146] Lister, N.W., and Powell, W.D., "Design Practice for Bituminous Pavements in the United Kingdom," *Proceeding of the 6th International Conference on Structural Design of Asphalt Pavements,* Vol. I, 1987, pp. 220–231.

[147] Love, A.E.H., *A Treatise on the Mathematical Theory of Elasticity*, 4th ed., Dover, New York, 1944.

[148] Mandal, B., "Curing and Flexural Fatigue Strength of Sea Beach Sand Stabilised with Lime and Fly-ash," M. Tech. Thesis, Dept. of Civil Engg., IIT Kharagpur, 1986.

[149] Mandl, C.E., "Evaluation and Optimization of Urban Public Transportation Networks," *Third European Congress on Operations Research*, 1979.

[150] Mandl, C.E., "Evaluation and Optimization of Urban Public Transportation Networks," *European Journal of Operational Research*, Vol. 5, 1980, pp. 396–404.

[151] *Manual for Asphalt Pavement*, 1989, Japan Road Association, Japan.

[152] *Manual on Uniform Traffic Control Devices for Streets and Highways*, 4th Revision, U.S. Department of Transportation, Federal Highway Administration, 1978.

[153] May, A.D., and Keller, H.E.M., "Non-integer Car-Following Models," *Highway Research Record 199*, 1967, pp. 19–32.

[154] May, A.D., *Traffic Flow Fundamentals*, Prentice-Hall, Inc., New Jersey, 1990.

[155] May, R., and Witczak, M.W., "Effective Granular Modulus to Model Pavement Responses," *Transportation Research Record*, No. 1022, TRB, National Research Council, Washington D.C., 1985, pp. 1–9.

[156] Mazumdar, B.K., "Lean Cement Concrete in Pavement Construction," M.Tech. Thesis, Dept. of Civil Engg., IIT Kharagpur, 1992.

[157] McShane, W.R., and Roess, R.P., *Traffic Engineering*, Prentice-Hall, Inc., New Jersey, 1990.

[158] Mertens, E.W., Wright, J.R., "Cationic Asphalt Emulsions, How they differ from Conventional Emulsions in Theory and Practice," *Highway Research Record*, Highway Research Board, 1959, pp. 386–397.

[159] Michaels, R.M., and Gozan, L.W., "Perceptual and Field Factors Causing Lateral Displacement," *Highway Research Record 25*, 1963.

[160] Miner, M.A., "Cumulative Damage in Fatigue," *Proceedings of ASME*, ASME, 1945, pp. A159–A164.

[161] Minty, E.J., Pratt, D.N., and Brett, A.J., 1980, "Aggregate Durability Tests Compared," *Proceeding of 10th ARRB Conference*, Vol. 10(3), pp. 10–20.

[162] Mitchell, J.K., and Monismith, C.L., "A Thickness Design Procedure for Pavements with Cement Stabilized Bases and Thin Asphalt Surfacings," *Proceeding of the 4th International Conference of Structural Design of Asphalt Pavements*, Vol. I, 1977, pp. 409–416.

[163] *Mix Design Methods for Asphalt*, The Asphalt Institute, 6th ed., Reprinted 1997.

[164] Monismith, C.L., Secor, K.E., and Blackmer, E.W., "Asphaltic Mixture Behaviour in Repeated Flexure," *Proceedings of AAPT*, AAPT, Vol. 30, 1961, pp. 188–222.

[165] Monismith, C.L., and Deacon, J.A., "Fatigue of Asphalt Paving Mixtures," *Journal of Transportation Engineering Division*, ASCE, TE2, 1969, pp. 317–345.

[166] Monismith, C.L., Secor, K.E., and Blackmer, W., "Asphalt mixture behaviour in repeated flexure," *Proceedings of Association of Asphalt Paving Technologists*, AAPT, Vol. 30, 1969, pp. 317–346.

[167] Monismith, C.L., "Pavement Evaluation and Overlay Design: Summary of Methods," Pavement Evaluation and Overlay Design: A Symposium and Related Papers, *Transportation Research Record*, No. 700, Transportation Research Board, National Research Council, Washington D.C., 1979, pp. 78–81.

