Lecture Notes in Economics and Mathematical Systems 460

Prof. Dr. G. Fandel
Fachbereich Wirtschaftswissenschaften
Fernuniversität Hagen
Feithstr. 140/AVZ II, D-58084 Hagen, Germany

Prof. Dr. W. Trockel
Institut für Mathematische Wirtschaftsforschung (IMW)
Universität Bielefeld
Universitätsstr. 25, D-33615 Bielefeld, Germany

Springer
Berlin
Heidelberg
New York
Barcelona
Budapest
Hong Kong
London
Milan
Paris
Santa Clara
Singapore
Tokyo

Bernhard Fleischmann · Jo A. E. E. van Nunen
M. Grazia Speranza · Paul Stähly (Eds.)

Advances in Distribution Logistics

Springer

Editors

Prof. Dr. Bernhard Fleischmann
University of Augsburg
Institute for Production and Logistics
Universitätsstraße 16
D-86135 Augsburg, Germany

Prof. Dr. Jo A. E. E. van Nunen
Erasmus University Rotterdam
Rotterdam School of Management
P.O. Box 1738
NL-3000 DR Rotterdam, The Netherlands

Prof. Dr. M. Grazia Speranza
University of Brescia
Department of Quantitative Methods
C. da S. Chiara 48b
I-25122 Brescia, Italy

Prof. Dr. Paul Stähly
University of St. Gallen
Institute for Operations Research
Bodanstraße 6
CH-9000 St. Gallen, Switzerland

Library of Congress Cataloging-in-Publication Data

Advances in distribution logistics / [editors], B. Fleischmann ... [et al.].
p. cm. -- (Lecture notes in economics and mathematical systems, ISSN 0075-8442 ; 460)
Includes bibliographical references.

1. Physical distribution of goods. 2. Business logistics. 3. Physical distribution of goods--Europe. 4. Business logistics--Europe. I. Fleischmann, Bernhard. II. Series.
HF5415.6.A38 1998
658.5--dc21 98-12669
CIP

ISBN-13: 978-3-540-64288-6 e-ISBN-13: 978-3-642-46865-0
DOI: 10.1007/978-3-642-46865-0

SPIN: 10649602 42/3143-543210 - Printed on acid-free paper

Advances in Distribution Logistics

Editorial

Distribution Logistics is concerned with the design and control of all processes necessary for delivering the products of manufacurers to the customers so as to satisfy their demand. These processes - transport, warehousing, administration and communication - are part of the supply chain where they are linked with the production and the purchase of materials. Physical distribution in its proper sense addresses a great number of customers spread over a large area, a country, a continent or all over the world, and is performed in a distribution network consisting of factories, warehouses, transshipment points, retail shops, etc. It involves different actors - manufacturers, carriers, retailers - with different but overlapping distribution networks and different logistics objectives.

Distribution Logistics has been a subject of research for some thirty years which has produced a rich body of literature concerning problem analysis, various quantitative models and planning methods both for the design of distribution systems and for the control of the operations, in particular inventory control and vehicle routing. However, the circumstances of the distribution business are subject to continuous change. In the seventies and eighties, several trends have gradually increased the complexity and the importance of the distribution tasks and costs: the concentration of the production locations in industry, which implies longer distances for the distribution, the increasing multiplicity of product variants and the growing part of just-in-time deliveries to the retailers. The recent trend of accelerated globalization of the markets favours, on one hand, the development of international distribution systems, on the other hand it has produced a tremendous cost pressure on all distribution processes for all participating parties. As a result, a strengthened confrontation between these parties, but also an increasing number of strategic cooperations can be observed. In addition, the installation of the Single European Market has strongly affected the distribution business in Europe: Transport tariffs have been deregulated, the international competition on the transport market has intensified, and barriers for border-crossing distribution networks have been removed. Finally, environmental aspects of freight traffic play an increasingly important role in the public and political discussion, in particular in view of the dramatic increase of the road traffic in Central Europe.

This development has stimulated an intensification of research on Distribution Logistics since the beginning of the nineties, in particular in Europe. The new conditions are being investigated, new instruments are being developed and first experiences with reorganization projects in practice have emerged.

The volume in hand takes this development into account. It presents recent work of a group of mainly European researchers who have come together at a series of workshops on Distribution Logistics since 1994. The primary orientation of the book is towards both the practice of Distribution Logistics and the decision support by quantitative models and techniques. Practice orientation requires a careful rather qualitative analysis of the planning situation as a first step prior to the development of planning methods. This is the subject of some contributions in Chapter 1 and, partly, of most of the other contributions. On the other hand, the majority of the papers presenting various mathematical methods do not deal with the simplifying standard models of plant location, vehicle routing or inventory control, but consider more or less complicated extensions of those models meeting the particular needs of Distribution Logistics in specific practical situations. Some articles focus on a particular application, but also most of the other articles contain a Section on applications. We therefore decided against organizing this volume in separate parts on „Theory" and „Applications".

The 21 articles have been arranged in five Chapters. The first one is concerned with general frameworks of Distribution Logistics, the other four deal with the main functions: Strategic design of distribution systems and location of warehouses; tactical and operational planning of transport; operational planning within the warehouse; and control of multi-stage inventory in a distribution system.

In **Chapter 1**, the articles of *Boutellier and Kobler* and of *Hagdorn-van der Meijden and van Nunen* both provide frameworks of the strategic planning process for Europe-wide Logistics Systems and define the role of quantitative decision support tools within this process. The SELD (Strategic EuroLogistics Design) model in the former article is intended to consolidate the more conceptual approach of Logistics and Operations Research. The latter paper reports on applications in the food and electronics industry. The paper of *Henaux and Semal* focuses on the delivery service provided to the customers and highlights its key factors. *Corbett, Blackburn and van Wassenhove* consider partnerships in the supply chain and analyze, by means of several real cases, their development and conditions of success. The paper of *Fleischmann* provides a framework for quantitative models for the design of freight traffic networks, comprising the different views of the actors involved. One focus is on modeling transportation costs after the deregulation.

In **Chapter 2**, the articles of *Bruns, Klose, Klose and Stähly* and *Tüshaus and Wittmann* present new models and algorithms for locating facilities, such as warehouses and transshipment points, in a one- or two-stage distribution system. While *Bruns* and *Klose* provide two different algorithms, the two other papers concentrate on modeling techniques and sensitivity analysis, which were applied in practical situations. Also *Daduna* analyzes a particular real-world distribution system and suggests a model for improving its structure. *Wlcek* considers a network of cooperating piece good carriers. He develops a local search heuristic

for the design problem, including the location of depots and hubs and the decision on the transport relations. Moreover, he reports on an application.

Chapter 3 is concerned with various aspects of transport planning within a distribution system. *Stumpf* considers the same network of carriers as *Wlcek* above. However, she tackles the daily control of the vehicles. For the long distance transports, this is a particular vehicle scheduling problem (VSP), but differs considerably from the classical VSP. *Bertazzi and Speranza* investigate the often neglected relationship between transport costs and inventories in a multi-stage supply chain. *Kleijn and Dekker* consider the typical distinction between large orders which are shipped directly from the factory or a central warehouse, and small oders delivered via regional stockpoints. They show the implications of the use of a „break quantity" as determinant for direct deliveries. The paper of *Kraus* presents a model for estimating the length of the vehicle tours from a depot to a given set of customers. This is an important interface between the strategic network design and the operational transport planning as well as a useful basis for the evaluation of distribution networks with respect to environmental aspects.

Chapter 4 contains two papers on the control of the internal transports in a warehouse. *De Koster and van der Meer* investigate different strategies for the control of the fork lift trucks in the distribution center of a computer wholesaler. *De Koster, van der Poort and Roodbergen* study the effects of algorithms for minimizing the length of orderpicking routes.

Chapter 5 is concerned with the inventories in a distribution system, which are required for providing a satisfactory service level to the customers. *Diks and de Kok* and *Tüshaus and Wahl* analyze control policies for two-stage distribution systems and provide approximation procedures for the optimization of the parameters. *De Leeuw, van Donselaar and de Kok* investigate the impact of different forecasting techniques on the inventory level. *Van der Laan, Salomon and van Nunen* consider a reverse logistics system, which includes, besides production and distribution, a return flow of used products to be remanufactured. They give an overview on inventory control models for this new field of research.

The editors are indebted to all authors for their valuable contributions and to the referees whose work, subject to tight deadlines, has been essential to guarantee the quality level of this book. Special personal thanks go to Dipl.-Inform. Kay Holte who gave substantial support in organizing and producing this volume.

Prof. Dr. Bernhard Fleischmann, University of Augsburg, Germany
Prof. Dr. Jo A. E. E. van Nunen, Erasmus University Rotterdam, The Netherlands
Prof. Dr. M. Grazia Speranza, University of Brescia, Italy
Prof. Dr. Paul Stähly, University of St. Gallen, Switzerland

Contents

Chapter 1: Frameworks for Distribution Logistics

Chapter 2: Warehouse Location and Network Design

Chapter 3: Transport Planning and Scheduling

Chapter 4: Operations within the Warehouse

Chapter 5: Inventory Control and Forecasting

Appendix

Chapter 1

Frameworks for Distribution Logistics

Strategic EuroLogistics Design

Roman Boutellier and Rochus A. Kobler

Universität St. Gallen, Institut für Technologiemanagement, Switzerland

Abstract. Both business practices and research approaches in the field of EuroLogistics actually exist. The research itself can be divided into a more conceptual approach and Operations Research, i.e., discrete location theory. Research activities at the Institute for Technology Management at the University of St. Gallen intend to consolidate the achievements of these two research approaches and business practices. The result is the Strategic EuroLogistics Design model - the SELD model. The SELD-model is a conceptual approach aimed for the reconfiguration of EuroLogistics structures. It goes beyond current conceptual frameworks for logistics design by providing detailed analysis and decision tools to the logistics practitioner. The SELD model distinguishes between five main phases. On the one side, this article provides a summary of the model. On the other, two of the five phases are described in more detail.

1. Introduction - Driving forces

Two main forces are driving substantial developments with respect to EuroLogistics. On one side, the *importance of distribution logistics* is increasing:

The ongoing concentration on core competencies throughout all industries leads to a further decrease in the company-owned, individual part of the total value chain. The number of organisations within the value chain is increasing; the same is true of the number of interactions among these numerous partners. Micro Car produces approximately 10% in house with 35 suppliers but with several hundred sub-suppliers.

Formerly, multi-product companies served local warehouses on a local basis. Nowadays, numerous specialised factories serve the local warehouses on an international basis of distribution logistics.

Other factors that lead to more (frequent) shipments and emphasise the importance of distribution logistics, are *just-in-time manufacturing* and the *increasing number of variants*.

Formerly critical competitive components in marketing, such as price and quality, are becoming useless for differentiation in many branches. Price and quality appear to be standardised prerequisites to stay in business. Physical products become more and more interchangeable.

On the other hand, specific service components, such as responsiveness, delivery times, delivery accuracy and delivery flexibility present powerful means to differentiate a product successfully from competitors' by meeting distinct customer service levels. Characteristically, *customer-service* components are heavily influenced by logistics management.

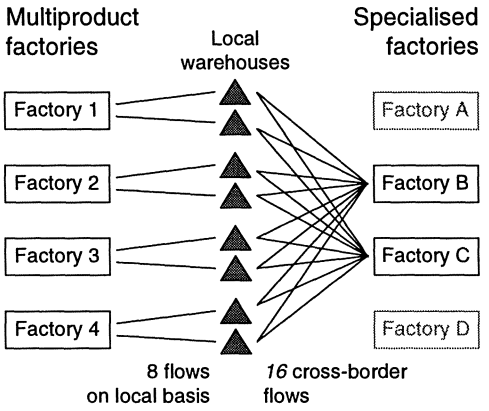


Fig. 1: Impacts of the ongoing specialisation and the concentration on core-competencies

The European political and economic environment has changed rapidly and will go through more transformation in the future. The most significant changes are:

- decreasing transportation costs
- decreasing communication costs
- free market zones (including standardisation of norms)
- increasing productivity
- stable currencies with the introduction of the EURO.

These five factors together with the development towards a *Single European Market* and *the opening up of Central and Eastern Europe's cheap labour forces* will change value chains within Europe dramatically over the next decade. From 1989 to 1993 ABB laid off approx. 40'000 workers in the Western hemisphere and created more than 20'000 jobs in Eastern Europe (see Thurow (1997)).

Some other circumstances within Europe will *not alter or be harmonised* in the immediate future: many national borders coincide with natural barriers. Wide diversity is seen in wealth, economic power, patterns of consumption, modes of ownership and languages.

As the economic and political environment in Europe leads to greater geographical integration of logistics activities, managers face basic structural logistics changes:

- where to place manufacturing processes, facilities, warehousing sites
- how to store and deploy inventories across a geographically dispersed facility network

- how best to serve each of their markets across larger geographic areas
- which transportation modes and carriers to employ.

While rationalising logistics systems often leads to fewer, more focused manufacturing facilities and fewer, distribution facilities, it may also result in a *second level of complexity*. Many more transportation moves result, a greater percentage of which are cross-border hauls. Managing the flow and storage of materials and information across this supply-chain network requires better information systems and more precise co-ordination.

2. The SELD-model

The objective is to establish a framework for Strategic European Distribution Logistics Design - EuroLogistics. The result is the Strategic EuroLogistics Design model - the SELD-model.

Both *business practices* and *research approaches* in the field of EuroLogistics actually exist. The research itself can be divided into a more *conceptual approach* and *Operations Research*, i.e., discrete location theory.

Research activities at the Institute for Technology Management at the University of St.Gallen intend to consolidate the achievements of Operations Research and conceptual design with existing best business practices. Diverse projects support these activities:

- Several User's Groups Logistics Benchmarking (UGLB) gave some insight in actual „best business practices" concerning the field of logistics in general and EuroLogistics explicitly[1]
- The co-operation with a leading *third party logistics provider* allowed the analysis of both the demand side and the supply side of third party logistics services.
- The approach was tested in a division of a mid-size company acting mainly in Europe (Turnover of the division CHF 130 mio; 46 mio items sold in 1995)

The SELD model distinguishes between five main phases. Feedback loops are foreseen between subsequent phases (see Fig. 2). It starts with the set-up of project management. The most important phases are 'analysis' and 'strategic alternatives'. These will be described in more detail in this article[2].

[1] Members of UGLB: ABB Industrie AG, ABB Power Production, Hewlett Packard, Europe, Hilti, ITT Automotive Europe, Landis & Gyr, Leica, Lista, et al.

[2] The full SELD model is published as a PhD dissertation in 1997, University of St. Gallen

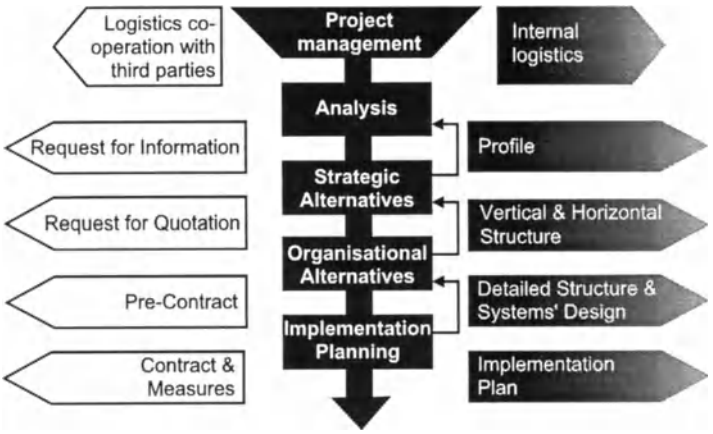


Fig. 2: The SELD-model's five main phases

2.1 Project management

Unfortunately, logistics functions very seldom acquire the appropriate level of strategic importance. It is the task of the project-steering group to clearly define the overall role of logistics. This clarification, in turn, can only be made by *most senior executives*. The main reason is that logistics is a typical "cross-function" within an enterprise (see Pfohl (1996), page 44, fig. 18). This requires *cross-divisional* or at least *cross-functional* co-ordination of logistics decisions, depending on the organisational approach regarding shared resources.

EuroLogistics projects therefore need the commitment of top management. It is vital to involve senior managers as members of the steering team and as project-sponsors.

In EuroLogistics projects not only the nationalities involved, but also the technologies and knowledge needed cover a broad range. A EuroLogistics project calls for the experience and knowledge of specialists in diverse fields, such as IT, corporate finance, EC regulationary frameworks, national and international customs, marketing, purchasing, production, etc., and last but not least *logistics* itself.

Stakeholders have to be identified. All stakeholders have a hidden agenda about what they expect from the project. These expectations have to be disclosed before a project is defined. The *political dimension* within a company cannot be avoided. Very often, this represents the most difficult barrier for cross-functional optimisation.

2.2 Organisational alternatives

Typical organisational alternatives are concerned with questions about the physical material flow, the information flow and the monetary flow. Traditionally, these three streams were organised in parallel. Nowadays, along with the

centralisation of inventories, they are conceptually developed in parallel and carried out in daily business *separately*. Today, a customer in country A gets the product delivered from a country C. The invoice is sent from a third country B. The money flows from bank D to bank E. The physical material flow is closely linked with border crossings and VAT procedures. The potential for optimisation is remarkable, as soon as non-EC countries are involved in distribution processes.

The problems and pitfalls concerning the monetary flow and fiscal aspects are not yet appreciated by many companies as to their full impact. These also include aspects of taxes on profits. However, both legal and practical constraints lead to more or less standardised practices closely linked to VAT procedures. One "good practice" implemented by HP and others is to have *one legal owner* for all goods in transit inside of Europe.

2.3 Implementation planning

The implementation of the new logistics structure has to be divided into programs. Site closing projects, facility expansion projects, new information systems projects are typical examples for the different programs managers are faced with. Typically, smaller pilot projects are initiated:

The implementation of a pilot project takes a shorter time. If problems arise, they can be solved more quickly and their impact on the firm's business can be managed.

Success generates confidence and trust, which are the most important inducements to make people accept cross-functional changes.

The project management, by running through a pilot scheme, learns to cope with bigger projects.

The project scope can be reduced either by limiting the *geographical area* in which the new structures are implemented first, or by restricting the enterprise's business field to one particular *business unit* or a specific *product line*.

A reduction of the geographical implementation scale for the pilot project has proved to be very beneficial. From a *logistical point of view* a pilot project should not represent the 'most difficult' area as far as implementation is concerned - from a *political point of view*, on the other hand, a pilot should certainly not represent the most complex implementation area as well. A failure of the pilot would be damaging for the entire EuroLogistics reconfiguration project.

2.4 Co-operation with third-party logistics-service providers

Companies have come to realise that good logistics performance can enhance the attractiveness of their products to customers. A questionnaire-study (see Gnirke (1995)) shows, that over 50% of 55 respondents perceive the *significance of outsourcing in distribution logistics* as „very high“ and „high“. As a result, they seek providers whose operations can add value rather than simply keep costs at a minimum. Some providers of logistics services thus in effect offer extensions to

the production line, performing activities such as relabeling, repackaging, or even final configuration at their distribution centres.

Providers of third-party logistics services in Europe face *major changes* because of two driving forces: Requirements of users of logistics services are growing rapidly and ongoing deregulation is altering the nature of the transport industry. The European transport industry is evolving toward logistics services that extend beyond transport alone. These services are very difficult to evaluate, because they offer combinations of services rather than the single activity of transport and only limited experience is available.

3. Analysis

The analysis phase has a twofold objective. Firstly, it should alert the project team to the several perspectives from which (distribution) logistics can be perceived. Secondly, it should point out opportunities within EuroLogistics.

The *company analysis* sets the focus for any further analysis and clearly states some first strengths, weaknesses and the goals to be achieved.

The *customer and customer-service analysis* illustrates the importance of customers' needs. The use of market research techniques in distribution has lagged behind their application in such areas as product testing and advertising research. We suggest to perform a *"double ABC analysis"* to improve customer-service effectiveness. The logic behind this approach is that some customers and some products are more profitable to the manufacturer than others. Consequently, the most attention to levels of customer service should be assigned to the most profitable customer/product combination. These reflect the "true" customer requirements that have to be considered most when reconfiguring distribution-logistics structures.

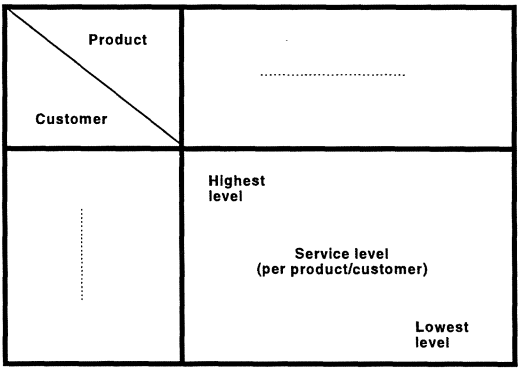


Fig. 3: Service levels, starting points of analysis

Existing *forecast models* and their measured deviations from reality give a first clue about safety stock levels needed.

Next, the current *supply chain* has to be analysed, and the distribution structure itself.

Concluding analysis then allows to identify critical factors either *supporting centralisation*, or respectively *requiring decentralised structures*.

In the following we summarise a few highlights of the analysis phase. They represent the wide variety of reflections necessary when reconfiguring EuroLogistics structures.

3.1 Customer service & customer analysis: Rules of thumb

The most rudimentary customer characteristic is their *geographical spread.*

Customer characteristics	***Require or allow decentralisation***	***Support centralisation***
Geographical spread	high	low
Accepted waiting time	short or *no* waiting time	long
Delivery accuracy	hours	days
Order quantities	heterogeneous, small	homogeneous, large
Order-quantity variability	low	high

Table 1: Customer characteristics and requirements influence distribution structures

The most important factor is the *accepted waiting time* or *product availability.* If buyers do not accept any waiting time but demand immediate product availability, only a decentralised structure is able to accommodate them.

The same is true for just-in-time orders. High *delivery frequencies* and *delivery accuracy* cannot be guaranteed over vast transportation distances. They call for a certain degree of decentralisation.

Order quantities "less than truck load" (LTL) normally account for much higher transportation costs than truck load (TL) quantities. Therefore, if customer orders are of LTL size, the distribution structure has to foresee some decentralised stock-keeping level or possibly a consolidation point.

Last but not least, the *customer order quantity variability* has to be considered carefully. To cope with the highest peak of customer-order quantities, the inventory level held, i.e., the safety stocks at the despatching stock-keeping point has to cover it. The greater the variability is, the greater the risk of immense inventory levels or even obsolete stock items becomes. The more centralised a distribution system is, the better this risk can be confined (law of big numbers).

3.2 Supply chain analysis

3.2.1 The internal value chain - The order penetration-point (see Boutellier and Kobler (1996))

The "time" factor is most important as products become more and more interchangeable, and availability is the main reason for customers to buy at all.

The *total lead time* to rebuild a product can be defined as the total amount of time necessary to order the components, to build a new product and to deliver it to the customer. This time includes the purchasing time, assembly times and transportation times for distribution. The customer, on the other hand, is prepared to *wait* a certain time, depending on the type of product and the market.

The difference between these two times, i.e., the *lead-time gap*, has traditionally been covered by inventory stocks. This, of course, implies huge inventory and obsolescence costs. The increasing number of product variants certainly does not support this approach. More recent approaches try to reduce lead times, or to increase the waiting time accepted by the customer. After all, a clear understanding of the internal logistics chain allows the lead-time gap to be coped with effectively.

The division of the total lead time into the lead time gap and the waiting-time accepted by the customers implies the *order penetration point* (OPP). This is the point on the value-adding chain where the *order meets the plan* (see Sharman (1984)); at this point customisation takes place.

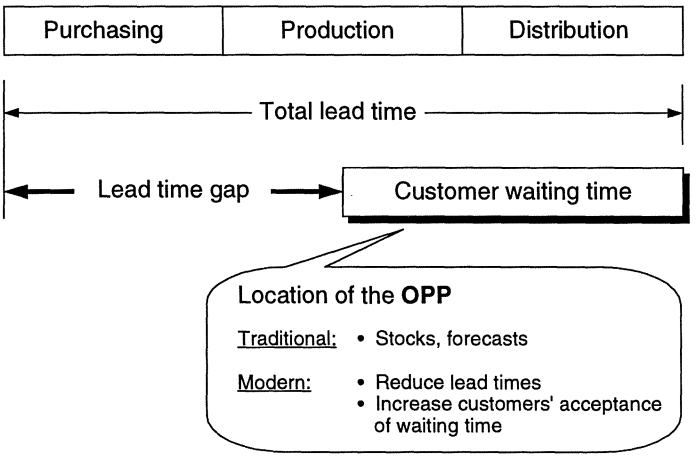


Fig. 4: Dividing the total lead time into the waiting time accepted by the customers and the lead time gap: The order penetration point

The principle to shift the OPP as far downstream as possible is also known as the *principle of postponement* (see Cooper et. al. (1993); Bowersox, Closs, Helferich (1986)). The concept is to avoid commitment through processing until the last possible moment before the customer's order is fulfilled. It has the

advantage of responding to the uncertainty of demand in particular markets through purposeful delay, and thereby avoiding the risks of inventory accumulation and obsolescence.

Processes upstream of the OPP are "pushed" by plan. They are oriented according to economies of scale and characterised by economical lot-sizes and long cycle times. Processes downstream of the OPP are optimised corresponding to economies of scope. Small lot-sizes reduce inventories, but require short set-up times (see Boutellier and Kobler (1996)).

The "appropriate" position of the OPP is vital for the whole logistics chain. The placement of the OPP has to be "designed" during the design-phase of the product. This concept has been introduced only recently by HP.