[168] Monismith, C.L., Seed, H.B., Mitry, F.G., and Chan, C.K., "Prediction of Pavement Deflection from Laboratory Tests ," *Proceeding of 2nd International Conference of Structural Design of Asphalt Pavements*, 1967, pp. 109–140.

[169] Moser, C.A., and Kalton, G., *Survey Methods in Social Investigation*, 2nd ed., Heinemann Educational Books, London, 1979.

[170] Mundrey, J.S., *Railway Track Engineering*, 3rd. ed., Tata McGraw-Hill, New Delhi, 2000.

[171] Naraian, A.D., "Road Development—Challenges Ahead," *Indian Highways*, The Indian Roads Congress, Vol. 24, No. 6, 1996, pp. 11–14.

[172] Nash, C.A., "Economic and Environmental Appraisal of Transport Improvement Projects," Chapter 14, *Transport Planning and Traffic Engineering*, Edited by C.A. O'Flaherty, John Wiley & Sons, 1997.

[173] Netterberg, F., and Paige-Green, P., *Carbonation of Lime and Cement Stabilised Layers in Road Construction*, Technical Report. RS/3/84, TRANSPORTEK, CSIR, Pretoria, 1984.

[174] Neville, A.M., and Brooks, J.J., *Concrete Technology*, International Student Edition, Addison-Wesley, Edition, 1990, Reprint, 1997.

[175] Neville, A.M., *Properties of Concrete*, Longman, Thomson Press (India) Ltd., 1st Indian Reprint, 2000.

[176] Newcomb, D.E., and Epps, J.A., "Statistical Specifications for Hot Mix Asphalt — What do we need to know?," *Hot Mix Asphalt Technology*, March/April, 2001, Volume~6, Number~2, published by National Asphalt Pavement Association, USA.

[177] O'Flaherty, C.A., "Design of Off-street Parking Facilities." Chapter 22 in O'Flaherty, C.A., (Ed.) *Transport Planning and Traffic Engineering*, Arnold, London, 1997.

[178] Otte, E., Savage, P.F., and Monismith, C.L., "Structural Design of Cemented Pavement Layers," *Transportation Engineering Journal*, ASCE, Vol.1–108, TE4, 1982, pp. 428–446.

[179] Pandey, B.B., and Naidu, P.K., *Elastic Modulus of Materials in Flexible Pavement Design*, Highway Research Board Bulletin, Vol. 50, The Indian Roads Congress, New Delhi, 1994, pp. 21–41.

[180] Pandey, B.B., "Bituminous Mix Design," *Proceedings of a Two Day Workshop on Design of Flexible Pavement with Emphasis on the New IRC:37–2001 Guidelines*, 9–10 February, IIT Kanpur, 2002.

[181] Pandey, B.B., and Bhattacharya, P.G., "Strength and Curing of Lime-Laterite Soil—Plain and Fibre Reinforced," *Highway Research Board Bulletin*, Vol. 24, Indian Roads Congress, New Delhi 1983, pp. 1–26.

[182] *Pavement Design*, AUSTROADS, Sydney, 1992.

[183] Pell, P.S., "Pavement Materials: Keynote Address," *Proceeding of the 6th International Conference of Structural Design of Asphalt Pavements*, Vol-II, 1987, pp. 36–70.

[184] Personal Communication, D.P. Gupta, Director (Research), Asian Institute of Transportation Development, letter dated 21st February, 1998.

[185] Picket, G., and Ray, G.K., "Influence Chart for Concrete Pavements," *Transaction of ASCE*, 1951, pp. 49–73.

[186] *Plain and Reinforced Concrete—Code of Practice*, IS 456:2000. 4th revision, The Bureau of Indian Standards, New Delhi, 2001.

[187] *Principles and Practice of Bituminous Surfacing*, Vol. 1, Sprayed Work, National Association of Australian State Road Authorities, 1980, Sydney.

[188] Ramkrishnana, R., "Prospects and Problems of Privatisation of Roads," *IRTDA Newsletter*, Vol. LXV, No. 3, March 1996.

[189] Rapp, M.H., and Gehner, C.D., "Transfer Optimization in an Interactive Graphic System for Transit Planning," *Transportation Research Record 619*, 1976, pp. 27–33.

[190] *Realising a Dream*, MORT&H Newsletter, Information Technology and Planning Division, NHAI, July 2001.

[191] *Recommended Design Criteria for the use of Cement-Modified Soil for the Road Construction*, 1st reprint, IRC:50–1973, The Indian Roads Congress, New Delhi, 1978.

[192] *Recommended Practice for Lime Flyash Stabilized Soil Base/Sub-Base in Pavement Construction*, IRC:88–1984, The Indian Roads Congress, New Delhi, 1984.

[193] *Recommended Practice for Traffic Rotaries*, IRC: 65–1976, The Indian Roads Congress, New Delhi, 1976.

[194] Reddy, K.H., and Chakroborty, P., "Procedure to Estimate the Origin Destination Matrix from Marginal Trip Totals and Ordinal Information on Matrix Elements," *Transportation Planning and Technology*, Vol. 22, No. 4, 1999, pp. 247–270.

[195] Reddy, K.H., and Chakroborty, P., "A Fuzzy Inference-based Assignment Algorithm to Estimate O-D Matrix from Link Volume Counts," *Journal of Computers, Environment and Urban Systems*, Vol. 22, No. 5, 1998, pp. 409–423.

[196] Richardson, A.J., Ampt, E.S., and Meyburg, A.H., *Survey Methods for Transport Planning*, Eucalyptus Press, Melbourne, 1995.

[197] Ruiter, E.R., "Towards a Better Understanding of the Intervening Opportunities Model," *Transportation Science*, Vol. 1, 1967.

[198] Rosello, X., "An Heuristic Algorithm to Generation an Urban Buses Networks," *Proceedings of the Second European Congress on Operations Research*, M. Roubens (Ed.), 1977, pp. 409–419.

[199] Sahu, C.S., "Design and Evaluation of Aggregate Gradings for Asphaltic Mixes," *Proceedings of the National Seminar on Emerging Trends in Highway Engineering*, Bangalore University, Bangalore, March, 1995, pp. 1.1–1.11.

[200] Sayers, M.W., "Profiles of Roughness," *Transportation Research Record*, National Research Council, TRB, Washington, D.C., No. 1260, pp. 106–111.

[201] Sayers, M.W., "On the Calculation of International Roughness Index from Longitudinal Road Profile," *Transportation Research Record*, TRB, Washington, D.C., No. 1501, pp. 1–12.

[202] Schreiber, H., *The History of Roads, from Amber Route to Motorway*, Barrie and Rockliff, London, 1961.

[203] Sennov, V.A., "Theory of Use of Granulated Materials in Road Construction," *Transportation Research Record*, No. 1119, 1987, pp. 1–10.

[204] Shaat, A.A., Kamal, M.A., and Matter, N.S., "Relationships Between Climatic Conditions and the Structural Parameters of Flexible Pavements," *Proceeding of the 7th International Conference on Structural Design of Asphalt Pavements*, Vol. III, 1992, pp. 326–340.

[205] Shackel, B., "Repeated Loading of Soils—A Review," *Australian Road Research Board Proceeding*, Vol. 5(3), 1973, pp. 22–49.

[206] *Shell Pavement Design Manual—Asphalt Pavements and Overlays for Road Traffic*, Shell International Petroleum Company Limited, 1978, London.

[207] Shifley, L.H., and Monismith, C.L., *Test Road to Determine the Influence of Subgrade Characteristics on the Transient Deflections of Asphalt Concrete Pavements*, Report No. TE-68-5, Department of Civil Engineering, Institute of Transportation and Traffic Engineering, University of California, Berkeley, 1968.

[208] Shook, J.F., Finn, F.N., Witczak, M.W., and Monismith, C.L., "Thickness Design of Asphalt Pavements—The Asphalt Institute Method," *Proceeding of 5th International Conference on Structural Design of Asphalt Pavements*, Vol. I, 1982, pp. 17–44.

[209] Singh R.P., and Pandey, B.P., "Analytical Design of Rigid Pavements," *Indian Highways*, The Indian Roads Congress, New Delhi, 1997, pp. 5–14.

[210] Sivaneswaran, N., Kramer, S.L., and Mahoney, J.P., "Advanced Backcalculation Using a Nonlinear Least Squares Optimisation Technique," *Transportation Research Record*, No. 1293, National Research Council, 1991, pp. 93–102.