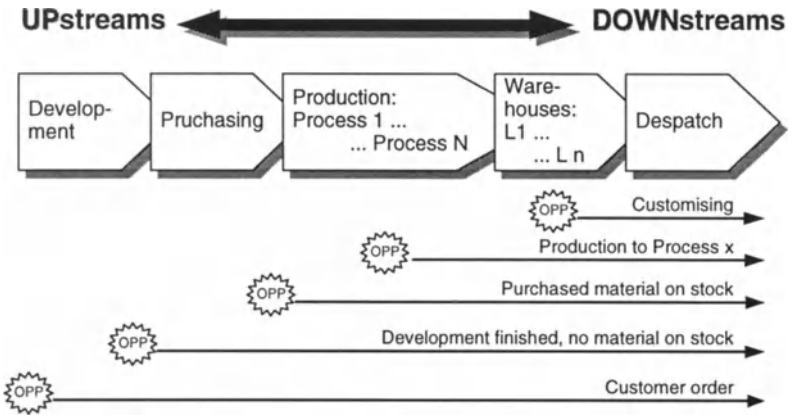


Fig. 5: Possible positioning of the OPP

In *modular production*, the aim of postponement is to retain product homogeneity as long as possible in the production process (multi-purpose modules).

In *postponed assembly*, the final configuration of the product for a particular customer is performed at a regional or local warehouse. This moves the final stage of production, i.e., customisation, into distribution. Customer-specific features and accessories are configured at the market. This is a typical practice in the computer industry[3], where individual requirements for hardware configurations, software and manuals vary over a wide range. Customising, i.e., postponed assembly, is frequently contracted out to logistics service providers who are also responsible for warehousing, order fulfilment and distribution.

Postponed assembly can also be performed in centralised warehouses. In the cases of Rank Xerox or Hewlett Packard, this strategy has been followed. Standard units are matched to individual countries by the insertion of appropriate language chips; and hardware configuration is accomplished according to specific customer requirements, as soon as the order is received in the central warehouse.

[3] UGLB 1995

Postponement reduces inventory costs in several ways. Value is not added until the order is called, reducing the number of different stock-keeping points and the total value of inventory. It also provides flexibility to tailor production processes to individual customers' orders, thus reducing the number of different items in distribution inventories in general.

One indicator which is useful to determine the OPP can be deduced from the *cost curve*, i.e., the increase in the product value according to the lead time. An ideal cost curve would rise slowly at the beginning and have a steep inclination at its end. Hence, the best position for the OPP is just before the curve starts to climb rapidly.

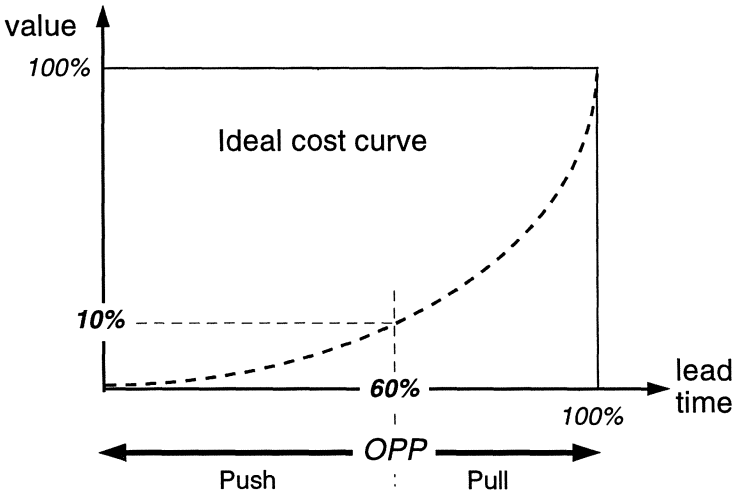


Fig. 6: The adequate position of the OPP: With a stock at 10% of the finished good's value, the lead time gap can be significantly reduced.

Considering centralisation versus decentralisation of distribution activities again, an ideal value curve would allow of positioning the OPP upstream and this, in turn, would support centralisation, since the lead-time gap can be reduced remarkably. Obviously, if production processes are very intensive as concerns fixed assets (e.g., the production or the transformation of raw materials), this is another reason to centralise. However, production processes in general are not objects of decentralisation. Customisation on the other hand, i.e., processes downstream from the OPP, are indeed objects of decentralisation. The fewer fixed assets they tie up, the more easily they may be copied, i.e., decentralised.

The *structure of distribution costs* has to be taken into consideration. An approximate distinction between transportation and inventory & warehousing costs allows one to emphasise either decrease of transportation costs, i.e., decentralisation, or a decrease of inventory & warehousing costs, i.e., to centralisation.

Supply-chain characteristics	*Require or allow decentralisation*	*Support centralisation*
Value-adding curve, i.e., cost curve	high & flat	ideal
Order-penetration point (OPP)	downstream	upstream
Main cost of production processes	variable	fixed
Main cost of customisation processes	variable	fixed
Production sites	multi-product	specialised
Predominant cost factor over total distribution structure	transport	inventory & warehousing

Table 2: Supply-chain characteristics and requirements influence distribution structures

3.2.2 Analysis of the current distribution structure

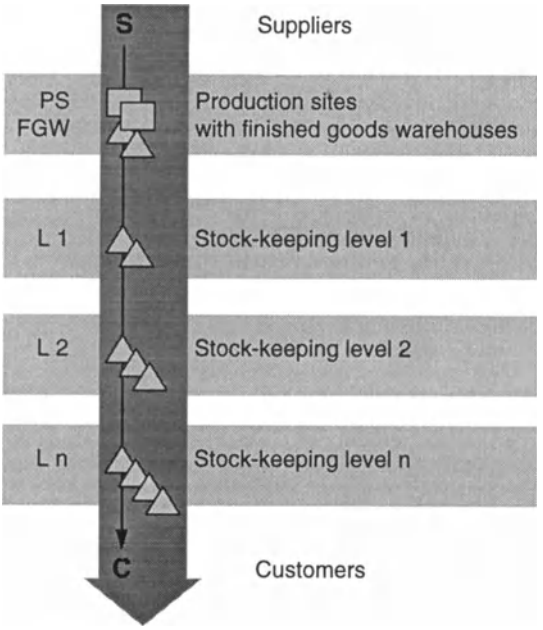


Fig. 7: Description of the current distribution structure

The first objective is to identify the *number*, the *geographic locations* and the *types* of all the stock-keeping points, such as *capacity figures*. Who is responsible to plan capacities and the inventory? Who takes financial responsibility?

A further analysis of the warehouses includes *stock turn rates*[4]. Is there an effective *stock-holding-policy* employed, i.e. are the right articles on stock in the appropriate amount?

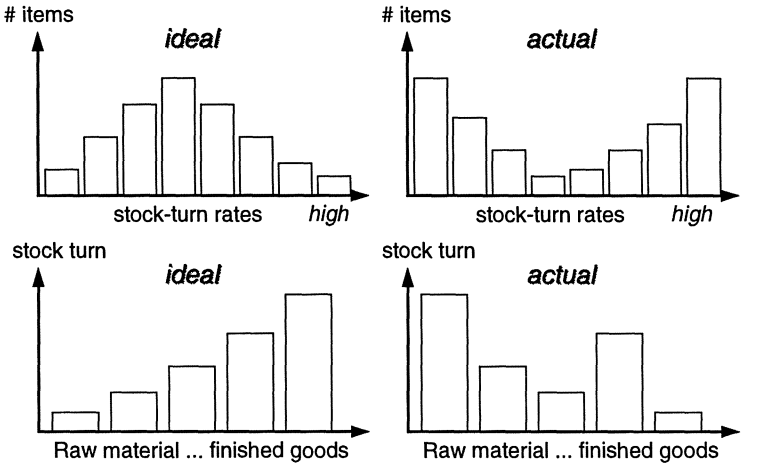


Fig. 8: Stock turns show improvement areas

3.2.3 Total distribution costs

A final step of the supply chain analysis is to assess the total distribution cost of the actual structure. Mainly three types of costs have to be considered: Warehousing cost, Trunking cost and Delivery cost. This is probably the most difficult and tedious task of the whole analysis. Obviously, this will be the basis of comparison for potential distribution structures.

Simplification can be obtained by the principle of "relevant costs": Relevant is only what is different between different alternatives. In most cases only approximate costs can be obtained.

3.2.4 Evaluating the effectiveness of a company's distribution structure

The *first question* to raise is whether the company's customers are satisfied with existing levels of customer service. Insight can be gained from customer loyalty, order cancellations, stock-outs and evaluating the company's general relationships with all channel partners.

Second question is how frequently a need for backordering or expediting occurs. The more frequently these occur, the less effective the distribution structure is presumed to be. The company's inventory management approach may not respond promptly to signals for reordering and re-supplying stock-keeping

[4] The stock turn rate is the ratio between the *inventory level* and the *yearly demand.*

levels. This is also reflected in the frequency of the employment of specific transportation modes.

The *third question* involves inventory turnover measures calculated for an entire product line, for individual products and for product groupings. It should be investigated whether these figures are increasing or decreasing and how they vary among different stock-keeping points in the distribution structure.

A *fourth question* to raise is whether overall inventory as a percentage of sales is directly proportional to a company's sales. Generally, given effective inventory management, this figure should decline as sales increase. Commonly, many firms or stock-keeping points experiencing a growing demand of their products will „over-inventory" those products.

3.3 Average product analysis: Rules of thumb

Several characteristics of the products themselves have to be considered when the degree of centralisation versus decentralisation of distribution activities is resolved. Since it is not really practicable to deal with each product individually, it makes sense to identify some *product-group* characteristics and to assess these.

Product characteristics	***Require or allow decentralisation***	***Support centralisation***
Monetary value	low	high
Monetary density	low	high
Volume	low	high
Physical density (weight-to-volume ratio)	high	low
Innovation rate (price - performance dynamics)	low	high
Environmental hazard	low	high
Need for protective packaging or storage	no	yes
Need for demonstration objects	yes	no

Table 3: Product characteristics influence distribution structures - Rules of thumb

4. Strategic alternatives

A logistics strategy concerning the distribution logistics structure incorporates a long-term commitment of financial and human resources to the movement and storage operations of an enterprise. The specific objective of the strategic decision is to provide an operating structure capable of attaining performance goals from both the internal and the external points of view - either at the lowest possible cost or the highest possible customer-service level. It is a strategic decision how many stock-keeping levels are to be maintained. Other strategic decisions include the number of stock-keeping points and to which locations they will be directed;

which assortments and quantities of materials and finished inventories will be stocked, and where; how transportation will be performed, etc.

The design, or respectively the decision-making process, concerning distribution-logistics structures embodies two principal steps. The first step is related to the vertical dimension of possible configurations, while the second and subsequent decision step deduces their horizontal dimension.

While *standard strategies* can be identified for the vertical structure design, the horizontal decision will include some in-depth *sensitivity analysis*.

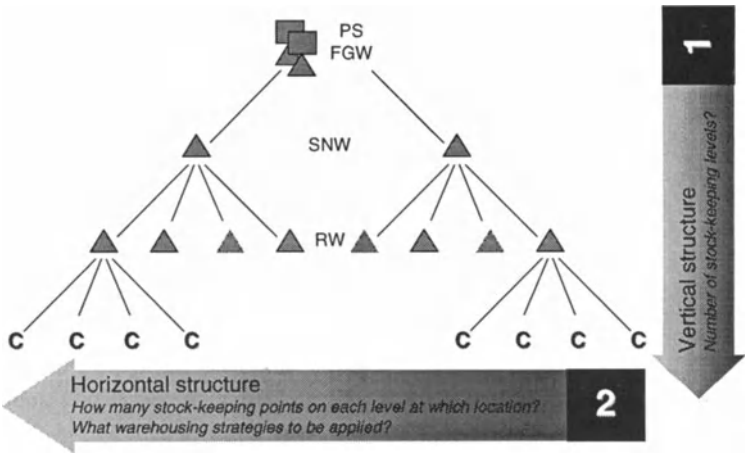


Fig. 9: Two-step design- and decision-making process

4.1 Structural trends

Warehouse networks, i.e., distribution structures, are affected by two main types of transport costs: trunking and delivery. Trunking-costs include those for the delivery from production sites to the warehouses. Trunking costs rise with the number of warehouses.

For the purpose of optimising the total cost of operating the distribution network, the number of warehouses has traditionally been quite high, because delivery costs (including national-border crossing) drew the distribution warehouses closer towards the markets.

A fundamental change is currently taking place in distribution in general. Both trunking and delivery costs are decreasing. This is partly a result of improved infrastructure, improved carrier efficiency due to better communication systems, removal of trade barriers, the rise of third-party carriers who combine truck movements, and last but not least deregulation. The future optimal solution will include fewer distribution warehouses, i.e. more centralised distribution structures.

This change involves moving from a traditional structure of reliance on chains of stock-keeping points connected by a multiple-level transportation system, towards planned delivery systems such as reduced, i.e., *centralised distribution*

structures, in which intermediate inventory is held only for processing, relying on an information system for co-ordination. They are not mutually exclusive.

Owing to the several *driving forces*, the number of stock-keeping levels in Europe can be reduced. If it was a 4-level distribution structure before, it is a 3-level distribution structure today, which theoretically reflects the greatest degree of decentralisation today[5].

This is owing to the fact that national warehouses no longer are indispensable from the political, i.e. border-crossing, point of view. In today's Europe, it is possible to achieve the same delivery performance by meeting cross-border demand from supranational or regional warehouses.

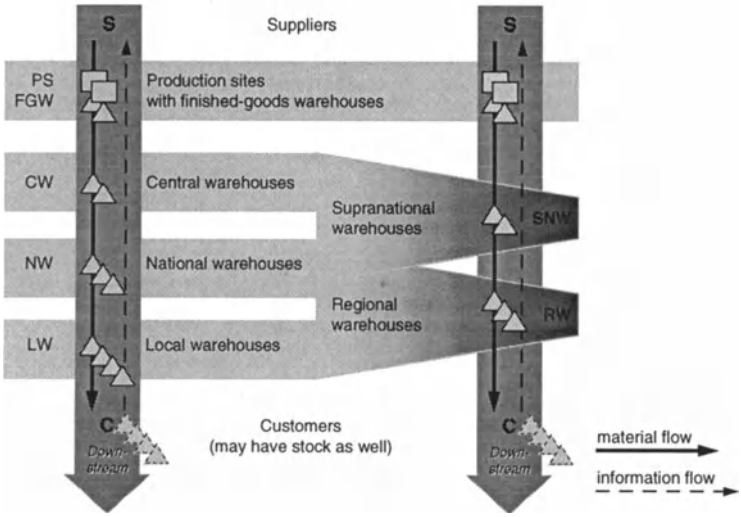


Fig. 10: The traditional distribution structures turn out to be obsolete: the most decentralised distribution structures within the "border-less" Europe has 3-stock-keeping levels.

4.2 A spectrum: National - regional - supranational - central

It is important to recognise the spectrum on which these standard strategies can be positioned in order to express their degree of centralisation. Obviously, the *type* of stock-keeping points applied, e.g., warehouses, does not indicate the structure as clearly as the *number of stock-keeping levels* employed.

---

[5] According to Gnirke, 1995, pp. 85-86
Most of the companies have not adapted their structures until today:
Two thirds of 56 responding international companies confirm having *initiated the reconfiguration of their distribution logistics in Europe*. On the other hand, only 25% of these are already in the implementation phase. The size of the companies is significantly correlated to their advance in EuroLogistics.

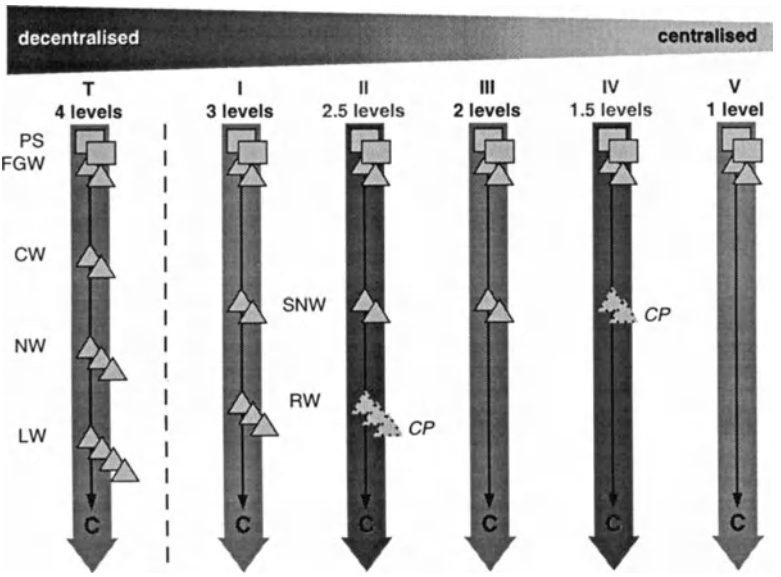


Fig. 11: The spectrum of alternative distribution-structure design.

4.2.1 National distribution structures (Traditional strategy T)

From a marketing perspective, the central issue is the extent of control by local sales organisations over distribution. The traditional system gives *autonomy and control to local operations*, but it also separates important elements of the supply chain. It may lead to multiple inventories. The advantages of integrating the supply chain are substantially abandoned.

The traditional path of product flow has been as follows: origination at the manufacturer's finished-goods warehouse (FGW), continuing through a series of intermediate stock-keeping points, such as central warehouses (CW) - national warehouses (NW) - and local warehouses (LW), and then delivery to customers, such as retailers or end-users. This represents a *four-level* distribution structure.

Even if each unit operates efficiently, the structure as a whole can generate *substantial inefficiencies* because of failures to match stock to actual demands, the length of time that stock remains in the distribution pipeline, and the delaying effects of cumulative decision rules (see Joiner (1994): „Costs of tampering", pp. 122-125), e.g. the whipsaw-effect (see Towill (1991)).

A further problem is that, as demand changes either by product or by volume, the structure is inherently *slow to respond*, leaving unsatisfied demands while accumulating needless inventories. As the number of stock-keeping levels increase, these structures become more difficult to manage, requiring close monitoring to ensure that stocks are available at the end of the pipeline to meet demand.

Traditional distribution structures appear to become *obsolete* in the Single European Market. Still, these structures do exist: on the one hand only a few companies adapted their distribution-logistics structures until today, on the other there are good reasons for some industries not to reconfigure their structures immediately.

Because of its close relationship with consumers, the retail and fast-moving consumer-goods industries[6] are especially influenced by local differences in culture, taste, consumer preference, and environmental regulation. At the same time easier border crossings are creating opportunities for retail firms to increase geographic integration. The phrase "think globally, act locally" implies distinct challenges for European retailers and fast-moving consumer-goods manufacturers.

4.2.2 Regional distribution structures (Strategy I)

This type of distribution structure comprises a *three-level* configuration. To eliminate one level as compared to the traditional structure type, either local and national stock-keeping points or national and central warehouses are merged.

However, the *disadvantages* concerning enormous inventories combined with a *tardy response* to market changes cannot be diminished to any great degree. Meanwhile, some manufacturers[6] warn against pursuing indiscriminately a pan-European approach to logistics.

Firms are starting to design and implement new network strategies that go beyond national boundaries, setting up "islands of integration" rather than developing fully integrated pan-European approaches for the distribution of their products. They are also streamlining logistics processes and rationalising warehousing mainly at national, and sometimes at cross-border levels.

In the view of a manufacturer in the consumer durable industry, companies must consider local / regional market differences in both purchasing patterns and market position[6]. This group's business strategy for the year 2000 is based upon "the desire to maintain a strong and differentiated brand image and to add value to the customer by reconfiguring distribution channels to achieve competitive advantage from 'direct-distribution service'".

The company aims at maintaining a differentiation strategy that applies to both products and services, and is ready to make trade-offs between standardisation and differentiation. Its distribution logistics strategy will be configured to deliver differentiated services in order to meet customers' unique buying criteria. This might mean one distribution system to serve the Northern European buying groups and another to meet the buying criteria of customers in Southern Europe.

[6] UGLB 1995 and 1996

4.2.3 Supranational distribution structures (Strategy III)

Two-level distribution structures centralise inventory mainly at one stock-keeping level. In general, finished-goods warehouses at the different production sites hold only a small range of finished goods, i.e., their own products. The supranational warehouses therefore take over the role of consolidation points with a complete, or at least very wide, range of goods in stock. Obviously, inventory holding can be reduced to a great extent within this configuration. Area-specific peculiarities and *customer requirements* can be taken partly into consideration when defining supranational regions.

Characteristic numbers of supranational warehouses within Europe range between three and eight, depending on the geographical spread of the market and the homogeneity of customer requirements. Most regional-structured distribution-systems are based upon supranational patterns, where regional or local storing facilities or sales points are supplied by large warehouses.

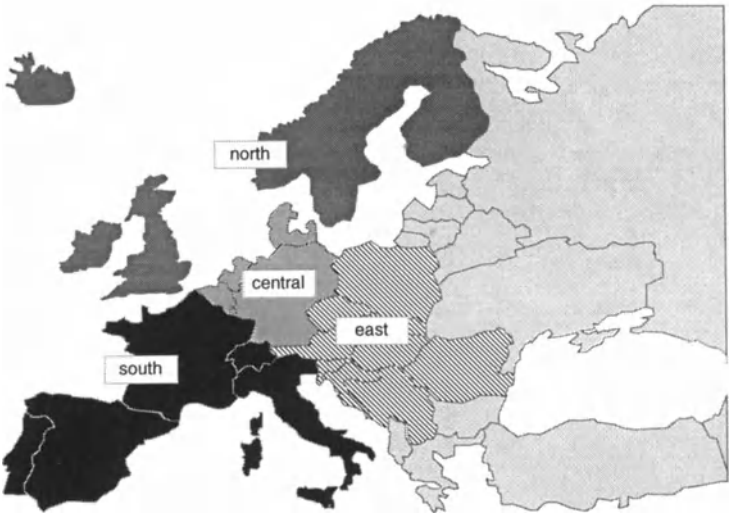


Fig. 12: Supranational European distribution structure divides Europe into four logistics areas[7]

4.2.4 Central distribution structures (Strategy V)

The "leanest" of all distribution structures is represented by the *one-level* configuration. Here, finished goods are shipped directly to customers.

Actually, the pure one-level configuration is only rarely employed for the distribution of products to *end-users*. It is more feasible to supply downstream

[7] UGLB 1995 and 1996

assembly plants or *retailers*. In the case of finished goods being transferred to end-users, so-called consolidation points are usually involved.

The greatest concern arising from these configurations is the *deterioration of customer service* and the consequent loss of business. Therefore, companies considering the centralisation of their distribution structures to this extent ought to make sure that neither of these concerns turns out to be well justified.

Unlike the fast-moving consumer goods industry, the business equipment industry[8] does not need to implement pan-European product strategies. Products have always been pan-European. The issue is that goods have been sold, monitored, and managed at national levels.

The strategy of focusing on efficiency rather than growth in *all aspects* of the business leads to clear objectives in the channel and the distribution-logistics network strategy:

- Centralise warehousing and cut the costs of maintaining local warehouses.
- Fill orders from a central warehouse or from cross-border warehouses to reduce inventory levels.
- Make use of configuration centres to retain a local focus and add value to the local customer.
- Develop a distribution network which can cost-effectively refill sales orders.
- Centrally co-ordinated information flows and centralised order-intake and processing.

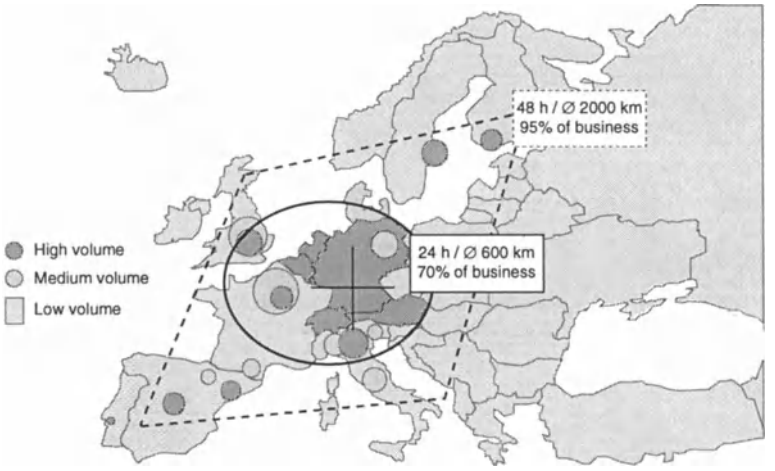


Fig. 13: Central distribution structure with a central warehouse in Germany is capable of covering 95% of the total business volume within 48h[8]

Last but not least, this type of distribution structure applies not only for finished goods. Some observers believe that the spare parts supply chain would be

[8] UGLB 1995 and 1996

a good candidate for this approach[9]. Generally, the number of those parts is huge and many are rarely needed. Demand fluctuation is high and there are many slow-moving items. Stocking them in as few places as possible improves efficiency dramatically. In fact, one case-study company has developed concepts for pan-European spare parts distribution simultaneously with the reconfiguration of the finished goods distribution.

4.2.5 Consolidation points and configuration centres (Strategies II and IV)

Consolidation points are signified as *semi-level* structures in the complete spectrum above. They arise for two reasons. On the one hand, distribution structures including distinct sources, i.e., production sites, for different products, may have to ensure that products from several sources reach the customer in one and the same delivery. If there is no stock-keeping level employed that stores all the products demanded between the production sites and the customer, they have to be merged at some other point, that is to say, at a consolidation point.

On the other hand, *transportation economies* often dictate full truck-loads (TL) being brought in whenever an inventory is replenished or a demand is satisfied. This leads to extensive cycle stocks. The alternative is to bring in smaller quantities and pay higher rates for less-than-truck-load (LTL) shipments.

Another alternative is to define a consolidation point, also known as *hub-and-spoke systems*, that were originally developed for the traditional freight systems for small packages and general cargo. However, the concept has been refined by express companies and the international forwarding companies and logistics-service providers. The principle is that all goods in an area are collected at a central point (consolidation point 1), consolidated and shipped to other consolidation points. There, the goods are consolidated with goods coming from other consolidation points and distributed to the local receivers. The system utilises the economies of 'full truck loads' between the consolidation points. Furthermore, the consolidation points employ *quick cross-docking operations* by automatic sorting and advanced information technology.

As companies move toward central warehousing, the question of what to do with national warehouses arises. To achieve economies, the number of these facilities is likely to be reduced[9].

Companies will retain some sort of physical location in the form of consolidation points or configuration centres in the country of final destination to meet country-specific customer requirements. In any case, European stock will be managed as a whole.

Configuration centres are becoming an increasingly important component of the supply chain. They are unique to the computer industry[9], and particularly important for those companies which seek to differentiate themselves from their competitors through value-added services. Work done on the basic box will

[9] UGLB 1995 and 1996

include installing extra memory cards, bigger or additional disk drives, communication cards, and the desired operating system and software, and ensuring that the resulting ensemble works with a particular printer or other peripheral device. Once the system is reconfigured and tested, it is repackaged and shipped to the customer.

4.2.6 Free points of storage

"Free points of storage", also referred to as "free points of delivery", denotes a concept of distribution in which the storing location of a particular product need not to be strictly determined.

In practice, however, companies do not (yet) apply this concept as a basic distribution strategy mostly because of different transfer prices and incentives at the different management levels. Therefore, the 'free points of storage' concept is only used to cope with *emergency cases*, where a particular product is not available in the right place and has to be expedited to customers directly. If a regional warehouse runs out of supplies of a certain product, for example, another warehouse may still have some stock available to ship either to the previous one or directly to the customer.

4.3 Horizontal structure design

On the one side, there is the allocation problem: *the number of* stock-keeping points, i.e. warehouses, and *their locations* in the distribution structure. On the other hand, *warehousing strategies*[10] have to be defined for the various locations. The latter decision component can be made on a local basis. However, the determination of the actual number of stock-keeping points is more intricate and requires a *sensitivity analysis*.

4.3.1 Sensitivity analysis: Number of stock-keeping points

At this stage, further in-depth analysis is indispensable to select the *most appropriate alternative* and to define the adequate *detailed structure*. The suitable *analysis tool* should be selected. A wide range of methodologies including *"manual" analysis* and the application of *Operation Research tools* (OR-tools) for decision support on the other are available.

Indeed, it makes sense to start sensitivity analysis by carrying out some manual analysis. This again leads to a more comprehensive understanding of the various forces activated, their impacts and affinities.

Customer satisfaction and total-cost analysis are the key to managing the physical-distribution function. Management should strive to minimise the total costs of physical distribution rather than attempt to minimise the cost of each

---

[10] Strategic decisions concerning warehousing include considerations about the *type of ownership, warehousing functions, stock holding policy, warehouse layout and design* and *warehouse handling systems*.

component. Attempts to reduce the cost of individual physical-distribution activities may be sub-optimal and lead to increased total costs. For example, consolidating finished-goods inventory in a small number of warehouses will reduce inventory-carrying costs and warehousing costs, but may lead to substantial increase in freight expense or lower sales volume as a result of reduced levels of customer service. Similarly, the savings associated with large, optimised batch-size production may be less than the increased inventory costs; or the reduction of field inventory may result in increased production set-up at production sites.

In order to achieve least-cost physical distribution, it is necessary to reduce *the total* of various cost elements. Besides trunking cost, distribution cost and warehousing cost, these include the following and others:

- Shortage costs
- Production-lot quantity cost
- Reduction of capital- and storage-space cost
- Reduction of safety stocks, i.e., risk costs
  The objective of setting safety stocks is to achieve the correct level to protect oneself against both the supply and demand uncertainties inherent in the lead time. The greater the number of warehouses, the greater the proportion of safety stocks to be carried.
  The "square root law"[11] states that safety stock can be reduced by the square root of the number of warehouses (see Maister (1993)).
- Economies of scale in central warehouses
- "[...] We expected the effects of economies of scale to be greater [...]"[12]. Studies of warehouse operation do not indicate strong economies of scale (see Pfohl et. al. (1992)). However, there are other arguments in inventory management, transportation management, automation and IT.

All these considerations above either focus on the inventory of *a single* stock-keeping point (or one stock-keeping level) in the supply chain or they represent *rough estimations* how cost curves and functions could develop. However, SELD consistently emphasises the need to consider any such inventory in the context of the *total supply chain* and to consolidate the effects from several cost curves.

There is no contradiction here. The fact is that such an optimisation of the total supply chain becomes extremely complex. It cannot be accomplished "manually". The goal is to determine a set of relevant factors for a specific company, to define an inventory strategy at one point in the chain, and then to model the entire supply chain to see how those factors and strategies coincide.

---

[11] The square root law assumes that each individual warehouse serves an exclusive market area, demand varies randomly and safety stock levels are statistically determined. Concentrating inventories in fewer locations aggregates demands, but safety stock requirements only increase as the square root of variation in demand.

[12] UGLB 1995 and 1996

References

Boutellier, R. / Kobler, R. A. (1996): Ganzheitliches Management der Wertschöpfungskette; in: Logistik im Unternehmen; VDI Verlag, 9, pp. 6-11.

Bowersox, D.J. / Closs, D.J. / Helferich, O.K. (1986): Logistical Management / A Systems Integration of Physical Distribution, Manufacturing Support, and Materials Procurement; Third edition, New York, Macmillan.

Cooper, J. / Cabocel, E. / O'Laughlin, K.A. (1993): Reconfiguring European Logistics Systems; Council of Logistics Management.

Gnirke, K. (1995): "State-of-the-Art" der Reorganisation der EuroLogistik internationaler Unternehmen / Forschungsbericht einer empirischen Untersuchung; Marburg, Lehrstuhl für Allgemeine Betriebswirtschaftslehre und Logistik der Philipps-Universität.

Joiner, B.L. (1994): Fourth Generation Management; New York, McGraw.

Maister, D.H. (1993): Managing the Professional Service Firm; New York, Free Press.

Pfohl, H.-Ch. (1996): Logistiksysteme; 5. Aufl., Berlin, Heidelberg, New York, Springer.

Pfohl, H.-Ch. / Zöllner, W.A. / Weber, N. (1992): Economies of Scale in Customer Warehouses: Theoretical and Empirical Analysis; in: Journal of Business Logistics; 13(1), pp. 95-124.

Schumacher, W. (1988): Die Entwicklung der betriebswissenschaftlichen Logistik und ihr Einfluss auf das zukünftige Leistungsbild des deutschen Speditions- und Lagereigewerbes; Köln.

Sharman, G. (1984): Rediscovery of Logistics; in: Harvard Business Review; September-October.

Thurow, L.C. (1997): The Future of Capitalism, New York, Penguin, p. 168

Towill, D.R. (1991): Supply Chain Dynamics; in: International Journal for Computer Integrated Manufacturing; Vol. 4, No. 4, pp. 197-208.

A Roadmap for Joint Supply-Chain Improvement Projects Based on Real-Life Cases

Charles J. Corbett[1], **Joseph D. Blackburn**[2] and **Luk N. Van Wassenhove**[3]

[1] Anderson Graduate School of Management, Box 951481, UCLA, Los Angeles, CA 90095-1481
[2] Owen Graduate School of Management, Vanderbilt University, Nashville, TN 37203
[3] INSEAD, 77305 Fontainebleau Cedex, France

Abstract

Forming closer partnerships with suppliers or customers can yield substantial benefits, as a slew of examples in the business literature show. Though several characteristics of successful partnerships are brought up time and again – mutual trust, commitment to the partnership, open information exchange – the literature remains strangely silent on the detailed mechanics of the *process* involved in getting there. For a large and all-powerful customer this may not matter much, but what if the *supplier* is driving the process? And how can the supplier make sure to reap the commercial benefits from their efforts?

We describe how a supplier initiated two cases of such joint supply-chain improvement efforts, the first guided primarily by trial and (frequent) error, the second seemingly much more streamlined, but both eventually leading to a mixture of success and disappointment. By contrasting the two, we identify various *critical factors* for smooth project progress, including the need to have a clearly defined process, the composition of the joint project team, and the importance of supply-chain mapping. We pull together the key learnings into a simple framework which can be seen as a "*roadmap*" for joint supply-chain improvement projects, and offer guidelines on managing the overall process and the individual steps. We briefly describe the company training programme through which the framework was implemented and the type of coordination structure needed to support such projects. Finally, we analyze why the first case was eventually *commercially* more successful, even though the second case was more successful at the *project* level: successfully improving relations with suppliers or customers also requires corresponding realignment of *internal* relations between departments.

1 Introduction

1.1 Advantages of partnership

Examples of successful customer-supplier partnerships are spreading fast. Among the most well-known partnerships are WalMart and Procter & Gamble (Hammer and Champy, 1993), Baxter and American Hospital Supplies (Byrnes, 1993), and Toyota with its first-tier suppliers. These and other firms have often reaped substantial benefits from these partnerships, through increased market share, inventory reductions, improved delivery service, improved quality, shorter product development cycles, etc. Even on a less grand scale than a full-fledged company-wide partnership, an improved customer-supplier relationship can already often lead to quick improvements in logistics, through the more open information exchange and better coordination compared to a more traditional arm's-length or adversarial relationship.

1.2 Characteristics of successful partnerships

Several studies have contrasted successful partnerships with more arm's-length customer-supplier pairs, finding some key characteristics that recur time and again. More open information exchange, eg. of cost and demand data, and more coordinated decision-making can go a long way toward reducing the inefficiencies inherent in arm's-length relationships. Mutual trust is crucial, eg. to reassure firms that information shared with a partner will not be used against them. Longer-term commitment to the partnership is needed, to encourage parties to invest in further improving the joint supply chain to mutual advantage. These issues are studied in among others Anderson and Weitz (1992) and Magrath and Hardy (1994).

1.3 But how to get there?

Although recognizing distinguishing characteristics of successful partnerships is important, this still leaves several key questions for suppliers wishing to embark on partnership drives with customers (or vice versa):

1. How can one build such a partnership over time, especially when the relationship with the customer or supplier in question has hitherto been more arm's-length? Ie., how to encourage information exchange, build trust, and create a longer-term commitment in such a context?
2. How can the partners-to-be overcome or avoid the obstacles they will almost inevitably encounter?
3. How can the supplier make sure to reap commercial value from his efforts, through e.g. increased volume or reduced price pressure from the customer?

To a large extent, existing literature ignores these "how-to" questions and focuses on successful partnerships rather than on the obstacles faced in getting there or on failed attempts. One exception to this rule is Byrnes and Shapiro (1991), and some of their findings on the critical factors in such projects do

indeed correspond to ours. We go a step further by integrating these critical factors into an operational framework to address the key questions above and to guide suppliers intending to initiate joint supply-chain improvement projects. It is based on our experience working with a large multinational chemical company, Pellton International,[1] seeking to improve supply-chain logistics jointly with their customers, primarily automotive suppliers. We start by describing two joint supply-chain improvement (JoSCI) projects initiated by Pellton. Their first, with Basco PLC, was driven largely by trial-and-error; progress was slow and faltered several times, but was eventually considered successful, leading to *logistics* benefits and to helping to turn around a traditionally fairly adversarial relationship which, in turn, translated into *commercial* benefits for Pellton. The second project, with Perdirelli Milan, promised similar logistics benefits in far less time but, in the end, did not yield the commercial value Pellton had hoped for.

1.4 Learning from experience

By contrasting the two projects, we (tentatively) identify some key steps in JoSCI projects, notably the importance of having an agreed-upon process to follow, of selection of the joint project team, and of including supply-chain mapping early on in the project. These and other findings were subsequently tested and refined by participating in and following several other JoSCI projects, which we do not describe here. We then pull the learnings together into a simple framework which can be seen as a "roadmap" for JoSCI projects, suggesting the key steps to be taken in such a project and offering guidelines for each. The steps of the roadmap may not come as a surprise in themselves, but the cases do suggest several critical lessons:

1. It is helpful to have a mutually agreed *process and objectives* before embarking on a JoSCI project.
2. Thorough *preparation* is key: team selection, benefit sharing agreements, analysis of opportunities, supply-chain mapping, choosing performance measures, recognizing and allocating the required resources.
3. Reaping the commercial benefits requires that the JoSCI efforts be well-integrated within the respective organizations, especially within the customer's purchasing group. Benefits also need to be gained by departments providing the resources, especially in matrix organizations.

These are the issues we focus on, though naturally project management and implementation are as important here as anywhere else. We sketch how this roadmap was implemented within Pellton through a company training program, focusing on supply-chain management in general and the JoSCI framework in particular. We also indicate additional requirements on the internal coordination structure for implementing and managing such efforts on a larger scale. Finally,

---

[1] Company names and various other details have been disguised for reasons of confidentiality.

we explain why the Basco project was eventually commercially successful despite being initially slow, whereas the Perdirelli seemed far smoother but led to commercial disappointment for Pellton.

2 Background of the case

2.1 Pellton International: supplying chemicals to automotive suppliers

2.1.1 Product and process

Pellton International is a multinational firm supplying a chemical Pell-Q to (among others) automotive suppliers, as depicted in Figure 1. Pell-Q is delivered in various different formulations and sizes; all told, several thousand SKUs. It uses global sourcing, with a small number of high-capacity production lines world-wide, where the various formulations are produced in batches ranging from several hours to several days. Changeover times vary from several minutes to a full day, depending on the products involved. Cycle times between batches of the same product range from 10 days to six months. Process quality and changeover times are highly variable. Pell-Q is cut to size and packaged to customer specifications on-line and stored until shipped. Pellton is the largest supplier with approximately 30% of the total market; four competitors share the balance. We focus on Europe, which Pellton supplies from its plant in Maastricht, The Netherlands. Basco PLC is Pellton's largest customer and accounts for 20% of their output; Perdirelli follows with 15%. Basco and Perdirelli and their competitors each supply to several auto assemblers. A very limited number of auto assembly plants use single sourcing. Pell-Q is a key raw material for Pellton's customers, most of whom are under heavy cost pressure, especially those supplying to OEM producers.

2.2 Their question: how to set up future joint supply-chain improvement projects?

2.2.1 How to capture learnings from first JoSCI projects?

In October 1995, Pellton was in the midst of supply-chain improvement projects with two key customers. By that time, the potential benefits of such projects had become clear, and they intended to initiate similar projects with many important customers world-wide. However, the first two projects had also proved highly resource-intensive, and had largely been driven by a single logistics manager. Moreover, the first project had encountered numerous obstacles and delays, many of which had later been avoided in the second. To set up such projects on a larger scale, they needed to capture the learnings from these experiences and carry them over to others within the organization. The aim of our project with Pellton was precisely that: documentation, improvement, and formalization of the supply-chain improvement process, and assistance in developing and

delivering the necessary training materials. Below we sketch the first two JoSCI projects, the first with Basco PLC in some detail, the second with Perdirelli's Milan plant only briefly. Chronological summaries of both projects are given in Figure 2.

3 Basco PLC: slow and painful, but eventually successful

3.1 Basco PLC: background

3.1.1 Basco moves to centralized purchasing

Relations between Pellton and Basco, its largest customer, have long been rather adversarial. Purchasing of strategic products, including Pell-Q, was being centralized at Basco Purchasing, which would manage negotiations with key suppliers and operate as a service center to Basco production sites. Basco Purchasing and Pellton came to realize that some form of partnership was needed; they reached a long-term commercial agreement, exchanging volume commitments for price concessions.

3.2 Logistics: aiming for SKU rationalization and consignment stock

3.2.1 Preparing the first workshop

During 1993, Pellton underwent major business redesign, which included setting up joint workshops with major customers to find out how to serve them better. Initial discussions with the Basco Purchasing group identified cost reduction as their overriding concern, so the objective of the first workshop was defined as "finding ways to reduce the cost of supplying Pell-Q".

3.2.2 The first workshop

The first workshop, in early 1994, had some 20 participants: from Basco Purchasing the managing director and chemicals purchasing director, the technical director of Basco and an assistant, and several Basco plant-level production and logistics managers. From Pellton the commercial director, sales manager, logistics manager, and several other sales and manufacturing staff attended. The commercial director opened by showing that a single Basco plant took over 60 different SKUs. That plant's logistics manager was sceptical: "we don't take all that, I don't believe it". No specific plans resulted, but they agreed that Pellton would draw up a list of proposals to be refined before a follow-up workshop two months later.

3.2.3 Working towards SKU rationalization

Afterwards, the sceptical Basco logistics manager did some homework and found, to his surprise, that Pellton's commercial director had not exaggerated. That convinced him of the need to work together. SKU rationalization was identified as a major opportunity: rather than supplying Pell-Q in every possible size, Pellton would offer only a more limited variety. This would allow substantial safety stock reductions, improved delivery service, and less rush-order-induced disruption for Pellton. They would have to reach agreement on how to compensate Basco for the additional trim losses they would incur by having a more limited variety of sizes to choose from; offering consignment stock would be part of that compensation.

3.2.4 The second workshop: agreeing on three projects

The second workshop saw changes in team members on both sides. The team agreed on three projects for further study: 1) inventory reduction; 2) bar coding; 3) packaging. A steering team was formed to oversee progress; it included the Pellton logistics manager and the Basco Purchasing chemicals purchasing director as main actors. Teams drawn from both companies were assigned to the three subprojects. Three years on, packaging has essentially become an internal Pellton affair; bar coding was abandoned as the team realized the goal and scope were not sufficiently clear.

3.2.5 Aiming for inventory reduction

The inventory reduction team was composed of logistics managers and sales staff from Pellton, a plant-level purchaser and logistics managers from Basco, and was led by the Basco Purchasing chemicals purchasing director. They first met in the Spring of 1994, and set three priorities: SKU rationalization, improving and integrating forecasting and ordering systems, and implementing JIT deliveries and consigned stock where possible. Total Pell-Q inventory at Basco sites ran into several tens of millions of dollars; Pellton expected the safety stock reduction from SKU rationalization to offset their additional inventory burden from consignment.

3.3 Organizational awkwardness, and failed implementation

3.3.1 "How shall we share the benefits?"

Although the implicit assumption had been that both sides should gain, no explicit agreements had been made. When Pellton's logistics manager raised the issue, the Basco Purchasing chemicals purchasing director responded by saying "let's focus on opportunities first and make sure there are benefits, then we can talk later about how to share". Six weeks after their first meeting, the team agreed that implementation at all Basco sites should be completed within half a year, by January 1, 1995.

3.3.2 Planning the roll-out in a decentralized organization

During subsequent meetings, it became clear that Basco was highly decentralized, and corporate headquarters could not simply impose new ideas on the plants. Their assistant technical director confirmed that "logistics processes vary enormously between sites within Basco and there is in most cases a disconnect between Pell-Q ordering and production planning and possibly even between production planning and sales". Eventually, the team decided to go ahead with SKU rationalization at all sites by year end, but to work on increasing forecasting accuracy on a site-by-site basis. The idea was that consigned stock would be implemented only after forecasting had been improved, to provide Basco with the incentive they said they needed.

3.3.3 Implementation delayed

By December 1994 the team had decided to start with a three-month pilot of SKU rationalization and consignment at a single site at their Antwerp, Belgium, plant; when successful, the project would be rolled out to the others. The pilot site was represented by the local logistics manager (the converted sceptic). To Pellton, it was still unclear exactly what his position was; "certainly in our early meetings with them, it was difficult to tell how much authority he had to implement changes. Our first couple of meetings at the Antwerp plant, we weren't even allowed out of the conference room, I don't know what he was afraid of."

3.3.4 Implementation almost derailed

When the Pellton sales representative responsible for the Basco Antwerp account was pulled onto the project team, he was surprised to find that the plant manager had not been involved so far. In April 1995, just after the pilot had started, he and a Pellton logistics engineer visited the Antwerp plant manager to explain the projects taking place and to discuss the proposal for an EDI link. Though not really enthusiastic, the plant manager agreed to study the project. However, the next day he opposed the EDI link in a meeting with the Basco steering team members. They, in turn, were upset that Pellton had acted without them. The tensions were soothed, but Pellton's sales rep concluded that "we underestimated the complexity of Basco internal communication and politics. In future, we have to try to refer to the Basco Purchasing chemicals purchasing director on everything."

3.3.5 Forecast accuracy was not good enough

The pilot had started on April 1, 1995. The SKU count was duly reduced, and Pellton took on ownership of Pell-Q inventory at the pilot site. Pellton replenished the consigned stock weekly, based on daily consumption data and weekly forecasts for the next three weeks. The agreement was that the consigned stock would cover the Antwerp plant's forecast needs for the next two weeks. When the pilot was evaluated after three months, the level of consigned stock

had not decreased. Actual requirements frequently deviated substantially from the forecasts on which deliveries were based. Pellton staff remembered that they had recognized early on the need to give Basco an incentive to produce good forecasts, but they had let the issue slip. They presumed Basco also needed to gain confidence in Pellton's ability to deliver reliably and responsively before feeling happy with a much lower stock.

3.4 Back to the drawing board

3.4.1 Putting the roll-out on hold

In a joint evaluation meeting in July 1995, Pellton announced they would not implement consigned stock at other sites until pilot site stock had decreased. Upon hearing the current stock levels, the Antwerp plant manager exclaimed "that's ridiculous, we don't need that much", and vowed to bring stock down. The steering team decided to map information flows from Basco Antwerp's customers back to Pellton, to design a better, more integrated planning and forecasting process making use of EDI.

3.4.2 Mapping the current process

To help with EDI, Pellton added an IT project manager to the team. They visited Basco Antwerp and another Basco plant and mapped information flows. The Basco team members (now including the Antwerp plant manager) were remarkably open; the Pellton team learnt a lot about Basco's physical process and planning procedures. Interestingly, during the mapping at the second site, the local logistics manager was amazed when he saw just how backward their own planning systems were. For instance, the planning systems at Basco were largely manual and paper-based; information was aggregated before being sent to Pellton, who would then guess how to disaggregate it for planning deliveries.

3.4.3 Designing the future process

The team met a month later to design a new integrated planning and forecasting system. Pellton would now receive all relevant information in the form of forecasts or updates as quickly and in as much detail as possible, and the Basco sites would get more visibility on Pell-Q stock available to them. Ideally, Pellton would like to see all EDI orders from Basco's customers, but the Basco team members explained that that information would not really help Pellton plan production and would tell them too much about the market.

3.4.4 Preparing for EDI

Soon after, the Pellton IT person returned to Antwerp to meet the local IT staff. Although Basco had sophisticated EDI links with its automotive customers, nothing similar was in place with suppliers, nor did that seem a high priority for the IT staff. Even setting up e-mail connections within the team took a long time, the Pellton IT person having to make up for the lack of IT support within Basco.

Gradually, the role of the IT person extended to overall project manager, keeping track of responsibilities etc., though the logistics manager remained Pellton project sponsor.

3.5 Success at last . . .

3.5.1 Getting joint management approval for the roll-out

By November 1995, stock at Basco Antwerp had decreased to the target level. The project team presented the proposed integrated planning system to joint senior management, who approved implementation at the two selected sites; two months later Basco finally decided to roll the project out to the remaining sites. The latter were generally enthusiastic, no doubt due in part to the strong advocacy by the Antwerp plant manager. The Basco Purchasing managing director took the opportunity to say they were "extremely pleased" with their cooperation with Pellton, that relations between them had improved enormously, and that he now wanted to follow a similar process within other divisions of Basco.

4 Perdirelli: a short-cut to success

4.1 Perdirelli: background

4.1.1 Perdirelli also going through re-engineering

Perdirelli is Pellton's second largest customer, with production sites in Milan, Italy, and elsewhere; the plants are fairly autonomous. Relations between Pellton and Perdirelli, organizational and personal, had generally been relatively good. In early 1995 Perdirelli's Milan plant also embarked on a re-engineering project; senior managers at both firms agreed that would be a good opportunity to jointly evaluate their supply-chain logistics. Aided by Perdirelli Milan's re-engineering consultants and implicitly drawing on Pellton's experience so far with Basco, a project plan was drawn up. Senior managers agreed that, in principle, costs and benefits would be shared 50-50, thus allowing the lower-level design team to focus on logistics issues while leaving commercial matters to others.

4.1.2 Switching to a standard product formulation

Perdirelli Milan took a non-standard formulation of Pell-Q unique to them, adding to production, inventory and logistics complexity for Pellton. In the past, various unsuccessful attempts had been made to switch to Pellton's standard formulation. This project was seen partly as a way to help Perdirelli Milan make the change.

4.2 Mapping the "as-is" supply chain

4.2.1 Getting a design team together

Given the business re-engineering turmoil in Milan, it was not obvious who should be on the design team, as their staff were already stretched to (or beyond) their limits. Eventually, a member of their core re-engineering team was dispatched, and a former plant manager, retired just a few months previously, was recalled to assist during the workshops, though he would not be involved any further. He had been at the plant for over 20 years, and had often been involved in purchasing during that period. Additionally, Perdirelli Milan asked one of their re-engineering consultants to facilitate the joint workshops. For Pellton the choice of team members was easier: the logistics manager who was also the key player in the Basco project, the local sales correspondent, and their plant manager.

4.2.2 Two workshops in quick succession

The project started with two two-day workshops, a week apart, in March 1995. The first, in Milan, kicked off with an outline of the process to be followed for the entire project. The rest of the time was largely devoted to "as-is" mapping of physical and information flows, and a first cut at identifying improvement opportunities. The maps were validated and opportunities evaluated in more detail before the second workshop, in which the "to-be" supply chain was designed and implementation planned; that included designating who, from Perdirelli Milan, would be on the implementation team. The proposals included SKU rationalization (to the same standard sizes as with Basco), consignment stock, with an EDI link to support it, and a new effort to help Perdirelli Milan switch to Pellton's standard formulation.