[211] Snedoon, I.N., *The Use of Integral Transforms*, 1st ed., Tata McGraw-Hill, New Delhi, 1974.

[212] Special Publication 30, *Manual for Economic Evaluation of Highway Projects in India*, The Indian Roads Congress, New Delhi, 1993.

[213] Special Report, *State of the Art: Application of Geotextiles in Highway Engineering*, The Indian Roads Congress, New Delhi, 1994.

[214] *Specification for Priming Base Course with Bituminous Primers*, IRC:16–1989. 1st revision, The Indian Roads Congress, New Delhi, 1989.

[215] *Specifications for Road and Bridge Works*, 4th Revision, Ministry of Road Transport and Highways, Government of India, published by IRC, 2001.

[216] *Specifications for Road and Bridge Works*, 2nd Revision, Ministry of Surface Transport, Government of India, published by IRC, 1993.

[217] *Specifications for Road and Bridge Works*, 3rd Revision, Ministry of Surface Transport, Government of India, published by IRC, 1997.

[218] Springenschmid, R., and Fleischer, W., "Recent Development in the Design and Construction of Concrete Pavements for German Expressways," *The Indian Concrete Journal*, Vol. 76, 2002, No. 2, pp. 81–89.

[219] *Standard Letters and Numerals of Different Heights for Use on Highways Signs*, IRC: 30–1968, The Indian Roads Congress, New Delhi, 1968.

[220] *Standard Specifications and Code of Practice for Construction of Concrete Roads*, IRC:15–1981, The Indian Roads Congress, New Delhi, 1981.

[221] Stouffer, S.A., "Intervening Opportunities: A Theory Relating Mobility and Distance," *American Sociology Review*, Vol. 6, 1940, pp. 845–867.

[222] Stovali, T., Larrard, F. De., and Buil, M., "Linear Packing Density Model of Grain Mixtures," *Powder Technology*, Vol. 48, 1986, pp. 1–12.

[223] *Superpave Mix Design*, Asphalt Institute, Superpave series no. 2, 3rd ed., 2001.

[224] *Superpave Performance Grade Asphalt Binder Specification and Testing*, 2nd ed., The Asphalt Institute, 1997.

[225] Taha, H., *Operations Research: An Introduction*, 4th. ed., Macmillan, New York, 1987.

[226] Taragin, A., "Driver Behaviour as Affected by Objects on the Highway Shoulders," *Proceedings of Highway Research Board*, Vol. 34, 1955, pp. 453–472.

[227] *Tentative Guidelines for Cement Concrete Mix Design for Pavements*, IRC:44–1976, 1st revision, The Indian Roads Congress, New Delhi, 1996.

[228] *Tentative Guidelines for the Use of Low Grade Aggregates and Soil Aggregate Mixtures in Road Pavement Construction*, IRC:63–1976, The Indian Roads Congress, New Delhi, 1976.

[229] *Tentative Guidelines for Lean-Cement Concrete and Lean Cement-Flyash Concrete as a Pavement Base or Sub-Base*, IRC:74–1979, The Indian Roads Congress, New Delhi, 1979.

[230] *Tentative Guidelines for Structural Strength Evaluation of Rigid Airfield Pavements*, IRC: 76–1979, The Indian Roads Congress, New Delhi, 1980.

[231] *Tentative Recommendations on the Provision of Parking Spaces for Urban Areas*, IRC Special Publication 12, The Indian Roads Congress, New Delhi, 1988.

[232] *Tentative Specification of Single and Double Coat Surface Dressing*, unpublished, The Indian Roads Congress, New Delhi, 2002.

[233] Teodorovic, D., *Transportation Networks*, Gordon and Breach Science Publishers, New York, 1986.

[234] The Little Book of Profiling, Research publication, University of Michigan, http://www.umtri.umich.edu/erd/roughness.

[235] *The Shell Bitumen Handbook*, Shell Bitumen, London, 1994.

[236] Theyse, H.L., Beer, M.de., and Rust, F.C., "Overview of the South African Mechanistic Pavement Design Analysis Method," Divisional paper No. DP–96/005, TRANSPORTEK, CSIR, Pretoria,1996. pp. 1–43.

[237] *Thickness Design—Asphalt Pavements for Highways and Streets*, The Asphalt Institute, Manual Series No. 1 (MS-1), 9th edition, reprint 1999.

[238] *Thickness Design of Concrete Pavements*, Portland Cement Association, Chicago, 1966.

[239] *Thickness Design of Concrete Pavements*, Portland Cement Association, PCA, Chicago, 1984.

[240] Thomlin, J., "Temperature Variations and Consequent Stresses Produced by Daily and Seasonal Temperature Cycles in Concrete Slabs," *Concrete Construction Engineering*, 1940, Vol. 35(6), pp. 298–307.

[241] Timoshenko, S., and Goodier, J.N., *Theory of Elasticity*, 2nd ed., International Student Edition, McGraw-Hill, Inc. Tokyo, 1951.

[242] Timoshenko, S.P., and Kriewer, S.W., *Theory of Plates and Shells*, 2nd ed., International Student Edition, McGraw-Hill, Kogakusha, 1940.

[243] Torquato, S., Truskett, T.M., and Debenedetti, P.G., "Is Random Close Packing of Spheres Well Defined?," *Physical Review Letters*, Vol. 84, No. 10, 2000, pp. 2064–2068.

[244] *Traffic Census on Non-Urban Roads*, First Revision, IRC:9–1972, The Indian Roads Congress, New Delhi, 1972.

[245] Transport and Road Research Laboratory, *A Guide to the Structural Design for New Roads*, Department of the Environment, Road Note 29, 3rd ed., HMSO, London, 1970.

[246] Tseng, K.H., and Lytton, R.L., "Fatigue Damage Properties of Asphaltic Pavements," *Transportation Research Record*, No. 1286. Transportation Research Board, National Research Council, Washington D.C., pp. 150–163, 1990.

[247] Ullidtz, P., *Pavement Analysis*, Elsevier, New York, 1986.

[248] Underwood, R.T., "Speed, Volume, and Density Relationships." *Quality and Theory of Traffic Flow*, Yale Bureau of Highway Traffic, Yale University, 1961, pp. 141–188.

[249] *Updating Road User Cost Data in India*, Final Report, Ministry of Surface Transport and the Asian Development Bank, L.R. Kadiyali and Associates, New Delhi, 1991.

[250] Uzan, J., "Characterisation of Granular Material, *Transportation Research Record*, No. 1022, TRB, National Research Council, Washington D.C., 1985, pp. 52–59.

[251] Valkering, C.P., and Stapel, F.D.R., "The Shell Pavement Design Method on a Personal Computer," *Proceeding of the 7th International Conference of Structural Design of Asphalt Pavements*, Vol. I, 1992, pp. 351–374.

[252] Veeraragavan, A., *Analysis of Structural Behaviour of Flexible Pavements*, unpublished Ph.D. Thesis, Bangalore University, India, 1989.

[253] Verstraeten, J., "Stresses and Displacements in Elastic Layered Systems, General Theory—Numerical Stress Calculation in Four-Layered Systems with Continuous Interfaces," *Proceeding of 2nd International Conference of Structural Design of Asphalt Pavements*, 1967, pp. 277–290.

[254] Verstraeten, J. "Moduli and Critical Strains in Repeated Bending of Bituminous Mixes Application to Pavement Design," *Proceeding of the 3rd International Conference of Structural Design of Asphalt Pavements*, Vol. 1, 1972, pp. 729–738.

[255] *Vertical Curves for Highways*, IRC Special Publication 23, The Indian Roads Congress, New Delhi, 1993.

[256] Vuchic, V.R., *Urban Public Transportation: Systems and Technology*, Prentice-Hall, Inc., New Jersey, 1981.

[257] Waller, M.F., Jr., "Emulsion Mix Design Methods: An Overview," *Transportation Research Record*, No. 754, TRB, National Research Council, Washington, D.C., 1980, pp. 1–9.

[258] Wang, F., and Lytton, R.L., "System Identification Method for Backcalculating Pavement Layer Properties," *Transportation Research Record*, No. 1384, TRB, National Research Council, Washington, D.C., 1993, pp. 1–7.