4.3 A handover, and another handover

4.3.1 Joint management approval followed soon

Two months later the opportunities had been validated and a more detailed implementation plan drawn up. The team presented their work to joint Pellton - Perdirelli Milan management, who approved the proposals. The Perdirelli Milan implementation team included the current production manager as team leader.

4.3.2 Implementation team disintegrated

During the Summer months the re-engineering storm really struck. Many Perdirelli Milan staff, including the entire implementation team except the production manager, were re-assigned. Fortunately their managing director had been instrumental in initiating the project and insisted that it continue; he saw it as an opportunity to learn how to do such projects with other suppliers in future, not as a one-off logistics project. He gave it high priority and visibility within

Perdirelli Milan and made sure sufficient resources were available despite the ongoing re-engineering effort. By September the production manager had selected a new team, consisting of a business planner in a newly formed supply-chain management group, the Pell-Q purchasing manager, a development engineer, and an IT expert. He gave the team a brief introduction to the project, with the documentation and plans from the design team.

4.3.3 The business planner became project manager

The production manager retained responsibility for the project and had the authority to ensure resource availability, but the business planner looked after day-to-day project management and documentation. He started by redefining the original implementation plan to fit the new team. The team members spent a month to get started and to review the opportunities identified by the design team, as they did not always understand the background of certain decisions or plans.

4.4 Back on the road, the project advanced rapidly

4.4.1 The project recovers quickly after a price increase

The Perdirelli Milan team was largely up to speed by October 1995. Meanwhile, Pellton had had their hands full dealing with the failure of the Basco Antwerp pilot. Just as the two sides were to restart the project, Pellton's sales department announced a price increase. Perdirelli Milan was naturally upset and threatened to stop doing business with Pellton; however they insisted that the project should continue regardless. An emergency commercial meeting settled the price issue. In November 1995 the two new teams met for the first time; Pellton had now added the logistics engineer and IT project manager from the Basco project. The meeting was a success and a long list of action points agreed on. For both sides the project would require substantial changes to their IT systems. Implementation was planned for March 1, 1996, starting with a one-month trial period during which the old and new systems would run in parallel.

4.4.2 A joint project manager to speed up the project

By late February it was becoming clear to the Pellton team that they could not meet the deadline for the IT changes. They confessed this to their counterparts in Milan, who admitted to having similar problems. During an emergency meeting Perdirelli Milan's production manager suggested a joint project manager, overseeing developments on both sides, to speed things up. For that role, he proposed Pellton's IT project manager who, as a result, was given direct authority over the IT resources in Milan. A few weeks later than initially planned, the new system was in place and is now operating successfully. Moreover, with Pellton's assistance, Perdirelli Milan has decided to switch to the standard Pell-Q formulation, which would lead to significant mutual benefits, including inventory reduction due to product standardization, and headcount

reductions in purchasing. Perdirelli also want to follow a similar process with their other sites.

5 Analysis of the cases

5.1 What can we learn from these two cases?

5.1.1 Mutual benefits?

Table 1 summarizes the main benefits obtained by Pellton, Basco and Perdirelli Milan. Although their exact magnitude is hard to pin down, all parties involved did consider the projects successful, and worth pursuing with other suppliers and customers. Therefore, the learnings one can extract from the two cases are valuable for future projects. To structure our analysis, we will discuss which problems were encountered, and for each of these why it occurred, how it was fixed, and what key learning this points to.

5.2 Preparation of a joint supply chain improvement (JoSCI) project

5.2.1 Need to agree on a process upfront

Especially in the early stages of the Basco project neither side had a clear idea of how to organize it, which activities to undertake, in which sequence, etc. This occurred because the team had not started out by defining a process to follow. The first step of the Perdirelli Milan project was to jointly agree on a process, which underlies the "roadmap" in the next section. As a result, the Perdirelli Milan project was far smoother (though it failed to deliver commercial success for Pellton). In addition, being able to describe the full process upfront has helped Pellton overcome scepticism in initial discussions with other customers, who were afraid a JoSCI project would be "yet another initiative that would never go anywhere". From this one sees the importance of simply having a process to follow, be it the one proposed here or another such as that in Byrnes and Shapiro (1991) or a re-engineering methodology as eg. that by Kodak (see Institute of Industrial Engineers, 1994) or "Rapid Re" (see Manganelli and Klein, 1994).

5.2.2 How to share the benefits?

At several points early on in the Basco project, team members were more concerned with how proposed changes would affect them, rather than considering the joint benefits. The Basco Antwerp logistics manager, for instance, initially had little incentive to cooperate. This can be explained by the benefit sharing agreement, or rather, the lack of it. The mandate given to the Pellton - Perdirelli Milan design team was explicitly to search for *joint* improvements and to leave commercial issues of sharing costs and benefits to senior management. In principle this would be on a 50-50 basis; though this may

be difficult to realize exactly in practice, it does clearly establish the goal of joint optimization. This allowed Perdirelli Milan to consider making the technical changes necessary to switch to Pellton's standard product formulation, knowing that both sides would benefit. From this one learns the importance of separating logistics and commercial issues and of agreeing on a joint-optimization-oriented sharing rule. Byrnes and Shapiro (1991, p. 21) found that the benefit sharing rule often evolves over time, with the customer initially often taking the lion's share but the vendor standing to gain as more customers adopt the new mode of operating and sales increase.

5.2.3 Functional representation on team

The IT-related projects agreed upon in the second Pellton - Basco workshop eventually petered out; as no IT staff had been involved they had not been thought through carefully enough. With Perdirelli Milan, IT staff from both sides joined the team much earlier. This is especially sensible as IT is frequently an important enabler in redesign (Davenport, 1993). Similarly, the Basco Antwerp plant manager almost killed the project when he was finally informed, but turned into a staunch supporter after having been involved for some time. It was the Pellton sales rep who had recommended his involvement, but he himself had not been pulled onto the team until late. The Basco Antwerp logistics manager who had been involved from the start seemed to lack the authority to agree to any changes. By contrast, the Perdirelli Milan production manager was involved at an early stage and provided strong support throughout. This points out the importance of having the appropriate functions and levels involved early in the project.

5.2.4 Project sponsors

Few resources (such as IT support) were made available within Basco; the Basco Purchasing chemicals purchasing director was their key player, but Pell-Q was by no means his only concern, and he had no direct authority over the production sites. A high-level project sponsor was lacking, in contrast to the Perdirelli Milan case where the managing director and production manager removed resource constraints and carried the project through potentially disrupting periods such as the double handover, Perdirelli Milan's redesign program, and the price increase.

5.3 From mapping to analysis and design to implementation

5.3.1 "As-is" mapping

The failure of the Basco Antwerp pilot was due, in part, to poor understanding (on both sides) of current ordering, forecasting and planning processes. The information mapping then performed was a true revelation for all concerned, and led to removal of many inefficient practices. The Perdirelli Milan project started with mapping exercises, thus avoiding such surprises and resulting time delays

during implementation. The importance of mapping is also recognized by Byrnes and Shapiro (1991) and the Institute of Industrial Engineers (1994). Additional, intangible but critical benefits of mapping lie in team building: it helps members replace typical individualistic perspectives by a system-wide one, a key message in Senge (1990).

5.3.2 Analysis and redesign

In the Basco project the team initially had little idea how to search for improvements. Unrealistic expectations of inventory savings resulted from poor understanding of the true causes of stock. This, in turn, was caused by the team's not knowing which tools to use. The quality control literature offers many suggestions for analysis; for instance, in a later project a fishbone chart (or Ishikawa diagram) was constructed for analysis of root causes of excessive inventories. Though often cast in redesign terms, the JoSCI projects so far were more geared toward continuous improvement, or streamlining the supply chain. Joint (radical) redesign is not impossible, but places far higher requirements on the partnership than the improvement methodology discussed here. Recognizing this distinction is important for selecting the appropriate analysis and design tools (see below) and setting realistic expectations. Careful estimation and measurement of benefits was not performed in either of these cases and has led to problems; Byrnes and Shapiro (1991) point out that this is crucial.

5.3.3 Managing the project

We have seen that progress in the Basco project was slow, deadlines were frequently extended, implementation and roll-out repeatedly delayed. One major contributory factor was the project management style: generally loose, with no tight deadlines or follow-up. The Perdirelli Milan project was managed more tightly throughout, especially when the new implementation team got started. Deadlines were tight (once even slightly too tight), and project managers on both sides followed up on all activities. This resulted in the Perdirelli Milan project being executed much faster (even though it was commercially less successful), which illustrates that appropriate project management is as important here as in any other project.

6 A roadmap for supply-chain streamlining projects

6.1 The roadmap

6.1.1 A roadmap based on key learnings above

Pulling together the critical factors identified in the previous section and combining them with existing literature, we construct the practical roadmap shown in Figure 3, suggesting step by step how JoSCI projects should be managed. The sequence is broadly similar to Byrnes and Shapiro's (1991) "awareness, orientation, implementation" and the Institute of Industrial Engineers' (1994) "project initiation, process understanding, new process design, business transition, change management", but we aim to provide a more detailed and operational roadmap here. The key lessons from the cases relate to the preparation stage, which is what we focus on; of course, project management and implementation are also critical but these are already widely discussed elsewhere.

6.2 Preparation

6.2.1 Team composition

Many factors need to be balanced here:

1. *Team size.* The first workshops with Basco included close to 20 people which was unwieldy and suggests a lack of focus in project scope. The Pellton - Perdirelli Milan team had between two and five people from either side, and worked relatively well.
2. *Functional representation.* Typical functions to include are logistics, sales/purchasing, production, and IT, all preferably from the very beginning of the project.
3. *Knowledge.* Especially during the mapping exercises and subsequent analysis, people with detailed knowledge of current processes need to be involved. Administrative sales and purchasing staff will know order volumes, patterns and procedures; sales reps and purchasing managers may be more aware of organizational issues within the other firm.
4. *Authority.* The project team should include people who can authorize changes suggested by the team, such as the Basco Antwerp pilot site plant manager and the Perdirelli Milan production manager.

5. *Team roles.* Various roles are needed in a team:
   - *Project sponsors,* senior managers to get the necessary resources and to authorize whatever changes are needed.
   - *Team leaders,* with accountability for team performance, with their own authority or invested with that of a sponsor.
   - *Project managers,* responsible for following up and documenting the project plan.
   - *Facilitation* of the workshops was sometimes done by consultants but later by team members. The facilitation role may be separate from the other team roles.
6. *Minimizing handover problems.* In general, three groups of people are concerned: design team, implementation team, and process owners (the people performing the work being redesigned). Ideally, the overlap between these three groups should be maximized. When handovers cannot be avoided, as with Perdirelli Milan, they should be managed carefully to avoid problems of acceptance and understanding, such as the not-invented-here syndrome (Allen, 1977).

Many issues concerning team selection and management are also discussed in Lynch and Werner (1992).

6.3 The first workshop: "as-is" mapping

6.3.1 Setting directions, and mapping

The primary goal of the first workshop is to establish directions for improvements and to map the current supply chain (the "as-is" mapping). A second critical aspect is team-building: it is the first occasion at which the team members from both companies sit together, and soon they should be working as a "one-company" team, jointly searching for opportunities.

6.3.2 Understanding customer needs

A good understanding of customer needs is required to guide the search for opportunities. These will be very different depending on eg. whether the customer competes primarily on low cost or on flexible response. Bowersox and Daugherty (1995) discuss how internal and external logistics structure should depend on the firms' strategic orientation. This is also central to Hammer and Champy's (1993) re-engineering principles: each process has a customer, and should be designed to meet that customer's needs. Note that this analysis should start from the *final* customer's needs, as the process is aimed towards making the entire supply chain meet those needs more effectively.

6.3.3 Supply-chain mapping procedure

Within Pellton, separate maps are typically drawn for physical flows and for planning and information flows. A plant tour helps to visualize the physical process being mapped. A major challenge in mapping is ensuring the right level

of detail throughout; mapping exercises often stay at too high and abstract a level, but time constraints will rule out too much detail. Breaking the process down into key subprocesses (eg. forecasting, production planning, scheduling, order picking, etc.) and then tackling these one by one has worked well. To induce a critical mindset, mapping each process "backwards" can help. Discussions of mapping with examples are given in Lynch and Werner (1992) and Manganelli and Klein (1994). The exact procedure is not as important as capturing the relevant information. As handovers between people often point to opportunities, time-function diagrams are appropriate. During the mapping, one person will actually draw the map (on a large sheet of paper, legible for the entire team); others should be capturing opportunities that come up and other relevant comments. Responsibility to follow up on these should be assigned by the end of the first workshop.

6.4 Analysis and "to-be" process design

6.4.1 Streamlining or redesign?

Probably the most important step in the entire process, this is where the supply chain is actually streamlined or redesigned. In either case, value-added analysis is useful. Three criteria must be met for an activity to be value-added: 1) the customer cares about it; 2) it transforms the product or brings it closer to the point of use; 3) it is done right first time. Lynch and Werner (1992) discuss many practical tools to use in a continuous improvement situation. Redesign, starting from a clean sheet of paper, is potentially more rewarding but also places much higher demands on the customer relationship.

6.4.2 Streamlining: look at re-engineering literature

Before this step, targets should have been set, based on customer needs uncovered earlier. Performance of the current supply chain along those targets and the gap with target levels required should be measured. Key performance indicators should be decided on, to answer the question "What will tell us if a change is an improvement?" The design step should then focus on meeting those needs on a routine basis, and dealing with exceptions or contingencies separately. Points to focus on in streamlining are eg. performing sequential activities in parallel, batching and other causes of inventories, and responsibilities for each process step. Harrington (1991) addresses these issues in more detail.

6.4.3 Redesign: "does this add value?"

Rather than stick to the as-is map and move or delete process steps, true redesign starts with a clean sheet of paper. The challenge is to construct a new supply chain using only the value-added activities; creativity will be a key asset. Ideally, a redesign team should be at least partly different from that during the as-is mapping, to provide the necessary fresh perspective. Though redesign may be

difficult in practice, most principles remain valid for streamlining: a vision of an ideal supply chain is always useful as a target to work towards. Going through a redesign exercise will often bring to light many assumptions underlying the current supply chain which are no longer valid, eg. due to advances in IT. General redesign principles can be found in Hammer and Champy (1993), Davenport (1993) and Harrison and Loch (1996).

6.5 The second workshop: "to-be" mapping

6.5.1 "To-be mapping"

As truly joint "to-be" design will be impractical due to time constraints, the approach taken by Pellton so far is to prepare a proposal for the to-be supply chain, if possible involving the customer, and then discussing it during the second workshop. Important steps are to check whether the to-be supply chain will deliver the target improvements, identifying the implementation team and carefully planning the handover (if any) and quantifying the resources needed.

6.6 Management review

6.6.1 Getting the green light

In both projects the design team had to defer to senior joint management for approval to implement their proposals. Organizing this management review as a presentation by the entire team (not distinguishing between Pellton team members and customer team members) to joint senior management reinforced the team-building effect. Senior management will generally have to decide on how to actually share the benefits, an issue the design team did not address other than by providing information necessary for the discussion.

6.7 Implementation

6.7.1 Change management and project management

Once implementation approval was given the Perdirelli Milan project shifted into a different mode; the Basco project continued to be loosely managed throughout. Clear senior management support will help overcome typical problems of resistance to change. Early involvement of the implementation team should reduce resistance associated with the not-invented-here syndrome (see Allen 1977, and remember the Basco pilot site plant manager); keeping a log of when and why decisions were made will help integrate new team members such as the Pellton sales rep for Basco Antwerp and the new Perdirelli Milan implementation team, and is also an important step in learning for future projects. IT support for e-mail links can enable appropriate communication, both formal and informal, within the team. Change management and project management are extensively discussed in the literature, see e.g. Meredith and Mantel (1985) or Cleland and King (1988).

6.8 Continuous improvement

The roadmap so far provides a way to achieve an initial supply-chain improvement. By opening more communication channels between Pellton and their customers, it should also be a basis for continuous improvement rather than a one-off project.

7 Integration of the JoSCI framework within the organizations

The JoSCI process is clearly time-consuming, and in order to implement it with multiple customers, more than one team is needed. Below we outline the training program developed and used within Pellton to prepare people throughout their world-wide organization to participate in such teams. To oversee this and to facilitate exchange of learnings, a coordination structure is needed. Moreover, the link between logistics improvement and commercial issues as benefit sharing should not be forgotten. The commercial benefits for Pellton from its JoSCI effort with Perdirelli Milan never materialized, despite the project being relatively smooth, due to insufficient integration within Perdirelli's purchasing organization.

7.1 Training

7.1.1 Structure of the training program

To prepare others within Pellton to participate in or manage joint supply-chain improvement projects, a 2 1/2-day training program was developed, based on more detailed versions of the Basco and Perdirelli case studies and the roadmap and supplemental documentation. An additional day can be reserved for facilitation skills training, to prepare participants for being team members or leaders. Participants are drawn primarily from sales, logistics, production, and IT. The training begins with a distribution game (the "beer game", see Senge, 1990), to make participants understand why supply-chain improvement projects are needed, and to serve as a vehicle for discussing the Pellton supply chain. A presentation of how the JoSCI process fits within Pellton's strategy follows. Next is a discussion of the Basco and Perdirelli cases, to bring out a number of logistics issues (eg. SKU rationalization, consignment stock). A first comparison of the projects introduces the key issues in the JoSCI roadmap. Then, after a brief preview of the roadmap itself, its key stages are discussed one by one in participant-facilitated sessions, drawing on the documentation and the two cases; mapping and redesign are done in group exercises. Lectures on topics as supply-chain management, business process re-engineering, time-based competition, etc. are injected where appropriate. The training concludes with a summary of the roadmap itself.

7.1.2 Style of the training program

The guiding philosophy behind the training was to let participants undergo the same learning process as the Pellton logistics manager and we did while developing the roadmap. The participants should not only know the "how", but particularly the "why" behind it. The program is designed to address a variation of learning styles, including case discussions, lectures, games, group exercises, and participant-facilitated discussions. The latter are included as facilitation skills were found to be important; participants practice facilitating supply-chain improvement workshops by facilitating the discussion of the key stages.

7.1.3 First experiences with the training program

Most of the first group of participants had been involved in the Basco or Perdirelli cases; a second training was conducted three months later. Feedback was generally positive (the main criticism was related to the facilitation training), and only minor changes in content were called for. Pellton intends to run the training throughout their organization, approximately six more times world-wide, to support applying the roadmap with many of their larger customers. More recently, a combined training/mapping workshop was held jointly with a customer. In fact, this supply-chain improvement process is now considered a key ingredient of Pellton's competitive strategy.

7.2 Coordination structure

7.2.1 Need for a superstructure

With half a dozen or considerably more supply-chain improvement projects running in parallel world-wide, a more formal coordination structure is needed. This has a number of reasons: selection of customers and ensuring resource availability within Pellton becomes more complex; projects with different customers should not move in conflicting directions; and learnings from projects must be captured and embodied in the process roadmap which, after all, is only based on two cases so far. A critical element of this superstructure will be to implement a performance measurement system to monitor the supply-chain improvements achieved, eg. resulting cost savings and market share increases, but also project performance itself, ie. project duration, resources consumed, etc. Also necessary for capturing learnings is a systematic project audit, as Clark and Wheelwright (1993) propose for new product development projects.

7.3 Reaping the commercial benefits

The JoSCI project with Perdirelli Milan was smooth precisely because it was almost entirely performed at the plant level within Perdirelli. Unfortunately for Pellton, though, when the next round of commercial negotiations started with Perdirelli's central purchasing group, the latter had hardly been involved in the JoSCI project and accordingly placed little value on Pellton's efforts, demanding a price decrease instead. In the event, Pellton was unable to convert its JoSCI investment into increased volume or reduced price pressure with Perdirelli.

With Basco, the opposite pattern emerges: initial progress was slowed because everything was done through the Basco Purchasing purchasing group rather than directly with the plants. However, the result was that Basco Purchasing did see the value of Pellton's JoSCI initiative and honoured this by increasing their volume. The conclusion is clear: for *project success*, direct contact with plant-level staff is essential, but to convert this into *commercial success*, the customer's purchasing group must be involved throughout too.

8 Conclusions

So what does a supplier need to do in order to improve supply-chain efficiency jointly with customers and move closer towards a true partnership with them, obtaining commercial benefits such as increased market share or reduced price pressure from doing so? We have described two projects in which Pellton, a chemical supplier, attempted to improve its relations with major customers. Based on a comparison of those cases, we derived critical factors for such projects, and integrated these into a joint supply-chain improvement framework. This can also be seen as a tentative roadmap of a practical process allowing firms currently in a more arm's-length relationship to jointly improve supply-chain logistics, and to move closer towards a true partnership in doing so.

The roadmap in Figure 3 outlines the critical steps in a supply-chain improvement project. At a deeper level, it simultaneously is designed to start from an arm's-length relationship and to build trust within the team as the project proceeds, moving the companies closer to a true partnership. We found that composition of the project team is critical and non-trivial, and that starting a project with a careful mapping of the current supply chain was important, as a basis for improvement but also for team-building. Also, just having such a roadmap already facilitated projects considerably. The roadmap has been implemented and tested within Pellton with several subsequent JoSCI projects world-wide. We also found that coordinating multiple parallel JoSCI projects and converting suppliers' efforts into commercial benefits requires careful integration of the JoSCI process within the respective organizations.

9 Acknowledgements

We are grateful to Pellton and several of their customers for their openness and cooperation during this research, and to Christian Terwiesch for several helpful suggestions. We would be very interested to hear from others on their experiences with this or similar procedures to supply-chain improvement.

10 References

Allen, T.J. (1977): Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information Within the R&D Organization, MIT Press, Cambridge, Mass.

Anderson, E and B. Weitz (1992): The Use of Pledges to Build and Sustain Commitment in Distribution Channels in: Journal of Marketing Research, Vol. XXIX, February, pp. 18-34.

Bowersox, D.J. and P.J. Daugherty (1995): Logistics Paradigms: The Impact of Information Technology in: Journal of Business Logistics, Vol. 16, No. 1, pp. 65-80.

Byrnes, J.L.S. (1993): Baxter's Stockless System: Redefining the Business (manuscript).

Byrnes, J.L.S. and R.D. Shapiro (1991): Intercompany Operating Ties: Unlocking the Value in Channel Restructuring in: Harvard Business School Report 91-058.

Clark, K.B. and S.C. Wheelwright (1993): Managing New Product and Process Development: Text and Cases in: The Free Press, New York.

Cleland, D.I. and W.R. King (eds.) (1988): Project Management Handbook, Van Nostrand Rheinhold, New York, 2nd ed.

Davenport, T.H. (1993): Process Innovation: Reengineering Work Through Information Technology in: Harvard Business School Press, Boston, Mass.

Hammer, M. and J. Champy (1993): Reengineering the Corporation (A Manifesto for Business Revolution), Nicholas Brealey Publishing, London.

Harrington, H.J. (1991): Business Process Improvement: the Breakthrough Strategy for Total Quality, Productivity, and Competitiveness, McGraw-Hill, New York.

Harrison, J.M. and C.H. Loch (1995): Operations Management and Reengineering, INSEAD working paper.

Institute of Industrial Engineers (1994): Beyond the Basics of Reengineering: Survival Tactics for the '90s.

Lynch, R.F. and T.J. Werner (1992): Continuous Improvement: Teams & Tools, QualTeam, Inc., Atlanta, Ga.

Magrath, A.J. and K.G. Hardy (1994): Building Customer Partnerships in: Business Horizons, January-February, pp. 24-28.

Manganelli, R.L. and M.M. Klein (1994): The Reengineering Handbook: a Step-by-Step Guide to Business Transformation, Amacom, American Management Association, New York.

Meredith, J.R. and S.J. Mantel, Jr. (1985): Project Management: A Managerial Approach, Wiley, New York.

Senge, P.M. (1990): The Fifth Discipline: the Art and Practice of the Learning Organization, Doubleday, New York.

Figure 1. Pellton International Supplies to Automotive Suppliers

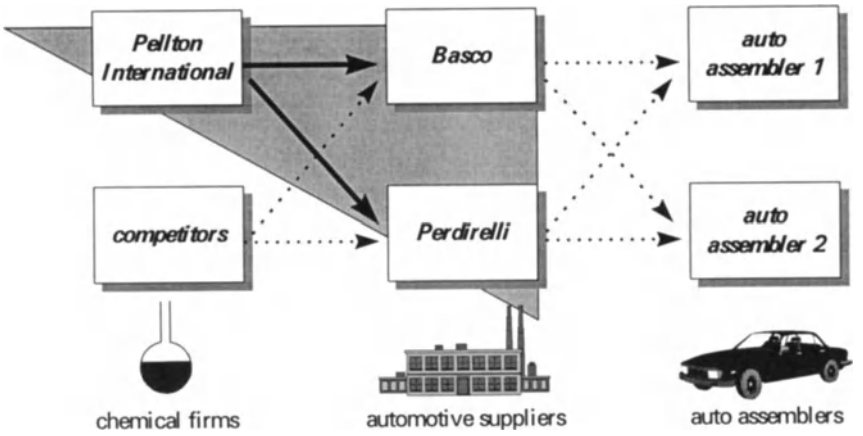


Figure 2. Chronology of Joint Supply-chain Improvement Projects with Basco and Perdirelli

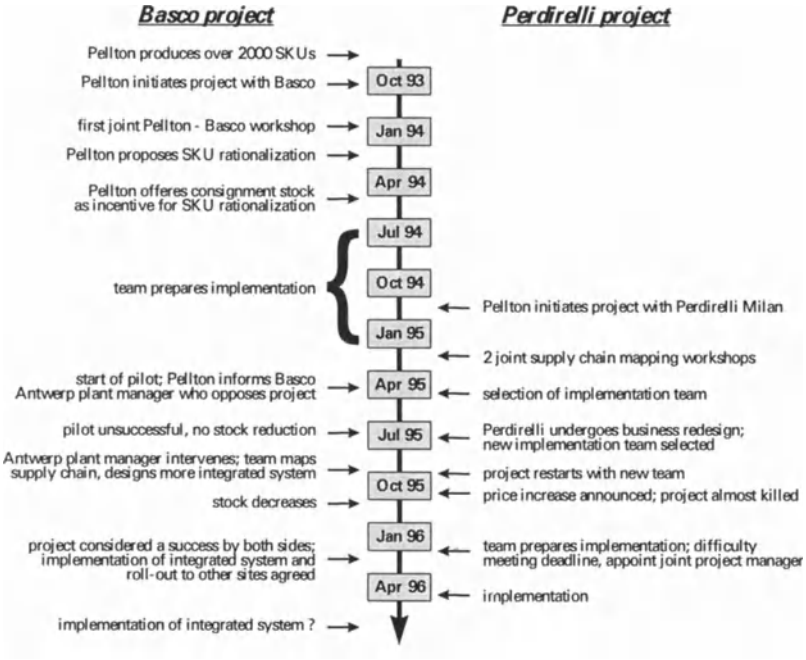


Figure 3. Roadmap for Joint Supply-chain Improvement (JoSCI) Projects

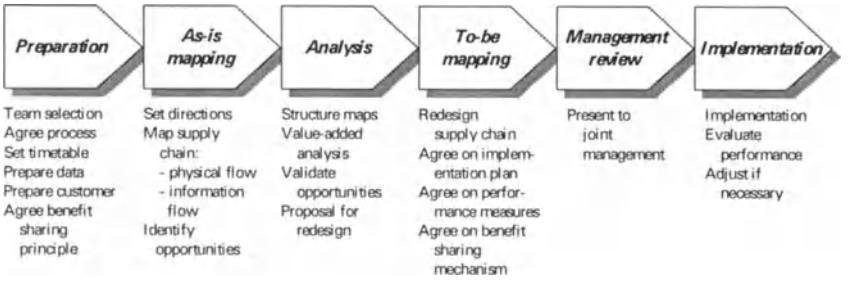


Table 1. Benefits achieved from JoSCI projects.

Pellton International	*Basco Antwerp*	*Perdirelli Milan*
• Radically improved relationship with key customer Basco, and have become preferred supplier • Projected elimination of unique product formulation for Perdirelli Milan, reduced safety stocks and scheduling complexity • SKU rationalization: potential for substantial safety stock reduction (if followed by enough other customers) • Better visibility on demand from Basco Antwerp allows keeping lower safety stock and helps prevent rush orders	• Consignment stock: eliminated millions of dollars of inventory • Potentially more reliable deliveries due to integrated planning and forecasting system	• Partial consignment stock: reduced safety stocks • Reduced headcount in ordering • Learnt about JoSCI process, plan to apply with other suppliers

Design of Freight Traffic Networks

Bernhard Fleischmann

Universität Augsburg, Lehrstuhl für Produktion und Logistik, Universitätsstraße 16,
D-86135 Augsburg

Summary. Most of the industrial freight traffic for the supply of materials and the distribution of products is performed in networks which serve to bundle the shipments and to reduce transport costs. Often these networks are operated by external carriers. We consider various types of such networks and the related design problems, both from the view of a manufacturer and of a carrier. A focus of the paper is on the modelling aspects within a common framework model, in particular on the definition of transport cost functions. Moreover, the recent literature on relevant network design methods is reviewed and two application cases are presented.

Introduction

The industrial freight traffic comprises the supply of materials to a manufacturing firm and the distribution of products from it. However, this distinction of input and output transports is not helpful as a characterization of the conditions of traffic, because the same transport relation can be distribution for one manufacturer and supply for another one. A better distinction is that between the relations from manufacturer to manufacturer (called **"supply"** in the following) and from manufacturer to wholesaler/retailer (**"distribution"**). An additional type of industrial traffic of increasing importance concerns waste disposal, which will not be considered in this paper.

Many transport orders are much too small to justify a separate shipment from origin to destination and therefore have to be bundled in **networks.** This is particularly true for the distribution traffic but also the regular frequent supply of certain materials is performed in networks. The networks are often operated by carriers, or rather logistics service providers, who are, besides the manufacturing and trading business, further important actors in the freight traffic.

The optimal design of networks is also a current question in other areas, such as public transit or telecommunication, where the problems lead to similar models and design methods. The mathematical methods of network design are a subject of intensive research.

This paper focuses on the problem setting and modelling aspects of the design of freight traffic networks. The relevant section of freight traffic and the objectives of the design problem depend on the views of the different actors. A consumer goods manufacturer looks at the total flow of his own products from his factories to his customers. A logistics service provider may combine in his network parts of the

flows of goods of various manufacturers, for supply and distribution. In both cases, the minimization of transport and handling costs is aimed at, whereas inventory is relevant for the manufacturer only. This paper will consider these two views of network design and, moreover, the macro-logistic view of a network of freight traffic centers, where a large number of manufacturers and carriers cooperate. In the design of such a network, ecological objectives are of great importance. Whereas the design of distribution networks has been discussed intensively in literature, the two latter views of freight traffic networks have been hardly addressed. We will define a common framework model for these problems, a non-convex multi-commodity network problem, which is a well known standard model. But the main interest is how to fill out this framework with details. In particular the appropriate cost functions will be discussed. Moreover, some recent algorithmic developments relevant to freight traffic networks are reviewed. Some case reports may help to illustrate the different problem structures.

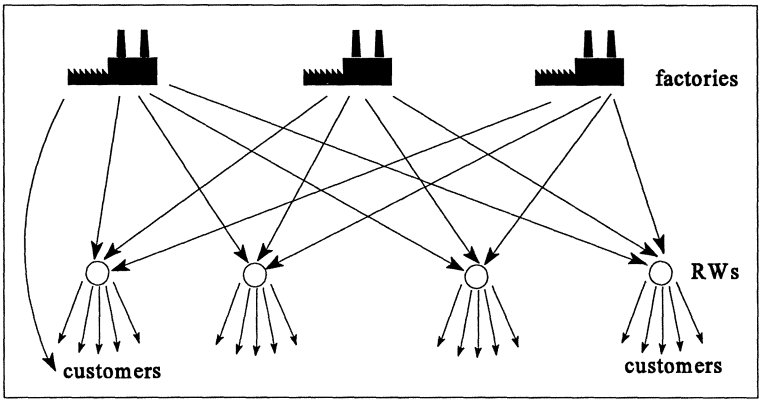
This paper is closely related to other papers in this volume: *Wlcek (1997)* presents new methods for the design of a network of carriers, *Kraus (1997)* investigates the ecological impact of freight traffic networks and *Stumpf (1997)* deals with the short-term operations in a network of carriers.

In the following, Section 1 describes three types of design problems, Section 2 deals with the modelling aspects. Section 3 gives a classified review of algorithms and Section 4 reports on real-life applications for two of the three problem types, while a case study for the third type can be found in *Wlcek (1997)* in this volume.

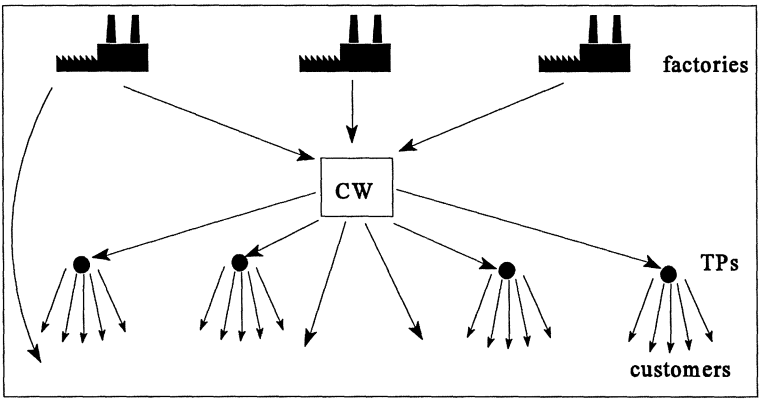
1. Problems

1.1 Distribution networks

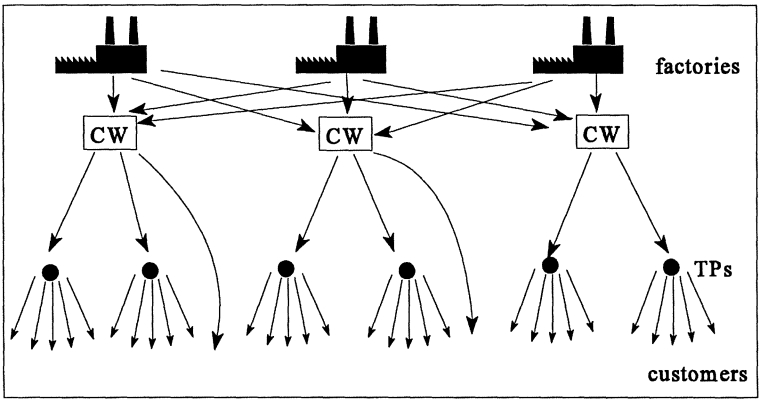
A distribution network comprises the flow of finished products of a consumer goods manufacturer from the factories up to the customers, i.e. wholesalers, retail-ers, department stores, etc. Each factory supplies a certain range of products, where the product programs may overlap between the factories. The orders of the custo-mers contain any mix of the products and vary remarkably in size with a rather low average (100-500 kg in many businesses). Typical network structures are therefore designed to bundle the transports over the long distances and to deliver the small orders in a radius of at most 100km in tours starting from a regional warehouse (RW) or from a stock-less transshipment point (TP). In the latter case, stock of all products is kept in one or several central warehouses (CW), which supply the TPs and have to be replenished from the appropriate factories. In addition, big orders (usually those exeeding one or several tons) are delivered directly from the factories or from a CW to the customers. Figure 1.1 shows basic structures, one with RWs and two transport stages, one with a single CW, TPs and three transport stages and one with a CW at every factory, but every TP supplied by a single CW.



2-stage network with RWs



3-stage network with single CW



3-stage network with CWs at factories

Figure 1.1 Typical Distribution Networks

In practice, any mixture of these structures occurs, but the trend goes to the three-stage CW/TP structures, which allow a centralisation of the stocks to a few points and nevertheless enable short lead times. The TPs are usually operated by external carriers as well as the incoming and outgoing transports (cf. Section 1.2). For a more intensive discussion of distribution structures and related problems see *Fleischmann (1993), Geoffrion/Powers (1995), Hagdorn-van der Meijden/van Nunen (1997)* and *Paraschis (1989).*

The design or redesign of the distribution system of a manufacturer may become necessary for many reasons: e.g. changes in the production sites, in the product allocation or in the customer structure; the switch from own vehicles to an external carrier or from one carrier to another one. Moreover, a recent trend is the cooperation of several manufacturers in a joint distribution system, which has to be designed carefully.

The design problem contains the decisions on the basic structure of the distribution network, the number and location of CWs, RWs, or TPs, and the distribution paths, depending on the products and the order size, or, equivalently, the delivery areas and the assignment of TPs to CWs. The objective is the minimization of costs for transportation, handling, administration and inventory.

1.2 Networks of carriers

In a distribution network, as discussed in the previous section, transports are strictly uni-directional. For an efficient use of the vehicles, however, these transports have to be combined with back freights. In Germany, vehicles owned by a manufacturing company are not allowed to transport goods for third parties. This is one reason for the increasing use of external carriers in the distribution business. In a carrier's network, freights of various manufacturers are combined, both in the same and in the opposite direction. Thus, a manufacturer's distribution network and a carrier's network are not completely different in reality, but overlap more or less. However, they differ in the basic structure. In a distribution network, goods flow from a few origins, the factories, to a great number of destinations, the customers, spread over a large area. In a network of carriers, both the origins and destinations of transport orders, i.e. the sending and receiving customers, are widespread and numerous. Therefore, a distribution network is a uni-directional **few-to-many network**, a network of carriers a bidirectional **many-to-many network**. as defined by *Daganzo (1991).*

A typical supraregional network for less-than-truck load (LTL) of a carrier or of several cooperating carriers consists of several terminals, each with a certain area where goods are picked up and delivered. Transports between different areas have to pass from one terminal to another terminal. Thus, a terminal is a TP linking long distance and short distance transports, like a TP in a distribution system, but both types of transports are going in and out; the short distance transports are called P&D (pickup and delivery) transports. In case of some 30 or 40 terminals (a typical

network size in Germany), the number of terminal-terminal relations is very high and the transport volume on many of these relations is too low for justifying direct transports. Therefore, additional intermediate TPs are required for consolidating and disaggregating orders; they may coincide with the terminal locations. Well-known special structures of this kind are (cf. Figure 1.2) the hub-and-spoke network, the hub-and-tree network and networks with several fully interconnected hubs where each terminal is assigned to exactly one hub. A hub is a special TP which differs from a TP in a distribution system by its functions: Long distance transports from various directions arrive synchronized at the hub, where the goods are sorted, and then the transports go back into the same directions. The system usually includes direct shipments of full and partial loads from origin to destination and the combination of direct shipments with terminal-terminal shipments on the same vehicle tour.

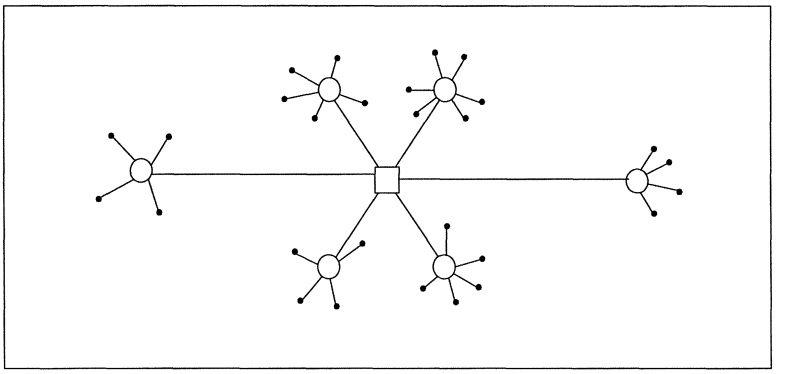
The design of networks of carriers concerns decisions on the basic structure, the number and locations of terminals and hubs, the routes between the terminals and the definiton of the areas. It has to consider the service requirements and the resulting restrictions, e.g. on travel times and on the number of transshipments per freight. The relevant costs include again the costs for transportation and for handling and administration in all locations. The transportation costs must take into account the complete vehicle round trips which may also involve direct shipments of full loads. Various strategic and tactical problems related to the design of many-to-many freight traffic networks are discussed in the recent survey of *Crainic/Laporte (1997).*

1.3 Networks of freight traffic centers

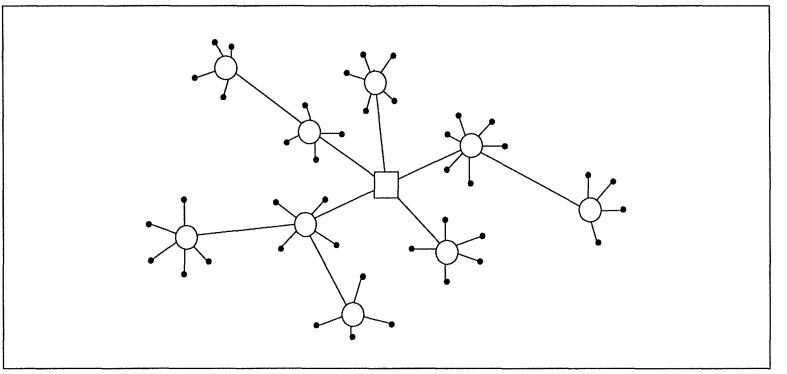
Freight traffic centers (FTC) for metropolitan areas serve as interface between the long distance traffic and the regional P&D traffic and at the same time as interface between different transport modes. Offering various high-capacitated facilities for handling, storing and administrating freight, a FTC is to attract a great number of carriers and to bundle the regional freight traffic. Situated on the periphery, it keeps the city free from heavy vehicle traffic.

Important factors for the location of a FTC are the availability of an appropriate site, the correspondance with the regional development plans, the connection to the road, rail and waterway networks and the acceptance by both potential users and residents. But the key determinant is the intensity of the various flows of traffic meeting the FTC, which do not only depend on the location of the considered FTC, but also on the number and locations of other FTCs in a global network. The design of such a network is a macro-economic decision problem, which is similar to the micro-economic design problem for a carrier network discussed before.

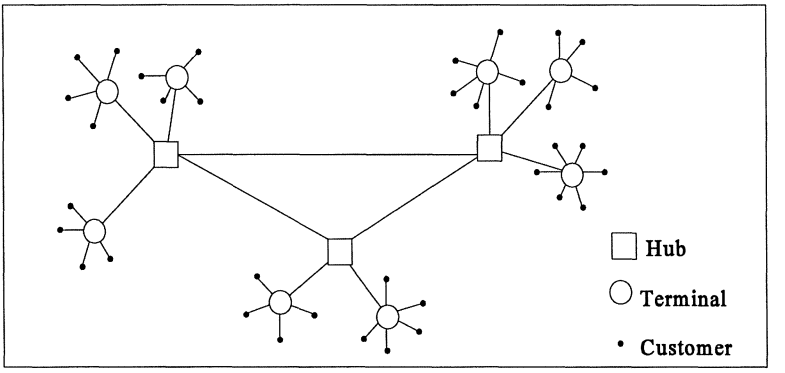
A major obstacle for the use of network flow models in this context is the extreme difficulty of estimating demand data, i.e. data on shipments between the single areas which would use the FTC network. The level of traffic in the network



Hub-and-Spoke-Network



Hub-and-Tree Network



Multi-Hub Network

Figure 1.2 Special structures of networks of carriers

depends on the degree of its acceptance by the potential users and on the proportion of freight suitable for FTCs. Moreover, the total distance travelled in the local pickup and delivery traffic depends on the degree of cooperation of the carriers. The efficient use of FTCs requires a concept of strict cooperation in the sense of "City Logistics", maximizing the load per vehicle and minimizing the number of trips.

In the following we will only consider the design of a local network of several FTCs for a large metropolitan area. (It will be illustrated by a case report in Section 3.2.) In this case, the use of several FTCs, instead of a single giant FTC, reduces the load per FTC, hence the annoyance for the residents, and the distances for the regional pick-up and delivery traffic. On the other hand, it leads to increasing peripheral traffic between the FTCs and to increased investment and operating costs.The network may be supplemented by city terminals, which permit to bundle the traffic between the FTCs and the customers and to shorten the distances of the uban P&D traffic.

Besides costs, the main criteria for the evaluation of a macro-economic network are the degree of annoyance of residential areas and the environmental impact.

2. Models

2.1 A framework model

In this Section, a framework model for the three types of network desing problems considered in Section 1 is presented. Common elements of all those freight traffic networks are transport orders, each to be carried out for a certain quantity of goods from a certain origin to a certain destination. Moreover, there are intermediate locations (warehouses, TPs) where transshipment or warehousing processes take place. By aggregating all transports from the same origin to the same destination in a given planning period, we obtain flows of goods and can make use of the rich theory of network flows. However, when looking at the transport and handling costs, we have to take into account the detailed shipments implied by the original transport orders. The resulting cost functions will be explained in Section 2.2.

First, the traffic in a freight network with given locations is represented by a classical network flow model. The extension of the model to decisions on the (intermediate) locations is explained at the end of this Section. In order to model the consolidation of flows from different origins and the splitting of flows to different destinations, we have to use a **multi-commodity network flow problem (MNFP)**, where the different "commodities" - or simply "products" p - have to show at least the origin of the flow. The destinations can then be expressed by specifying demands for certain products in the corresponding nodes. The following model is a slightly more generalized and more compressed version of the framework model of *Magnanti and Wong (1984)*, where every origin/destination pair is considered as a product. MNFP models with the latter structure are also used by *Amiri/Pirkul (1997)*

and *Balakrishnan/Graves (1989)* for transport optimization in a given network.

The network (A,V) consists of the following elements:

- nodes $i \in V$:
  - **sources** which supply the quantity b_i^p of product p
  - **destinations** which are to receive the quantity $-b_i^p$ of product p $(b_i^p < 0)$
  - **intermediate nodes** with $b_i^p = 0$

- arcs $(i,j) \in A$

  with a cost function $C_{ij}(x)$ for flow x.

In a distribution network, the sources are the factories with a certain range of products. In a network of carriers, the sources are the origins of the transports, and every source node corresponds to exactly one product. The intermediate nodes are warehouses, TPs, hubs, FTCs or P&D terminals.

All quantities (supply, demand, flows) refer to a certain planning period of n days. A node can be at the same time a source for some products and a destination for other products. Parallel arcs are admitted, in particular for modeling different transport modes. (In this case, the notation (i,j) for the arcs is not unique and needs a modification, which we suppress for simplicity).

The MNFP consists of

- the variables:

  $x_{ij}^p \geq 0$ flow of product p through arc (i,j)

- the flow conservation constraints:

  outflow - inflow = b_i^p for every node i and every product p

- the objective:

  minimization of costs.

With the simplifying notations

H	node/arc incidence matrix of (A,V)
x^p	flow vector of product p
b^p	supply/demand vector of product p

the model can be written:

MNFP

$$minimize \sum_{i,j} C_{ij}(\sum_{p} x_{ij}^{p})$$

s. t. $H\, x^p = b^p$, *for every p*

$x^p \geq 0$.

The following special cases permit an easy solution:

- If the cost functions $C_{ij}(.)$ are linear, the problem decomposes into single-commodity linear network flow problems, solvable by standard algorithms.

- If the cost functions are linear and every product p is supplied by a single source, we have a shortest-path problem for every product p from its source to every destination.

In the following, some modelling details and extensions of the MNFP are discussed.

Customer locations

In order to reduce the network size and to get practicable results, it is preferable to aggregate customers which are spread in the plane to **customer locations**, e.g. w.r.t. postal codes. This concerns the destinations in a distribution network and both the origins and destinations in a many-to-many network of carriers. Every customer location then forms a node in the network. In a many-to-many network, this aggregation also reduces the number of "products".

Attention must be paid to the case when a customer location is subject to the restriction of unsplit inflow or outflow (as explained below). This restriction usually concerns the destinations (and origins, for a carrier network) of smaller shipments and forces these shipments to be consolidated at the same intermediate node. Therefore, the destinations (and origins) of the larger shipments, above a certain size limit, which may go directly from origin to destination, have to be aggregated into a separate customer location with split inflow and outflow allowed.

Costs and capacities of a node

Nonlinear costs (and/or capacity restrictions, see below) assigned to a node can be modeled by duplicating this node, one copy receiving the inflow, the other one sending the outflow, and by adding an arc (**throughput arc**) between them representing the throughflow affected with the cost and capacity of the original node.

Unsplit inflow, outflow or route

It is often required that certain nodes must receive their inflow from a single other node ("**single source**" restriction, e.g. for customer locations; for transshipment

points in a distribution network) or send their outflow to a single other node (e.g. sources in a network of carriers). Sometimes the transport between a pair of nodes must follow a **single route,** (e.g. in a non-aggregated carrier network, where the demands represent single orders). These restrictions can only be modeled by means of additional binary variables. The solution space becomes non-convex and has a combinatorial structure.

Capacity restrictions
Separate capacity restrictions per product are easy to consider. Common capacity restrictions for several products are an additional coupling factor between the products, besides the nonlinear objective function, which complicates the solution procedure.

Common cost function for opposite arcs
In a many-to-many network, transports on opposite arcs may be carried out by the same round trip of a vehicle. The additive cost term

$$C_{ij}\,(\sum_{p} x_{ij}^{p}) + C_{ji}\,(\sum_{p} x_{ji}^{p})$$

has then to be replaced by a common function

$$\tilde{C}_{ij}\,(\sum_{p} x_{ij}^{p}\,,\ \sum_{p} x_{ji}^{p}) \quad (i < j).$$

This interdependence of flows on different arcs leads again to a combinatorial extension of the model.

Locational decisions
The design of a freight network involves decisions on the number and locations of intermediate nodes with a certain function. Formally, the decision on the selection of locations from a given set of **potential locations** is already contained in the MNFP: Every potential location is represented by two nodes, linked by a throughput arc, which bears the fixed costs for operating the location and for investment. Nevertheless, most solution procedures differentiate between two or three levels of decisions in MNFP:

- selecting the locations,
- routing in the network,

and, in case of forbidden splits,

- allocation of customer locations to warehouses, terminals etc.

2.2 Cost functions

Freight network design is a strategic planning problem, where it is usually not possible to consider all operations, in particular the use of the vehicles, in detail. But

the costs which are to be considered in strategic planning are strongly influenced by those operations. Therefore, appropriate cost functions are needed which approximate the operational costs. These cost functions represent an interface between the strategic and the operational planning levels.

For the **handling costs** (assigned to the throughput arcs) it is not difficult to define a cost rate per ton, which may decrease with increasing throughput. Also fixed or piecewise fixed costs for administration and investment can be assigned to the throughput arcs.

Inventory costs are only relevant in a distribution network from the viewpoint of the manufacturer. Modelling inventory costs as function of the network flow is complicated, because different types of stock, such as safety stock, stock due to transport units and to order picking, are to be considered (cf. *Fleischmann (1996)*). To our experience the impact of the network structure on the inventory can be well summarized in the number of stock-keeping warehouses. Using this number as a parameter one can determine the best network with w warehouses for various w without considering inventory cost and, in a separate step afterwards, inventory for selected network configurations. Therefore, we do not deal with modelling inventoy cost in this paper. Models for determining inventory levels and allocation in a two-stage distribution system are presented by *Diks/de Kok (1997)* and by *Tüshaus/Wahl (1997)* in this volume.