[259] Webster, F.V., "Traffic Signal Settings." *Road Research Technical Paper 39*, Road Research Laboratory, 1958.

[260] Wells, A.T., *Airport Planning and Management*, 4th ed., McGraw-Hill, New York, 2000.

[261] Westergaard, H.M., "New Formulas for Stresses in Concrete Pavements of Airfields," *ASCE Transactions*, ASME, Vol. 113, 1948, pp. 425–444.

[262] Witczak, M.W., "Design of Full-Depth Asphalt Pavements," *3rd International Conference on the Structural Design of Asphalt Pavements*, Vol. I, 1972, pp. 550–567.

[263] Wonnacott, T.H., and Wonnacott, R.J., *Introductory Statistics*, 2nd. ed., John Wiley & Sons, New York, 1972.

[264] *World Bank Development Indicators 2001.* http://www.worldbank.org/data/databytopic/databytopic.html.

[265] Yandell, W.O., "Residual Stresses and Strains and Fatigue Cracking," *Proceedings of ASCE*, Vol. 108, No. TE1, 1982, pp. 103–116.

[266] Yang H., Huang, *Pavement Analysis and Design*, Prentice Hall, 1993.

[267] Yang, N.C., *Design of Functional Pavements*, McGraw-Hill, New York, 1972.

[268] Yoder, E.J., *Principles of Pavement Design*, 1st ed., John Wiley & Sons, Inc., 1959.

[269] Yoder, E.J., and Witczak, M.W., *Principles of Pavement Design*, 2nd ed., John Wiley & Sons, Inc., New York, 1975.

[270] Zimmermann, H.J., *Fuzzy Set Theory and Its Applications*, 2nd ed., Kluwer Academic Publishers, Dordrecht, 1991.

[271] Zoramliana, H., "A Study of Lean Cement Concrete Pavement with Soft Aggregates of Mizoram,'' M.Tech. thesis, Dept. of Civil Engg., IIT, Kharagpur, 1996.

[272] Kikuchi, S., and Vuchic, V.R., "Transit vehicle stopping regimes and spacings," *Transportation Science*, Vol. 16, No. 3, pp. 311–331.

Index

PRINCIPLES OF TRANSPORTATION ENGINEERING

PARTHA CHAKROBORTY
ANIMESH DAS

This detailed introduction to transportation engineering is designed to serve as a comprehensive text for undergraduate as well as first-year master's students in civil engineering. In order to keep the treatment focused, the emphasis is on roadways (highways) based transportation systems, from the perspective of Indian conditions.

Salient features include:

- Analysis of those characteristics of vehicles and drivers which affect traffic and design of traffic facilities.
- Principles of road geometry design and how to lay a road.
- Characterization and analysis of flows on highways, unsignalized intersections, and signalized intersections.
- Design principles of traffic facilities.
- Modern problems of mobility, access, and congestion in transit systems.
- Engineering characteristics of pavement materials.
- Structural analysis and design of highway pavements.
- Construction techniques of highways with special reference to the Indian conditions.
- Maintenance strategies for highways.
- Concise reference to elements of highway economics.

The book includes many figures, worked-out examples, and exercises that highlight practical engineering design considerations in real-time problems.

About the Authors

PARTHA CHAKROBORTY, Ph.D., is Professor, Department of Civil Engineering, Indian Institute of Technology Kanpur. His areas of specialization include *optimization in transportation*, *traffic facilities design*, *uncertainty modelling in transportation*, and *traffic flow modelling*. He has several publications to his credit in reputed international journals. Dr. Chakroborty is the recipient of various prestigious awards from international bodies for his work in the area of Traffic and Transportation Engineering.

ANIMESH DAS, Ph.D., is Professor, Department of Civil Engineering, Indian Institute of Technology Kanpur. His areas of interest include *pavement materials, pavement design* and *evaluation*. He has several publications to his credit in various national and international journals. He has received AICTE Career Award for Young Teachers in 2003, INAE Young Engineer Award in 2004, and IRC–Pt. Jawaharlal Nehru Birth Centenary Award in 2006 and IIT Kanpur Batch of 1970 Research Fellowship in 2009.

ISBN: 978-81-203-2084-0

www.phindia.com