Transport cost functions are more complicated. As the transport costs depend on the single shipments, the cost function $C_{ij}(x)$ in every arc (i,j) has to be modeled in two steps:

a) derivation of number and quantity of the shipments along (i,j) from the flow x;

b) definition of the cost $T_{ij}(q)$ of a single shipment of quantity q along (i,j).

Step b) can be replaced by using a **tariff** $T_{ij}(q)$ to be paid to an external carrier. Until end of 1993, official tariffs for long distance road transports were obligatory in Germany. This case applies also for rail transports. The use of these tariffs in distribution planning is presented in detail in *Fleischmann (1993)* and in *Paraschis (1989)*. Since the deregulation, however, the strong competition on the transport market forces the carriers to apply cost-oriented tariffs, i.e. tariffs close to the real transportation costs. Various transport cost models incl. standard values for the basic parameters are presented by *Ebner (1997)*.

For modelling the real transportation costs, the following three cases are to be distinguished: A. Arcs not incident with a customer location, B. Pick-up or delivery arcs, i.e. the regional transports to (and from) the customers, and C. Direct delivery arcs, i.e. long distance transports to (and from) the customers.

A. Arcs (i,j) not incident with a customer location

Transports between factories, warehouses and other intermediate nodes are assumed to occur regularly. Let

n	length of the planning period in days
n_{ij}	minimal number of shipments required in the planning period (due to service restrictions or perishable goods), $1 \leq n_{ij} \leq n$;
Q_{ij}	full load of a vehicle;
c_{ij}	cost of a single shipment.

Then we get in step a) the quantity per shipment

$$q = min\ (x/n_{ij}\ ,\ Q_{ij}\ )$$

and the number of shipments

$$max(x/Q_{ij}\ ,\ n_{ij}\ )$$

and in step b)

$$T_{ij}\ (q) = c_{ij} \quad for \quad 0 < q \leq Q_{ij}$$

and hence

$$C_{ij}\ (x) = c_{ij}\ max(x/Q_{ij}\ ,\ n_{ij}).$$

This assumes that partial shipments can be combined over several days, e.g. if shipments occur daily with q equal to 1½ loads, one can ship alternately 2 and 1 full loads per day. This assumption usually holds if node j is a warehouse in a distribution system. Note that in this case $C_{ij}\ (x)$ is linear for $x \geq Q_{ij}\ n_{ij}$ which is usually true for the transports between factories and central warehouses.

However, for the transports to transshipment points and for the regular (daily) line services in a carrier network, the above assumption does not apply, and we have (with $n_{ij} = n$)

$$C_{ij}\ (x) = c_{ij}\ n \left\lceil \frac{x}{nQ_{ij}} \right\rceil ,$$

which is a non-concave staircase function. Another possibility for transports to transshipment points in a distribution network is the combination with direct delivery tours (see case C below).

B. Pick-up or delivery arcs (i,j)

For deliveries and pick-ups the number and size of the shipments is determined by the orders of the customers. For designing the network, they are usually taken from a characteristic period of the past and are either used as they are - step a) is then

omitted - or aggregated in step a) , e.g. into order size classes (cf. *Fleischmann (1993)).*

But step b) is difficult, as deliveries and pick-ups are carried out in vehicle tours combining many orders in various customer locations. Thus, there is no direct shipment in (i,j) in reality, and the cost function $T_{ij}(q)$ has to estimate the contribution of delivering the quantity q in location j to the total cost of the tour from depot i.

By the way, the same kind of function is relevant for a carrier who calculates the prices that single customers have to pay for pick-ups and deliveries. This is a topical problem in Germany after the deregulation of tariffs.

We use the ring model, explained in the following, for defining a function $T(d,q)$ for the cost of delivering or picking up an order of size q at a location with distance d from the depot and set $T_{ij}(q) = T(d_{ij}, q)$ where d_{ij} is the length of arc (i,j) (cf. *Fleischmann (1979)*). An alternative model for the same purpose is proposed by *Tüshaus/Wittmann (1997).*

Ring Model

The following parameters are needed which may vary from region to region:

- v speed on the way to the first customer and from the last customer to the depot (may depend on d).
- v_0 speed between the customers
- d_0 average distance between two consecutive customers
- s average stop time per customer (may depend on q)
- Q vehicle capacity
- Z maximal duration of a tour
- c_1, c_2, c_3 cost per km, per hour and per tour, resp.

The model considers a certain order with distance d and size $q \leq Q$ and assumes that it is delivered (or picked up) on a tour containing only orders with the same properties; i.e. the customers are situated on a ring with radius d around the depot. Considering both the capacity and the time limit of a tour, the maximum number of orders in the tour is

$$k = min\left(\frac{Q}{q}, \frac{Z - 2d/v + d_0/v_0}{s + d_0/v_0}\right).$$

The cost per order is 1/k of the total cost of the tour, hence:

$$T(d,q) = \frac{1}{k}[c_1(2d + (k-1)d_0) + c_2(2d/v + (k-1)d_0/v_0 + ks) + c_3].$$

Note that the model estimates at once the length and the duration of the tour (the factors after c_1 and c_2, resp.). It can also be applied for a fleet of vehicles with different sizes Q_r $(r \in R)$ by defining $\tilde{T}(d, q, Q_r)$ for every vehicle size Q_r and setting

$T(d,q) = min \{ \tilde{T}(d, q, Q_r) : r \in R \}.$

The model tends to overestimate the distance traveled from the depot to the first customer of a tour and from the last customer to the depot: Its average in the model is twice the average customer-depot distance over all customers, whereas in reality, the customers nearer to the depot are preferred as first or last customers in a tour. A better approximation is usually achieved if the term 2d in the above formulae is replaced by 2df, with an appropriate factor f between 0.7 and 1.0.

The model has been tested by comparison with the results of a vehicle scheduling algorithm using real data with 50 - 120 customers. For the sum of all tours of a day from one depot, the average (maximum) absolute deviation was 12.7 (23%) for the length, 7.9 (16%) for the number of tours and 7.4 (15%) for the costs. It has been used in several distribution planning studies and showed a good fit with actual data for the present situation.

C. Direct delivery arcs (i,j)

In most freight networks, orders which exceed a size limit, say 1 or 2 tons, are shipped directly from the origin (or a central warehouse) to the destination. They may be combined to tours, if smaller than full loads, but those tours contain only 2, 3 or at most 4 deliveries and, in case of a carrier network, the same number of pick-ups. The composition of the tours is mainly restricted by the vehicle capacity. In a distribution network, transports to transshipment points are often combined with direct delivery tours, so that the cost function derived in the following also applies to this case, but q determined as in Case A. The structure of these trunk haulage tours is quite different from the regional P&D tours, which may contain some 10-20 deliveries and are mainly restricted by time. Therefore, the ring model is not appropriate in this case. Models and methods for short term scheduling of trunk haulage tours are presented by *Stumpf (1997)*. An approximation can be derived from a sample S of direct shipments with sizes q_s ($s = 1,..,n_0$) in the following way:

With the notations

Q	size of vehicles
$\bar{Q}$	observed average load of a direct delivery tour ($\bar{Q} < Q$)
q_0	minimum order size for direct deliveries
$\bar{q}$	mean of the sample
n_0	number of orders in the sample
n_t	average number of orders per tour = $\bar{Q} \ / \ \bar{q}$
c_{ij}	cost of shipping a full load along (i,j)
$\alpha(q)$	proportion of the cost of a full load assigned to an order of size q (parameter function)

the cost function is

$$T_{ij}(q) = \alpha(q)\, c_{ij}.$$

$\alpha(q)$ has to satisfy the conditions

$$q/Q \le \alpha(q) \le 1 \text{ and } \alpha(q) = 1 \text{ for } Q - q_0 \le q \le Q.$$

If all direct deliveries occur on the same arc (i,j), the condition

$$\bar{\alpha} := \frac{1}{n_0} \sum_s \alpha(q_s) = \frac{1}{n_t}$$

ensures the correct total cost for the sample. For a linear function $\alpha(q)$ this would imply $\alpha(q) = q / \bar{Q}$ which contradicts $\alpha(q) \le 1$ *for* $q > \bar{Q}$ (cf. Figure 2.1). Considering the detour caused by combining orders to different destinations in a tour, $\bar{\alpha}$ must be even greater, say $\bar{\alpha} = \beta/n_t$ with a detour factor $\beta > 1$. Thus $\alpha(q)$ must be a concave function. This is the reason why any cost-oriented trunk haulage tariff $T_{ij}(q)$ must be nonlinear-degressive.

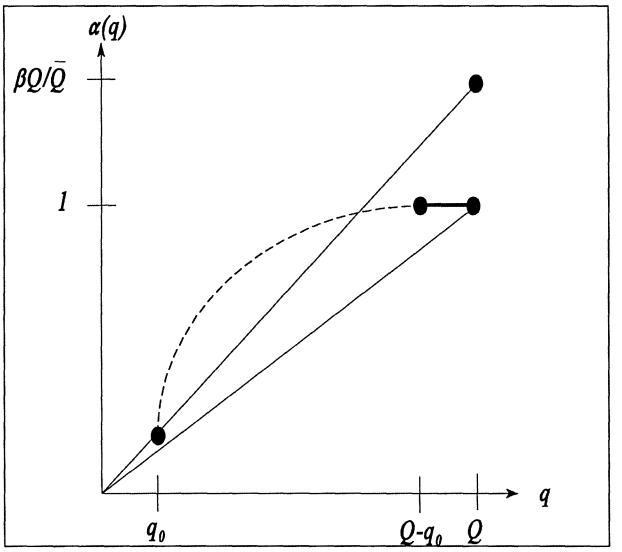


Figure 2.1
Different forms of the function α (q)

An appropriate form for $\alpha(q)$ is a quadratic function, where the three parameters are easy to determine by the equations $\alpha(Q-q_0) = 1$, $\alpha'(Q-q_0) = 0$ and a given $\bar{\alpha}$. Several variants of this function are studied by *Kraus (1997)* (in this volume) and validated with real data.

3. Methods

3.1 Overview

The MNFP models considered before differ from a simple linear model by one or several of the following complicating properties:

A. Non-convex cost functions
B. Non-split constraints
C. Locational decisions.

Case C can be considered as a special case of A, as explained in Section 2.1, but most algorithms for A are not designed for this case. Without any one of the properties A, B, C, the MNFP is a special linear program, for which efficient optimization algorithms are available. We do not consider pure facility location problems (only property C) like the Warehouse Location Problem (WLP) and related problems. WLP type models with linear costs are not appropriate for freight transportation networks, as the optimal number of warehouses, terminals, etc., is mainly influenced by the economies of scale in the transportation cost. The locational decisions, however, can often be taken by comparing a limited number of alternatives in practice.

Depending on the properties A, B, C of the problem, the following algorithmic principles are mainly used. Except for the first one, these are heuristic principles:

- **Exact algorithms**, enumerating the tree structure of optimal flows by means of Dynamic Programming or Branch and Bound (for A and A+C, concave costs, single commodity); these algorithms seem to be practicable for small problems only *(cf. Magnanti/Wong, 1984).*

- **Benders Decomposition** (for B+C) *(cf. Magnanti/Wong, (1984))*

- **Column Generation** (for A+B), where the "columns" corespond to routes *(cf. Barnhart et al. (1996))*

- **Local linearization** (mainly for A)

- **Iterative Decomposition** of the problem into
  - Routing and allocating the non-split-nodes (**Route-Alloc**, for A+B)
  - Fixing the locations and allocating (**Loc-Alloc**, for B+C)

- **Lagrangean Relaxation and/or Dual Ascent** (for A and A+B)

- **Local Search** (for all types).

In the following we shortly review some recent algorithms based on local linearization (Section 3.2), Decomposition (Section 3.3), Lagrangean relaxation (Section 3.4) and Local Search (Section 3.5). Earlier reviews are given by *Magnanti/Wong (1984), Minoux (1989)* and *Paraschis (1989).*

3.2 Local linearization

Local linearization, as first suggested by *Yaged (1971)*, consists in alternately fixing cost rates for every arc and solving the linear problem with these cost rates (cf. Figure 3.1). For concave differentiable costs, it converges to a local optimum, if the marginal costs of the current solution are used as cost rates for the next iteration. Its advantage is the ease of implementation and the fast convergence in a few iterations. According to our experience, this still holds for general cost functions. The disadvantage is the sometimes poor quality of the solution, which strongly depends on the starting cost rates. Additional improvement steps are therefore advisable.

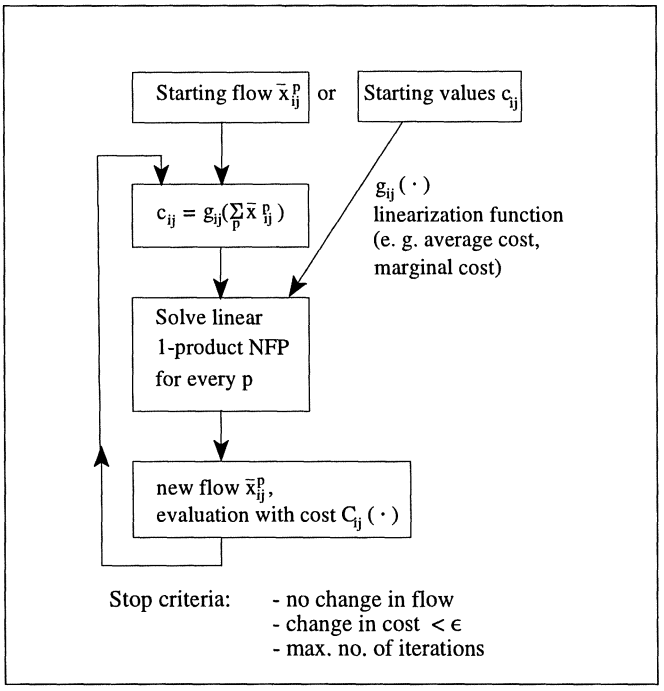


Figure 3.1
Solving MNFP by local linearization

Paraschis (1989) has investigated various types of linearization functions for the general case. For uncapacitated networks, it can be extended to single-source restrictions in the destinations (cf. *Fleischmann (1993)).*

We have used local linearization in many real-life distribution studies for

networks with up to 100 factory and warehouse nodes, 1500 customer locations, 25000 arcs and 20 products and realized CPU times of a few minutes on a PC.

3.3 Decomposition

Decomposition can be used for separating the assignment of no-split-nodes to other nodes from the remaining routing and/or location problem. In this remaining problem the no-split-nodes can be neglected, their supply or demand is aggregated to other nodes by the assignment.

In a few-to-many network, the assignment problem is trivial, if there are no capacity restrictions; otherwise it is a general assignment problem (GAP). The same holds for the assignment of node pairs to a single route in any network.

In a many-to-many network, however, except for the one-hub structure, the node assignment problem is quadratic: For instance, for

i,j	no-split customer locations
w_{ij}	the quantity to be shipped from i to j
c_{kl}	the cost rate for transports from terminal k to terminal l, determined by the routing (approximately, if nonlinear costs)
y_{ik}	the assignment variable of location i to terminal k

the cost of the assignment is

$$\sum_{i,j,k,l} c_{kl}\ w_{ij}\ y_{ik}\ y_{jl}$$

like in the Quadratic Assignment Problem (QAP). This problem is very difficult on its own and is usually solved by heuristics, similar to the QAP heuristics.

The remaining routing and/or location problem is reduced in size and without no-split constraints. If the location problem is solved by exchange heuristics or by branching, the overall procedure alternates between changing the locations and solving the assignment problem.

Algorithms of this type are proposed by *Klincewicz (1991)* for a multi-hub network with no-split terminals and by *Aykin (1993)* for a capacitated multi-hub network with single-route restrictions. Both models contain decisions on the location of hubs, but linear costs. *Leung/Magnanti/Singhal (1990)* consider a given general many-to-many network with nonlinear costs, capacity restrictions and single-route restrictions.

3.4 Lagrangean relaxation and dual ascent

As opposed to other heuristic procedures, the Lagrangean relaxation and the related dual ascent technique provide lower bounds (LB) for the cost of the optimal solution. They are mostly used for solving the piecewise-linear MNFP without no-

split constraints. By relaxing the flow conservation constraints, the problem decomposes into single-arc subproblems, which are easy to solve. Both Lagrangean relaxation and dual ascent yield values of the dual variables and modified cost coefficients, which are then used for finding heuristic solutions.

The most general case of a many-to-many network is considered by *Balakrishnan/Graves (1989).* For networks with up to 60 nodes, 60 commodities and 360 arcs, their algorithm yields an average gap between LB and heuristic solution of 1.7%; the CPU times are rather high, but reduce to less than 10 minutes in case of "layered networks", where every route contains at most two transshipments. These results have still been improved by *Amiri/Pirkul (1997)* using a slightly modified relaxation. *Leung/Magnanti/Singhal (1990)* solve routing problems in real networks with 210 terminals, 40 distribution centers and 2000 arcs and detailed vehicle cost functions in about 1 hour on a Prime 850 (in combination with a decomposition for single-route constraints, cf. Section 3.3). *Larsson/Migdalas/Rönnqvist (1994)* consider various relaxations for the single-commodity case with concave costs. *Crainic/Delorme (1993)* consider many-to-many networks for the transport of empty containers with customers and terminals, locational decisions on the terminals, but linear costs; they solve, by dual ascent, problems with about 250 nodes, 44 potential locations, 5000 arcs and 12 commodities in less than 1 minute CPU on SUN Sparc and yield gaps of 1-3% in most cases.

3.5 Local search

Local search algorithms try to improve a solution by exploring neighbourhood solutions. Classical **descent algorithms** admitting only strict improvement steps are presented for the concave MNFP by *Gallo/Sodini (1979)*, who define neigborhood as adjacency of extremal flows, and by *Minoux (1989)* who redirects the flow of a single arc through an appropriate path.

Modern search strategies admitting temporarily increasing costs can be applied to all decision levels of network design - location, routing, assignment - with all complicating properties A, B, C of the MNFP. The difficulty is, that the cost evaluation of given locations requires the solution of a (non-linear) MNFP, and the evaluation of a given routing requires the solution of a (quadratic or general) assignment problem. However, for the search in the neighbourhood of the current solution, it would be much too expansive to solve the problems of the lower levels for every neighbour. Instead, simple cost approximations have to be used for the evaluation. After selection of a new solution, the lower level decisions have to be adapted, possibly again by local search.

Tabu search is used by *Crainic et al. (1993)* and *Skorin-Kapov/Skorin-Kapov (1994)* for a multi-hub network with linear costs and locational decisions on the hubs, and by *Bazlamacci/Hindi (1996)* for the concave MFNP, based on the search of extreme flows of Gallo/Sodini. The latter paper reports on the solution of

randomly generated problems with up to 48 nodes, 174 arcs and 5 commodities in less than 60 seconds and a remarkable improvement of the solutions (up to 14%) compared with the pure descent algorithm. *Wlcek (1997)* (in this volume) applies the **Deluge search** scheme to the design of multi-hub networks of carriers including all problems A,B and C.

4. Applications

4.1 Distribution system design

We have implemented the local linearization algorithm (cf. Section 3.2) in a decision support system, DISI, for simulating, evaluating and optimizing a distribution system for consumer goods. It is written in Fortran and runs on a PC (cf. *Fleischmann (1993)*).

The underlying model allows up to 3 transport stages from the factories to the customers and arbitrary cost functions for transportation on every relation and for handling in every factory, warehouse or transshipment point. The system contains 1500 customer locations, based on postal codes and geographical coordinates, and aggregates automatically given customer data into locations. Distances are calculated as Euclidian distances times a street factor, using the coordinates, or can be specified exactly. Potential locations for warehouses and transshipment points can be opened or closed in dialogue, and the optimal flow is calculated for given locations. The resulting quantities and costs can be shown in a more or less detailed form, as desired by the users, and graphically in network diagrams.

DISI has been used in about 30 cases in industry (food, tobacco, electronics, domestic appliances, carriers). Typical questions to be answered by distribution studies are:

- adaptation of the distribution system to changes in demand structure or distribution area
- decision on switching from own vehicles to external carriers, evaluating offers of carriers;
- effects of changes in factory locations and allocation of products;
- potential savings by cooperation of several producers, design of a joint distribution system
- supporting the calculation of new tariffs of a carrier.

In the following we report on a recent case from the food industry .

Present situation

The company has two factories, one in the North of Germany producing all products and 83% of the tonnage, one in the South producing only a part of the product

program and 17% of the tonnage. All orders exceeding 10 pallets are shipped directly from the factories to the customers, this concerns 15% of the orders and 57% of the tonnage; the orders up to 10 pallets are shipped via 16 RWs, possibly with an additional transshipment at one of 7 TPs.

Objectives of the study

The main question was the optimal number of warehouses and TPs, as the business has changed remarkably in volume and structure since the definition of the present system. Moreover, the optimal order size limit for the direct deliveries should be determined. Another idea to be investigated was to ship the total production of the southern factory first to the northern factory warehouse in order to avoid the rather small shipment sizes from the southern factory to the RW and to the customers.

Procedure

The study was based on the data of 153.000 orders of the last year. As the current fixed-rate transportation prices of the carriers are not appropriate to show effects of changes in the distribution system, they were replaced by cost-oriented tariffs, as described in Section 2.2. Further relevant costs are the handling costs in the factory warehouses, RWs and TPs and fixed costs in the RWs. The only inventory kept in the RWs are the started pallets for order picking, which can be included in the fixed costs, whereas the safety stock kept in the factory warehouses is not influenced by the considered variations of the distribution system.

First the present distribution system, including the assignment of customers to RWs and TPs, i.e. the delivery areas, is evaluated by its total cost. In a second step, the present network is still kept and only the delivery areas are now optimized, yielding a **reference model** for the structural changes. This way, the effects of a pure change of delivery areas and of changes in the RW and TP locations can be distinguished. Using DISI, starting from 110 potential locations, about 25 configurations with 2 - 10 RWs have been determined, 5 of which have been selected by the project team and presented to the board. Moreover, the effects of changes in the limit for direct deliveries and of the centralized supply were evaluated.

Results

Table 4.1 summarizes the results. The figures are costs per year, normalized so that the total cost without direct deliveries in the present system is 1000. The optimization of delivery areas saves 2.7% of these costs, the selected configurations, denoted A to E, up to 10% in addition. The costs proved to be rather insensitive against changes in the direct delivery limit, the best limit turned out to be 8 pallets. The central supply structure enables more direct deliveries and reduces the delivery costs slightly, but the cost of the additional transport between the factories make it disadvantegeous. It was decided finally, to implement configuration C.

Table 4.1 Results of the distribution study.

	Present configuration		Configurations					Reference Model with	
	Present delivery areas	opt. areas (**Reference Model**)	A	B	C	D	E	direct delivery from 8 Pal.	central supply
# RW	16	16	8	8	6	4	2	16	16
Costs # TP	7	7	14	20	21	26	22	7	7
RW	188	188	125	124	97	65	32	178	186
factory → RW/TP	256	251	247	260	277	314	355	230	194
RW/TP → customers	556	534	548	506	506	498	522	500	516
$\sum$ deliveries via RW	1000	973	920	890	880	877	909	908	896
direct deliveries	477	477	like in present configuration					530	540
factory S → factory N									107
Total	1477	1450						1438	1543

4.2 Design of carrier networks

In the scope of a common research project of the Universities of Augsburg and Nürnberg and several SME carriers the decision support system BOSS for designing and evaluating LTL networks has been developped (cf. *Hemming et al. (1996), Wlcek (1997)*). It contains a MNFP model for a general many-to-many network, various vehicle cost models, as described in Section 2.2 and uses local search heuristics (cf. Section 3.5). In addition, the main characteristic values of the ecological impact are calculated for every solution, based on the resulting mileage, transport mode and type of vehicles (cf. *Kraus (1997)*). A pilot application is described by *Wlcek (1997)* in this volume.

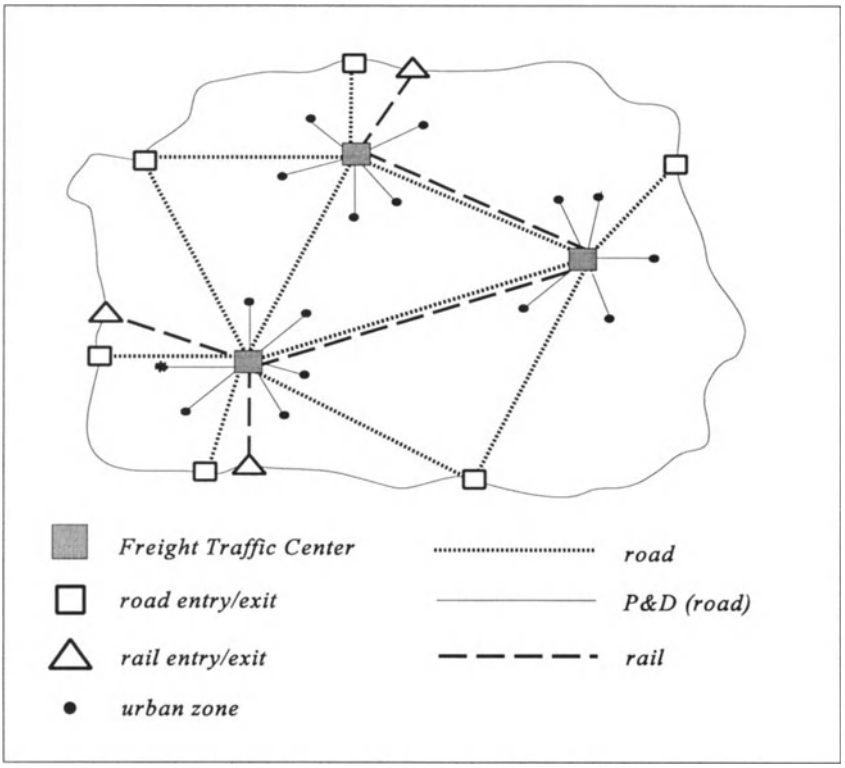


Figure 4.1 Freight Traffic Centers for a metropolitan area

4.3 Location of freight traffic centers for the Berlin area

After the German unification, the city of Berlin charged a consortium of consultants and research institutes to elaborate a concept of freight traffic centers (FTC) for the

Berlin area. One task was to evaluate various locations of FTCs w.r.t. the expected traffic flow. The following MNFP model was used (cf. Figure 4.1):

Nodes

- 10 points where main road and railway lines enter the considered area (entry/exit points, EEP),
- 1 to 4 FTCs at 12 potential locations (outside the city),
- in some scenarios 1 to 3 additional city terminals at 8 potential locations,
- 764 urban zones (destinations).

Arcs

- from EEPs to FTCs (rail or road, depending on type of EEP),
- from FTCs to FTCs and to city terminals (rail and road),
- from FTCs and city terminals to destinations (road).

Assumptions

1. There is an ideal cooperation of the carriers using the FTCs so that the total freight on a certain relation is transported by the minimum number of vehicles required.
2. Transports between a certain EEP and a certain zone follow the same route in both directions.

These assumptions allow to reduce the bidirectional freight network to a unidirectional model: For every arc, the maximum of the flows in either direction is considered, because it determines the number of vehicle round-trips required. The EEPs are the sources, the zones are the destinations with demand equal to the maximum of the arriving and the leaving freight.

Demand data

Estimations of the freight tonnage arriving and leaving were available per zone and per day of the week as well as the percentage of the total freight per EEP. It was assumed that these percentages were the same for the freight shipped between the EEPs and every zone. Thinking of the freight from the 10 EEPs at 5 days of the week as 50 different "products", the demand per zone and product could be specified.

Restrictions

All locations for FTCs and city terminals have limited capacities depending on the available sites.

The rail EEPs are subject to an unsplit outflow constraint, because trains cannot be split at the EEP, but only at the FTC. For the road EEPs different scenarios with and without no-split constraints were considered. For a split flow from one EEP to several FTCs the truck loads would need to be presorted.

Costs
The objective was to minimize the total mileage of the vehicles. Hence the mileage per arc is used as cost of the arc. For

d_{ij} length of arc (i,j)
v_{ij} vehicle size along (i,j)

the cost function is

$$C_{ij}(x) = d_{ij}\ x/v_{ij}$$

As the flows are large compared with vehicle sizes, rounding up the number of vehicles to an integer is not relevant, hence the cost function is linear.

Vehicle sizes are 19.2 tons for the road traffic between EEP and terminals and 2.7 tons for the P&D traffic to the zones. For the rail arcs, the ecological advantage against trucks was estimated by the factor 3.6, so that the "vehicle size" is 3.6*19.2 on rail arcs.

Procedure
The distribution planning system DISI (cf. Section 3.1) was used to calculate the optimal flow for about 40 configurations of FTC and city terminal locations, under varying assumptions on the use of rail traffic and the possibility of splitting flows at the EEP.

Results
The optimal number of FTCs (w.r.t. minimal vehicle mileage) turned out to be 3 to 4. The use of one additional city terminal reduces the traffic by up to 18%, the use of three city terminals by up to 25%. Meanwhile, a decision has been taken for 3 FTC locations, which differ from the best configuration in one location and by about 5% more vehicle mileage.

References

Amiri, A./ Pirkul, H. (1997): New formulation and relaxation to solve a concave-cost network flow problem. In: Journal of the Operational Research Society 48, 278-287

Aykin T. (1994): Lagrangian relaxation based approaches to capacited hub-and-spoke network design problem. In: European Journal of Operational Research 79, 501-523

Balakrishnan A./ Graves S. C. (1989): A Composite Algorithm for a Concave-Cost Network Flow Problem. In: Networks 19, 175-202

Barnhart C./ Hane C.A./ Vance P.H. (1996): Integer Multicommodity Flow Problems. Working Paper, MIT

Bazlamacci C.F./ Hindi K.F. (1996): Enhanced Adjacent Extreme-point Search and Tabu Search for the Minimum concave-cost Uncapacitated Transshipment Problem. In: Journal of the Operational Research Society 47, 1150-1165

Crainic T.G./ Delorme L. (1993): Dual-ascent procedures for multicommodity location-allocation problems with balancing requirements. In: Transportation Science 27, 90-101

Crainic T.G./ Gendreau M./ Soriano P./ Toulouse M. (1993): A tabu search procedure for multicommodity location/allocation with balancing requirements. In: Annals of Operations Research 41, 359-383

Crainic, T.G./ Laporte, G. (1997): Planning models for freight transportation. In: European Journal of Operational Research 97, 409-438

Daganzo C.F. (1996): Logistics Systems Analysis. 2nd ed. (Springer) Berlin et al.

Diks, E.B./ de Kok, A.G. (1997): Transshipments in a divergent 2-echelon distribution system. In this volume

Ebner, G. (1997): Controlling komplexer Logistiknetzwerke. Doctoral dissertation, University Nürnberg. GVB-Schriftreihe 34

Fleischmann B. (1979): Distributionsplanung. In: K.-W. Gaede et al. (eds) Proceedings in Operations Research 8. (Physica) Würzburg, 293-308

Fleischmann B. (1993): Designing distribution systems with transport economies of scale. In: European Journal of Operational Research 70, 31-42

Fleischmann B. (1996): Management of finished products inventory in the consumer goods industry. Working paper, Universität Augsburg

Geoffrion, A.M./ Powers, R.F. (1995): Twenty years of strategic distribution systems design: an evolutionary perspective. In: Interfaces 25, 105-127

Hagdorn-van der Meijden, L./ van Nunen, J. (1997): Strategic Decision making for logistics network design. In this volume

Hemming H./ Ebner G./ Kraus S./ Wlcek H. (1996): Kosten- und Umweltorientierte Optimierung von Güterverkehrsnetzen. Bericht zum AIF-Projekt Nr. 9767, Universität Nürnberg

Gallo G./ Sodini C. (1979): Adjacent extreme flows and application to minimum concave cost flow problems. In: Networks 9, 95-121

Kliencewicz J. G. (1991): Heuristics for the p-hub location problem. In: European Journal of Operational Research 53, 25-37

Kraus S. (1997): Estimating the length of trunk tours for environmental and cost evaluation of distribution systems. In this volume

Larson T./ Migdalas A./ Rönnqvist M. (1994): A Lagrangean heuristic for the capacitated concave minimum cost network flow problem. In: European Journal of Operational Research 78, 116-129

Leung J. M. Y./ Magnanti T. L./ Singhal V. (1990): Routing in Point-to-Point Delivery Systems: Formulations and Solution Heuristics. In: Transportation Science 24, 245-260

Magnanti T. L./ Wong R. T. (1984): Network design and transportation planning: Models and algorithms. In: Transportation Science 18, 1-55

Minoux M. (1989): Network Synthesis and Optimum Network Design Problems: Models, Solution Methods and Applications. In: Network 19, 313-360

Paraschis I. (1989): Optimale Gestaltung von Mehrprodukt-Distributionssystemen: Modelle - Methoden - Anwendungen. (Physica) Heidelberg

Skorin-Kapov D./ Skorin-Kapov J. (1994): On tabu search for the location of interacting hub facilities. In: European Journal of Operational Research 73, 502-509

Stumpf P. (1997): Vehicle routing and Scheduling for Trunk Haulage. In this volume

Tüshaus, U./ Wahl, Ch. (1997): Inventory positioning in a two-stage distribution system with service level constraints. In this volume.

Tüshaus, U./ Wittmann, S. (1997): Strategic Logistic Planning by means of simple plant location: a case study. In this volume

Wlcek H. (1997): Local Search Heuristics for the Design of Freight Carrier Networks. In this volume

Strategic Decision Making for Logistics Network Design

Lorike Hagdorn - van der Meijden and Jo A.E.E. van Nunen

Erasmus University Rotterdam, Rotterdam School of Management,
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands,
Telephone: (31)10 4082032, Fax: (31)10 4523595,
E-mail: lhagdorn@fac.fbk.eur.nl, jnunen@fac.fbk.eur.nl

Abstract. The process of creating competitive logistics networks is often very complex: a range of alternative networks needs to be developed, analyzed and compared; a wide variety of quantitative and qualitative criteria are considered and several parties with differing functional backgrounds and differing disciplines are involved. To reduce this complexity, we present a framework for logistics network design, which, as we have experienced, enhances the quality of the design process as well as the quality of the resulting network.
After analyzing previous research, we present our framework for logistics network design. The framework starts with an analysis of external and internal developments and uncertainties. Via this analysis, business choices are made on entrepreneurial topics, technological issues, organizational aspects and human resource questions. These business choices are combined into scenarios and lead to the developments of alternative logistics networks. In the development and in the evaluation of alternative logistics networks, the decision support system SLAM plays an important role. The final steps in the framework are the final selection of a logistics network, the authorization and the implementation and review of this selected network.
Besides these steps in the framework, also the different types of participants and their role in the framework are worked out and the routing of the decision-making process through the described steps is considered.
Finally, the framework is illustrated by two applications, one in the food industry and one in the business electronics industry.

Keywords. Framework, scenarios, external developments, internal developments, decision criteria, decision support, cases

1 Introduction

This paper is concerned with the analysis and design of logistics networks for industrial firms. A logistics network is comprised of suppliers, plants and warehouses which, by a systematic transfer of raw materials, semi-finished and

finished products, accomplish the delivery of the final product to the customer at the right time and in the right place (see figure 1).

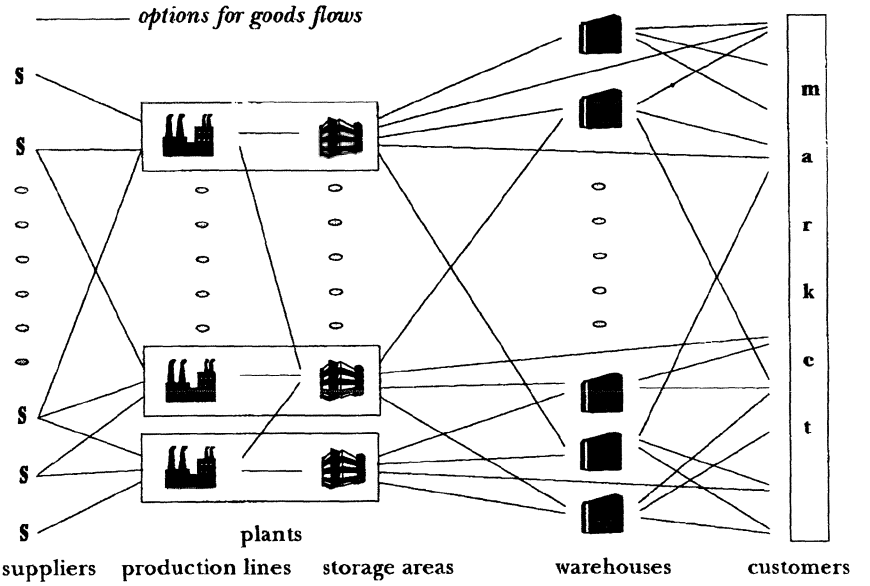


Fig. 1 Diagram of a logistics network

In the turbulent environment of today, in which new markets are emerging and customers are asking for high quality products, produced according to customer specifications and delivered within short lead-times with high reliability; in which technologies are evolving fast; and in which environmental issues cannot be ignored, companies are looking for new opportunities to enhance their competitive advantage (see Bowersox, 1992, 1995, Hagdorn, 1996, Boutellier et al., 1998, Fleischmann, 1998, Tüshaus et el., 1998).

It will be clear that theory about support in the complex process of creating competitive logistics networks is of vital importance, both for industrial companies and for research in the field of business administration.

When designing a competitive logistics network for a specific company, the first question that needs to be answered is to determine the number of echelons and the type of facilities needed. In addition, the number, location and role of each type of facility has to be decided on (see also Geoffrion and Powers, 1995).

To give an idea of the type of problems we have in mind, we consider the following example of a multinational European company that manufactures, sells and distributes 200 different types of consumer electronics products, such as

faxes, printers, copiers, personal computers, etc. to about 3,000 European dealers and wholesalers. The products are produced on 75 production lines at five European production plants. The finished products are stored at national warehouses, which deliver the products to the customers. Some semi-products that are used in the production process are purchased from external suppliers, others are produced by suppliers owned by the multinational company itself. In all, about 50 suppliers and 100 types of parts are involved in the multinational's logistics network. Some suppliers supply parts to just one plant, others serve several plants. It will be clear that this logistics network results in numerous goods flows across Europe.
The managing board of the multinational company anticipates a range of opportunities and threats in the near future as described previously. Therefor, the managing board commissions an investigation into the possibilities of improving service and reducing logistics costs and delivery times by restructuring the companies European logistics network. The suggest to consider a reduction of the number of suppliers, to cut down the number of twelve national warehouses to a few large European warehouses and to consider customer delivery direct from a plant.

In this investigation, several questions of the following type have to be addressed:

- How many plants and warehouses are needed?
- What are the best locations for the plants and the warehouses?
- What should be the size of the plants and the warehouses?
- Which suppliers are needed and which parts should be purchased from which supplier?
- Which products should be produced by which plants?
- Through which warehouses should the products flow from the plants to the customers?
- etc.

Answering these questions, while taking account of developments such as described above, inevitably leads to a complex design process. The importance of the logistics network that is designed during this process for a company like the one indicated in the example, brings us to the central problem of this paper "*How to design a competitive logistics network for a specific industrial company?*"

In the logistics network design (LND) processes in which we have been involved, a range of alternative networks needed to be developed, analyzed and compared. This comparison was based on a wide variety of quantitative and qualitative criteria, in order to arrive at the most suitable logistics network for a specific company. Moreover, several parties with differing functional backgrounds (board members, staff departments, operational managers, etc.) and from differing disciplines (logistics, marketing, finance, etc.) were involved in the design process. The result was a complex and often time-consuming process with many

interruptions and feed-back loops. We have experienced that the use of a framework for LND enhances the quality of the design process as well as the quality of the resulting network. Such a framework can provide a sound basis for structuring the decision process, developing scenarios and gaining insight into relevant decision criteria. In this paper a framework for the design of a competitive logistics network for a specific industrial company will be described.

2. Previous research

In the literature several authors have described a framework for LND, often with a specific focus (see table 1). If we combine our experience gained from real-life cases with the descriptions found in the literature, we can make a number of observations:

- *Logistics focus*

In the classical logistics approach, production and distribution are considered separately. From the point of view of an integrated supply chain, production and distribution are closely connected. Table 1 shows that in most existing frameworks the main emphasis is either on production or on distribution.

- *Organization of the design process*

The design of a logistics network is a complex process. Looking at the design process as a decision-making process may offer additional insights that may improve the design of the logistics network. Table 1 shows that existing frameworks do not provide a detailed explanation of the LND in terms of the decision-making process.

As mentioned earlier, there are many parties, with different types of functional backgrounds, involved in the LND process.

Incorporating this aspect into the LND framework may also produce additional insights. Van de Ven and Ribbers (1993) and Mourits (1995) comment briefly on this aspect.

- *Development of alternatives*

In each of the LND processes in which we were involved, several alternative networks were designed, analyzed and compared. The need for a structured approach to developing these alternatives and integrating them in scenario's is reflected by four of the frameworks in table 1, but is not worked out in detail.

- *Evaluation criteria*

During the LND process, a range of evaluation criteria, both quantitative and qualitative, are applied. Nearly all frameworks in table 1 take some criteria of both types into consideration.

- *Support by DSS*

A Decision Support System (DSS) is very helpful in specifying the values of the quantitative evaluation criteria. The frameworks in table 1 enable the development of several quantitative models for simulating or optimizing logistics networks. However, only Bender (1985) considers

how these can be incorporated in a DSS and how a DSS can support the decision-making process. Although most of the frameworks in table1 pay little attention to the potential role of DSSs in the design of logistics networks, several DSSs for LND exist (see Hagdorn and Warffemius, 1995).

	Bender (1985)	Cohen, Fisher, Jaikumar (1989)	Cook, Burley (1985)	Fine, hax (1985)	Rushton, Saw (1992)	Ven, Ribbers (1993)	Vos (1993)	Mourits (1995)
Logistics focus								
Production	no	yes	yes	yes	yes	yes	yes	yes
Distribution	yes	yes	yes	yes	yes	yes	yes	yes
Organization of the design process								
Considered as a decision process	no	no	no	no	no	yes; not in detail	yes; not in detail	yes; not in detail
Different parties and disciplines distinguished	no	no	no	no	no	yes; not in detail	no	yes; not in detail
Development of alternatives								
	yes; not in detail	no	no	no	yes; not in detail	yes; not in detail	no	yes; not in detail
Evaluation criteria								
Qualitative aspects	some	no	yes	yes	yes	yes	some	some
Quantitative aspects	yes; focus on costs	yes; focus on costs	yes; focus on costs	no	yes	yes	yes	yes; focus on costs
Support by DSS								
Models	yes; optimization and simulation	yes; optimization	yes; simulation	no	yes; simulation, but not in detail	yes; simulation	yes; not in detail	yes; optimization
DSS	yes; not a specific one	no	no	no	no	no	no	yes
Framework based on cases								
	a small one	no	no	yes	yes	yes	yes	one on distribution

A shaded area shows the main topic of the corresponding framework

Table 1 Comparison of frameworks for LND.

It will be clear that there are many questions in the field of LND that are still waiting to be answered. In this paper we hope to provide an answer to some of them or to improve the solutions proposed by other authors (see also Fleischmann, 1998 and Boutellier et al., 1998). In our framework, we will try to:

- Focus simultaneously on production and distribution.
- Analyze the design process as a strategic decision-making process.
- Take account of the involvement of different parties and disciplines.
- Structure the process of scenario development, in which both qualitative and quantitative evaluation criteria are used.
- Specify the valuable role of a DSS in the design process.

3. Classification of developments and uncertainties

As illustrated in the introduction, there are many uncertainties related to the future that need to be considered when designing a logistics network.

To classify these uncertainties, we will use Broekstra's Consistency Model for Organizational Assessment and Change (1984, 1989) depicted in figure 2. This model divides the developments and uncertainties into an external category, related to the environment, and an internal category, related to the company. It also shows the links between these two categories. Several other models exist for analyzing developments and uncertainties (see Aaker, 1984, Wheelen and Hunger, 1995), but these do not show the relationships between external and internal developments and uncertainties as clearly as Broekstra's model.

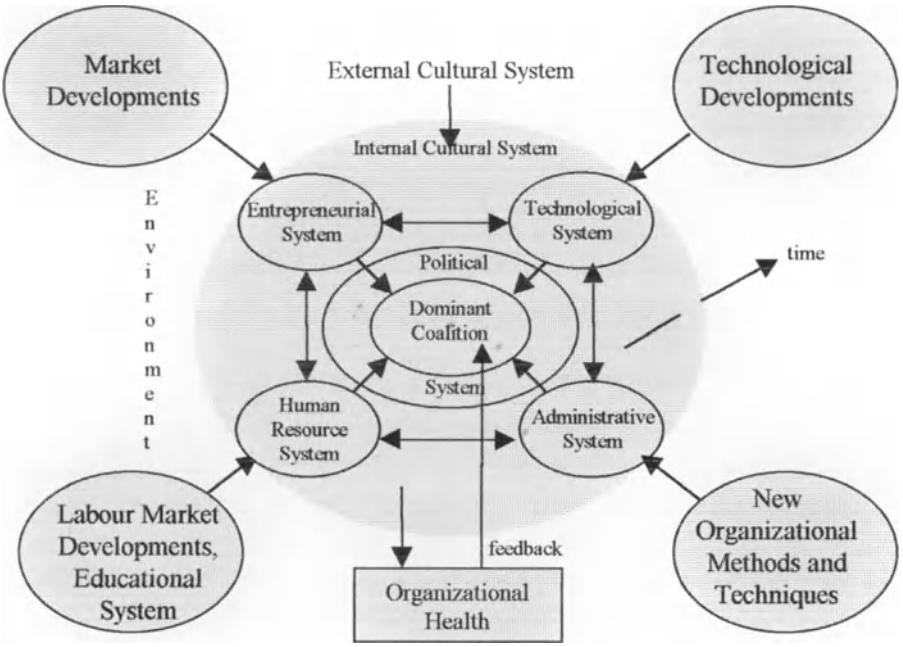


Fig. 2 A Consistency Model for Organizational Assessment and Change (Broekstra, 1984, 1989).

Broekstra's Consistency model allows us to classify the developments and uncertainties into four subcategories with each an external component and a company component:

1. The *`Market Developments'* represent the external developments with respect to markets, customers, suppliers, competition, governmental policies, regulations, environmental issues and economics.

These external developments determine to a large extent the potential internal options of a company regarding *`Entrepreneurial Elements'*. Broekstra's entrepreneurial aspect system basically refers to product-market combinations. It

also covers the options of a particular competitive strategy, for example, differentiation in cost leadership or in customer valued items (see Porter, 1985). The entrepreneurial choices reflect important requirements that must be fulfilled by the logistics network that is being designed (see also Kraus, 1998 and van der Laan et al., 1998).

2. '*Technological Developments*' concern innovation in products, production methods, distribution methods, computer facilities, electronic data interchange, multi media applications, electronic commerce, etc. How a company can take advantage of these innovations is expressed by Broekstra in the business options regarding 'Technological Elements'. Broekstra's technological aspect system refers to the company's production and distribution 'hardware', i.e. its primary conversion process. It also includes the tools for daily operations, such as automation of manufacturing or warehousing processes, flexible manufacturing processes, techniques for remanufacturing or recycling used products, appropriate transportation facilities, tools for information processing within the logistics network like EDI, etc (see also van der Laan et al., 1998 and Daduna, 1998).

3. '*New Organizational Methods and Techniques*' refer to developments and new insights in organizational structures, for instance centralization versus decentralization, product-oriented approaches versus market-oriented approaches, mergers, take-overs, co-makerships, global network corporations (Maljers, 1995), etc. New management control techniques, changing accounting methods, etc., also belong to this category of external changes. In the company's administrative aspect system, these external developments are evaluated and options for their implementation in the organizational structure developed in terms of distinct '*Administrative Elements*'. With respect to this aspect system, the logistics network needs to consider such options as national or international structure, local plants for local production or special purpose plants for international production, co-makerships with suppliers and transport companies (see also Corbett et al.,1998), centralized versus decentralized activities, etc.
This aspect system also includes the systems that administer and control the activities of the technological and human resources (socio-technical) system. This means that it also includes the options for logistics planning concepts, such as MRP II, JIT, OPT, DRP II, ERP, ECR, the cost accounting method, management information at the various control levels, decision-making processes, etc.

4. '*Labour Market and Educational System Developments*' illustrate the changes and developments in the labor market with respect to the available labor force, its educational level, wage rates, developments in the flexibility of the labor force, etc. These external developments are reflected in the company by Broekstra's internal aspect system of '*Human Resource Elements*'. This aspect system refers to the organizational 'software', the characteristics of employees (age, skills, knowledge, turnover, motivation, satisfaction, leadership styles, flexibility, etc.) and the nature and quality of the social relations (the socio-psychological system).

Human resource elements are, for example, training of employees in order to benefit from changed technology, and control concepts. It also includes the dismissal, recruitment and transfer of employees in case of closure, reduction or expansion of plants or warehouses.

This classification of the external and company related components helps us to structure the process of designing a logistics network by using scenarios. An important step in the framework consists of analyzing relevant developments and uncertainties in the external environment. These are divided into the four categories, as explained above: market developments; technological developments; new organizational methods and techniques; and developments in the labour market. The values of the external factors of these four types are combined into consistent sets. Each set of consistent values represents a view of the future and is defined as an external scenario.
The external scenarios provide the basis for the strategic choices concerning the logistics network. These are elaborated into alternative company scenarios. These are sets of mutually consistent values of company factors, which Broekstra also classified into four categories: entrepreneurial elements, technological elements, administrative elements and human resource elements in the logistics network.
The external and the company scenarios play an important role in the framework we will describe.

4. Participants in the framework

Due to the complexity of the problem, many participants with different types of functional backgrounds are involved. Based on the cases we were involved in, we distinguish in our framework three multi-disciplinary composed groups of participants: the top management, the task force and the field.

The top management
The top management plays a major role in the process of designing a competitive logistics network for the company:
they initiate the process of developing an LND, often on the basis of suggestions from the field; they determine the strategic directions of the firm and possibly develop some ideas for the design of the logistics network; they establish a task force to elaborate their ideas and to develop alternative LNDs; and they take the final decision as to the LND that will be implemented.

The task force
The responsibility of the task force is to develop alternative LNDs, which involves providing support to the top management in working out their objectives and ideas, as well as communicating and coordinating the process with the business units (the field).
The task force is usually made up of logistics, finance and marketing experts. In addition to corporate experts, the task force often includes representatives from the field, to facilitate coordination with the field. Depending on the phase of the

design process, other specialists may be involved in the work of the task force. Within the task force often several working groups are appointed whose task it is to explore new options and side constraints for an LND, such as standardization of products and packaging materials and the introduction of EDI.
Sometimes the task force is embedded in the existing organization (e.g., the European logistics department), sometimes a special working group is established which is allocated to the top management team for this special purpose.

The field
The field consists of all actors involved in the operations. Sometimes the field indicates to the top management to improve the logistics performance and usually the field is involved in the process of designing a new logistics network. They provide the task force with data on cost rates, demand forecasts, customer locations, specific local market requirements, etc. Moreover, they assist in the evaluation of the operational feasibility of the LND, by investigating the requirements for its implementation.

5. The DSS SLAM in the framework

To develop a competitive logistics network for an industrial company, a DSS that calculates the structure of the logistics network that fulfills the market requirements against lowest possible variable logistics costs, can be very helpful. The DSS SLAM, developed by us around a Strategic Location Allocation Model, proved to be very useful. Also other models and DSSs can be used within our framework (see for example Diks et al., 1998, Bertazzi et al., 1998, Klose, 1998, Bruns, 1998, Fleischmann, 1998, Wlcek, 1998, Daduna, 1998, Tüshaus et al., 1998).
Figure 3 provides a detailed picture of the contribution of the DSS SLAM and the incorporated MILP model (see also van Nunen, 1984 and Hagdorn, 1996).

The minimum information needed for the SLAM is information on products, markets, demand forecasts, optional number and types of echelons in the logistics chain, optional types and locations of facilities and cost levels of the activities (1a). If more details on the business choices are available, information on customer service levels (represented by lead time and inventory levels), fixed locations and activities of facilities, minimum and maximum sizes of facilities and transportation flows between facilities can be added as input for SLAM (1b).
On the basis of this information, the MILP model determine the values of the factors concerning the facilities and the flows in an LND with the lowest possible level of total variable logistics costs (2 and 3).

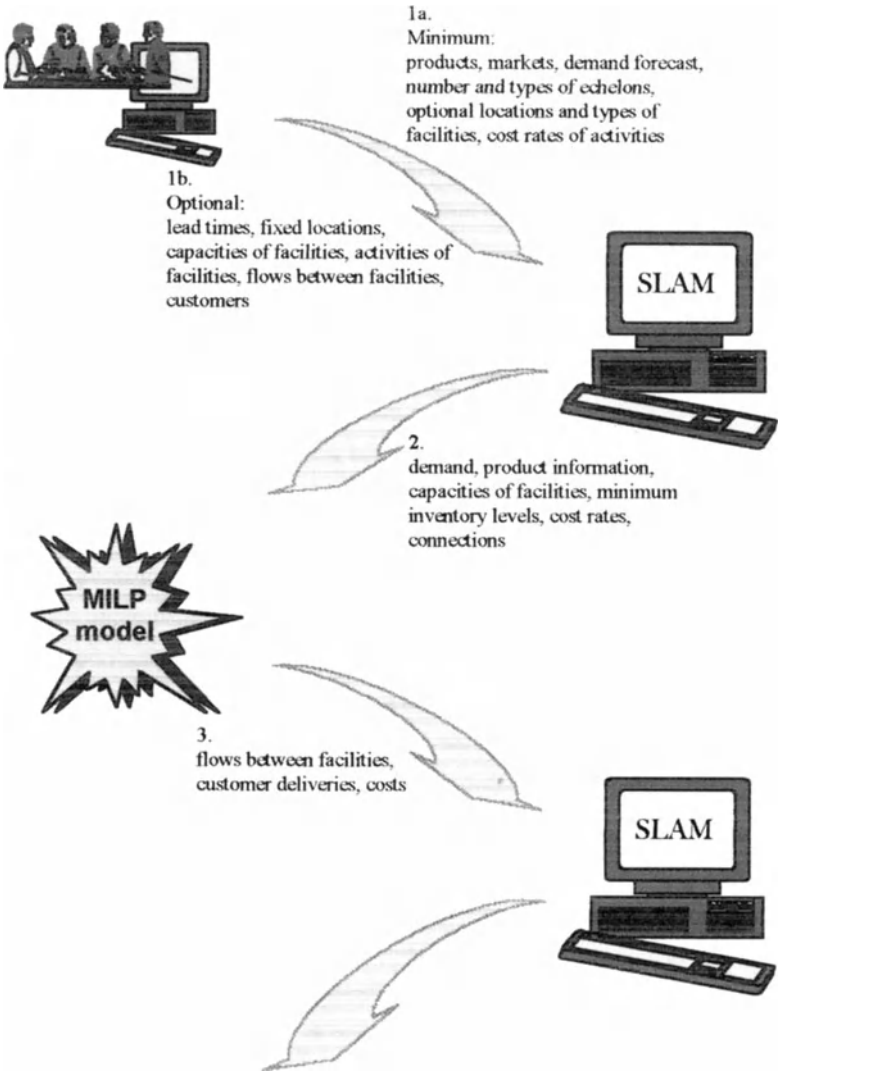


Fig. 3 Contribution of DSS SLAM and its MILP model.

The values concerning the facilities and flows constitute a logistics network structure, which is reported by SLAM on a range of quantitative aspects (4), such as number and sizes of facilities, customer deliveries, lead times and logistics costs.

6. Steps in the framework

Figure 4 shows the new framework in detail. The steps in the framework will be described by integrating the use of scenarios, the roles of the participants and the contribution of the DSS SLAM. The ultimate goal of the framework is to create for a company the most competitive LND for a company. This is achieved by developing alternative external scenarios, translating them into company scenarios, evaluating and comparing these scenarios (especially their LNDs), and selecting the most competitive LND for implementation. Besides the development and comparison of new LNDs, the existing LND and its underlying business choices are also described and used as a reference for the new LNDs and their underlying business choices. The existing LND is the starting point for the reorganization process. It gives insight into the business improvements that can be made and it shows which investments are needed to achieve these improvements.

Below, the steps of the framework are described. The numbers between brackets in the descriptions of the steps refer to the numbers of transitions in figure 4.

Why redesign the LND?

The process of designing an LND usually starts with the identification of opportunities, problems or crises related to the existing logistics network. Top management starts the LND process, prompted by signals from the field (1) or by external signals.

In this phase, management information systems are often useful tools to identify and specify the motives for redesigning the logistics network, such as a declining or increasing market share, competitors' innovations, increasing costs, declining returns on investments, etc.

External environment

At this stage external scenarios are developed. The top management gives a preliminary idea of the factors and the ranges of factor values

that will play a role in the external scenarios. Information is needed on such issues as market developments, technological innovations, interest rates and political developments. On the basis of this information, experts may extend the required factors and values to describe alternative external scenarios in detail. At this stage in the process, the task force, of which these experts are members, is installed (3).

To gather the required information, the task force can use executive information systems, research companies and other sources such as information available on the internet.

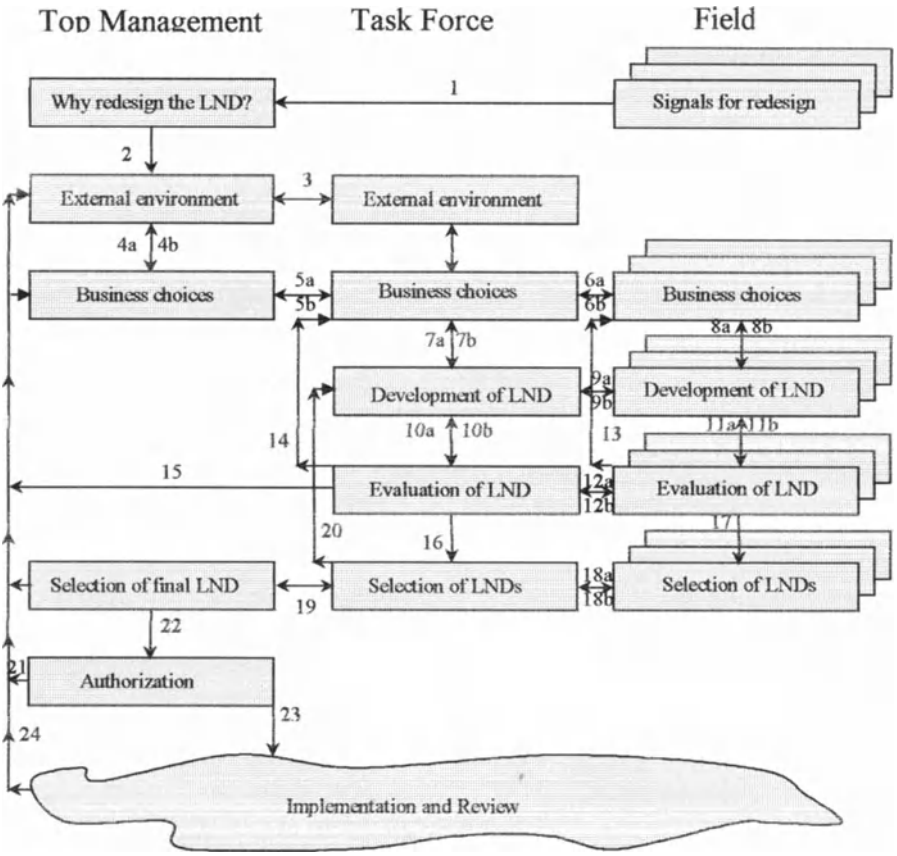


Fig. 4: A framework for logistics network design

The outcome of this phase is a set of sometimes as many as 20 external scenarios. A selection of two, three or four widely diverging external scenarios are used as a starting point for the development of company scenarios.

Business choices

Through business choices, a start is made with the translation of each of the selected external scenarios into one or more company scenarios (4a,c). The focus is on the entrepreneurial choices in which several objectives are set regarding cost reductions, customer service improvements, time limits for the reorganization, etc.

Note that also a thorough investigation of the existing situation is needed to gain insight into the starting points for the reorganization process. In this stage, the focus is on the entrepreneurial choices that were made and are still valid in the current situation.

Often the top management and the task force first discuss the business choices and especially the entrepreneurial choices in fairly general terms (5a,b). Following this, the task force is asked to elaborate these global settings. To prepare the proper entrepreneurial choices, the task force also cooperates with the field. Together, they investigate the signals from the field (if any) regarding the redesign of the existing LND and they gather data on the existing business situation and options for future entrepreneurial choices (6a,b).
Sometimes additional investigations on the external environment are needed (4b,d).

The information needed in this phase concerns existing and forecasted market shares, lead times, cost rates, etc. Again, information systems may help to generate the necessary data. The result of this phase is a description of previous business choices that led to the existing business situation.

Development of LNDs
In this phase (7a), alternative LNDs are developed. Moreover, a detailed specification of the present LND is made. In this specification the input parameters for the MILP are given. Here, data like the current plant locations and production capacity locations and sizes of warehouses are specified together with cost components which depend on e.g. frequency of delivery and routing tours (see Bertazzi et al., 1998). This present LND is the reference for all the alternative LNDs that will be developed.

The task force elaborates the top management's guidelines for LNDs. They cooperate with the field, especially in gathering additional data on the existing and potential LNDs (8a). Sometimes, the field also provides suggestions for alternative LNDs (9a,b).
Note that the top management team does not play a role in these phases of detailed development and evaluation of LNDs. This shows that design of the logistics network is largely delegated to the task force, which underlines the importance of its supportive role.

In the development phase of the framework the DSS SLAM fulfills an important role. Its first task is to help specify the existing LND as a reference for the comparison of the alternative LNDs. Then, SLAM and especially its MILP model, can help define alternative LNDs, based on the business choices made in the previous phase in the framework. In case that an unfeasible solution of the MILP model is generated, often changes are needed in the business choices (7b, 8b).

Two alternative LNDs that are often developed are the `green field' alternative and the `optimized' existing LND. The green field alternative shows an LND which sets no limits on the capacities of the facilities and which does not fix flows of goods or allocations of customers in advance. From the costs perspective, this is

often seen as the ideal LND. The `optimized' existing LND contains the capacities of the existing plants and warehouses, but with the 'optimized' quantities of semi-products and finished products flowing between the facilities and without the existing allocation of customers to warehouses.
In the next step (11a), an alternative LND is evaluated. If the evaluation gives rise to modifications, a reiteration of the development phase (10b, 11b) will take place to improve the LND.

The result of the development phase is that the existing logistics network is described as a reference for the alternative LNDs developed in this phase. These LNDs are evaluated in the next phase; a return-loop to the development phase may be needed to improve them until they meet the evaluation criteria specified in the evaluation phase.

Evaluation of LNDs
In this phase, the LNDs are evaluated by the task force, often supported by the field (12a,b). Also the existing situation is evaluated. Figure 5 presents the evaluation criteria.
The first criterion is the *operational feasibility* of the LND. Here, the need for human resources, technological concepts, administrative processes and management control activities of the logistics network and its implementation are investigated (see Ackoff et al., 1984, Anthony, 1992).
A second criterion, based on Broekstra's Consistency Model for Organizational Assessment and Change (1984, 1989), is the *political feasibility* of a scenario. Broekstra describes the `political aspect system' of a company as the distribution and use of power and influence across the organization. The feasibility of a proposed LND with respect to this political system is important for its success.
A third aspect of the realization of an LND is the *time schedule*. If the time path set for the implementation is too long, changes in the LND are required.
While the operational, political and time aspects are related to the internal implementation aspects of an LND, Porter (1985) focuses on the sources of *competitive advantage* in an LND (e.g., cost advantage, buyer value, technology, first mover advantage) and the *competitors' reactions* to each alternative. These criteria can be seen as belonging to Broekstra's concept of `organizational health', which in fact is the main evaluation criterion. This concept is often explained as *return on investments*. Expectations concerning the financial returns on investments are based on expected future operational costs of the suggested LND, the long-term investments that are necessary (like new technologies, new buildings etc.) and the costs of the reorganization process.
Besides this `hard' performance, Broekstra also includes `soft' performance (or *social effectiveness*) in his concept of organizational health. Social effectiveness refers to the complex of the firm's value system, mission, philosophy, behavioral norms, belief systems, climate, etc. This internal cultural system is crucial in attaining effectiveness.

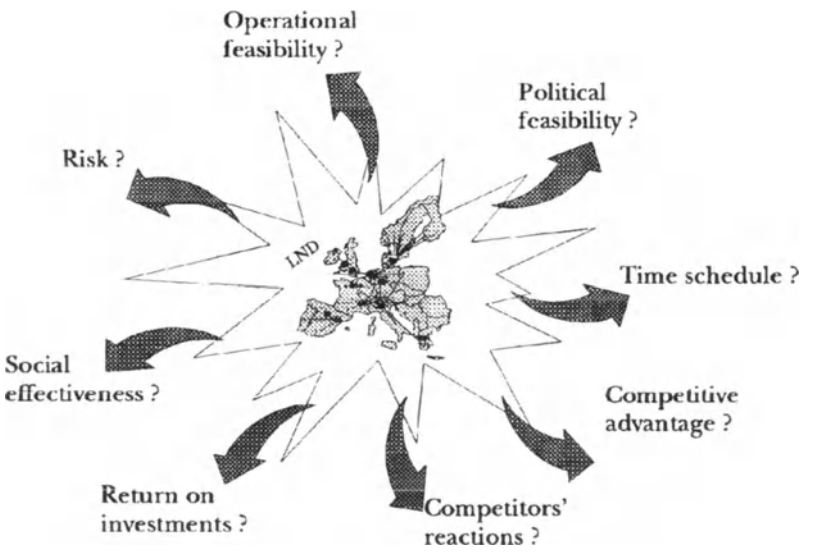


Fig. 5 Criteria for the evaluation of an LND.

The final evaluation criterion we propose is *risk*. Risk is a function of how poorly an LND will perform if a `wrong' external scenario occurs (Porter, 1985). Risk also depends on the degree to which a company is tied up, once it has committed itself to an LND by setting its product line, customer service levels, facilities and so on. This means that risk is closely related to the sensitivity of an LND to uncertain future external developments. In the evaluation of the risk of an LND, the external scenarios that were not selected for the development of LNDs are also used.
As a result of this evaluation it may be necessary to adjust the LND (10b,11b), to adjust the business choices (13, 14, 15) or to adjust the external scenario (15). This process of adjustment continues for each of the alternative LNDs until the evaluation criteria are met.

The evaluation phase can be supported by the DSS SLAM and its MILP model, especially as regards the evaluation of variable operational logistics costs, use of facilities, deliveries to customers, lead times, etc. Moreover, the SLAM and the MILP model facilitate the analysis of the sensitivity of an LND to future changes, for instance in interest rates, transport rates, demand, product ranges, etc.
To assess the operational feasibility of the selected alternatives, field actors are often asked to examine the alternatives and to propose changes. SLAM can assist in this process by showing the effects of the proposed changes and suggesting further adjustments. These suggestions are gathered by the task force and incorporated into the alternatives. Again, SLAM and its MILP model play an

important role in combining the field suggestions, showing their effects and evaluating new alternatives. This shows the close relationship between the phases of `development of LND' and `evaluation of LND'.

Selection of the final LND
When for each external scenario several company scenarios (and incorporated LNDs) have been developed, finally one specific LND has to be selected (16, 17). In the previous step, a range of evaluation criteria were presented. In this multi-criteria decision-making problem, the decision dilemma is, according to Porter (1985): ``A company does not know which scenario will occur, so it must choose the best way to cope with uncertainty in selecting its strategy, given its resources and initial position". In this strategic decision-making problem, the theory of multi-criteria decision-making, in which all criteria are measured quantitatively, does not apply (see Korhonen et al., 1992). Porter (1985) describes five main approaches to the selection of a scenario. We apply them to the selection of an LND:

- *Bet on the most probable LND*
  Select the LND that is based on the external scenario that is considered to be the most probable.
- *Bet on the `best' LND*
  Choose the LND in which the most sustainable long-term competitive advantage is established.
- *Hedging*
  Choose a LND that produces satisfactory results under all external scenarios.
- *Preserve flexibility*
  Select the LND that preserves flexibility until it becomes more apparent which external scenario will actually occur.
- *Influence*
  Select a desirable LND that can be brought about by using company's resources.

To these five policies, Mintzberg (1994) added a sixth one:

- *Contingency planning*
  The creation of alternative LNDs to deal with different external scenarios.

In fact this comes close to a mixture of Porter's 'Preserve flexibility' and `Hedging' approaches.
In the real-life cases in which we were involved, only the `Bet on the most probable scenario' and `Contingency planning' approaches were used.

At this stage, a shortlist of about four most-preferred alternative LNDs is made by the task force, sometimes in cooperation with the field (18). The final selection is made by the top management (19).

The DSS SLAM is used in the discussions with the top management team as well as in the cooperation with the field. Using the report facilities, SLAM shows the structure of the designed networks and enables comparison of the evaluation of each LND by the quantitative criteria. SLAM is often used to provide objective information on the alternative strategies. Sometimes, improvements in an LND are needed (20), or reconsiderations of business choices, or investigations in the external environment (21).
If small changes are needed in an LND, SLAM is often used in meetings (even in the top management meetings) to show the effects of the changes instantly.
At the end of this phase of the framework, the final selection of an LND is made.

Authorization
While the final decision with respect to the selection of an LND is often prepared by the task force, the authorization of the decision is usually given by the top of the organization (22). Often the proposal should be approved by several parties that have the power to enforce modifications or to influence the acceptance process (e.g., trade unions, consumer organizations). Note that a preliminary acceptance by the field is often partly ensured by the composition of the task force and by the interaction with the field during the process.
By approving the final selection of the LND, authorization is given for the implementation of the complete LND, or part of it (23). The selected LND is often first set up as a pilot for one country or one business unit.

Implementation and Review
On the basis of the experience gained during the implementation, minor or major modifications may be made to the new LND, or the reorganization process may be suspended. In the case of a pilot implementation, the review phase is important for making any changes needed in the LND before further implementations are started. In case of major modifications this may be the start for a new loop in the framework (24).
The review may also focus on the decision process. On the basis of the review, lessons can be learned and future decision processes can be improved.

7. Loops in the framework

Mintzberg et al. (1976) distinguish three simultaneously occurring driving forces that affect the routing of the decision-making process through the described steps in the framework:

- The *decision control procedures* guide the way in which the decision process evolves and the allocation of the organizational resources.

- The *decision communication procedures* determine the exchange of information during the decision process. These procedures range from general scanning (exploration) and focused searches for information (investigation) to the distribution of information among the involved parties.
- The *political procedures* represent the way a decision process evolves in an environment of influencing and sometimes hostile forces.

The influence of these three types of procedures on the decision process manifests itself in the form of interruptions, scheduling delays, feedback delays, timing delays and speed-ups, comprehension cycles and failure cycles.
In terms of Mintzberg's 'driving forces', the three parties involved in our framework, their tasks and their interactions belong to the decision control procedures; the management of information exchange during the decision process belongs to the decision communication procedures; the influence of the field, trade unions and consumer organizations may lead to political procedures.

The sequence and number of loops and cycles in the framework strongly depend on the decision situation. Sometimes a large number of development and evaluation steps are needed by the task force to develop a first set of interesting LNDs. In other situations, the field has all kinds of suggestions, ideas or political objections, resulting in several loops between the task force and the field before a compromise is reached on proposals for LNDs that are ready for evaluation by the top management. Sometimes the top management is closely involved in the process and asks for frequent feedback from the task force. Interaction between the task force and the top management may also be prompted by new insights presented by the task force that compel the top management to reconsider its objectives.

8. Applications of the framework

In this paragraph we will show how the framework has been applied to a specific European company in the food industry and a specific European company in the business electronics industry. Both companies were considering a rationalization of their production and distribution strategies in Europe.

8.1 An application in the food industry

The company is a large European company producing food products at several locations in the world and selling them worldwide. We will consider the situation in Europe. Forced by shortages in production capacity at some locations and overflows at others, the board has established a task force, named `European Production Coordination'. This is a multi-disciplinary, cross-organizational task force with no responsibility for operations. Its mission is to show the way towards cost reduction through rationalization of the European production and distribution structure and standardization of products and packaging types. In the design of a

new logistics network for this company, 20 plants with a total of 100 production lines, 16 existing and about 25 potential warehouses, 500 customer groups, 15 types of purchased products and 50 different finished product groups are involved. The task force has organized itself in four working groups, focusing respectively on sales forecasting, product standardization, packaging standardization and a new production and distribution structure. We were especially involved in the latter one. The four working groups are coordinated by a coordination team, which reports to the board. Figure 6 shows an overview of the framework that emerged from this project.

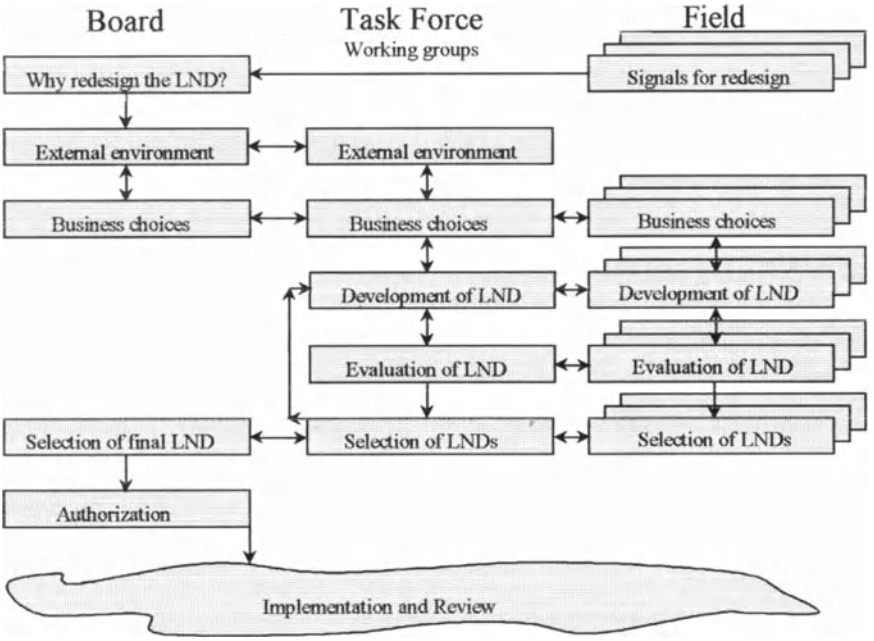


Fig. 6 Framework for an application in the food industry.

Numbers of loops and cycles

The start of the process was initiated by the field. The board, the task force and its working groups considered the complete logistics network, whereas each field party considered only their geographical region and product groups of the logistics operations. Several members of the working groups were representatives of the field.

The external scenarios and the business choices were developed simultaneously in three cycles in which the board and the task force were involved.

In the development and evaluation phase, our working group, which dealt with the new production and distribution structure, reported eight times to the tasks

force's coordination team. Both in the phase of making business choices and in the evaluation phase, our group consulted with the field twice even though the field was represented in the working group. The proposals for a new LND were discussed by the board and the task force in four cycles. In our working group, each time the board had suggestions for improvement we made about three return-loops from the evaluation phase to the development phase. In this loops, not only the LND was improved, but also business choices were adapted or worked out into greater detail.
As a result of the organization of the task force in four working groups, there was a great deal of communication between the coordination team and each of the working groups, among the working groups and between each working group and the field. The coordination team of the working groups met about 15 times.

Time needed
The total process from problem identification until authorization of the new LND by the top management took about 18 months. In our working group, we used the DSS SLAM and its MILP model intensively. We needed one month to gather raw data from eight countries and 36 locations and, at a later stage, two months to gather the detailed data we needed. When the data collection was finished, the eight cycles of scenario development that followed took about two weeks each.

Results
The result of this project was a strategy to realize product standardization and packaging standardization for all European countries and a plan for a new production and distribution structure. In this new structure, production and warehousing capacities were reallocated, several plants and warehouses were closed down, a few new warehouses were built, and the product flows from plants to customers were reallocated. The project is expected to lead to a reduction of total variable logistics costs by about 15% on an annual basis.

8.2 An application in the business electronics industry
This company produces and sells business electronics products in most parts of the world. In this description, we will consider the European market and the European production and distribution facilities. A request for a large investment in one of the European production plants was the motive for the board to ask the European Logistics Department (ELD) to reconsider the LND in Europe. The ELD's mission was to reduce the total logistics costs by 30% and to reduce the order-to-install lead times from 72 to 24 hours for as many products and markets as possible. We were asked to support the ELD, which is responsible for the operations in Europe, in developing a new LND which would contribute to these objectives.

The company serves about 500 market areas in Europe, which demand five types of products. As the production facilities at three European locations (three plants with each about seven production lines) were fairly new, no new production

locations were considered. The company had 16 national warehouses for distribution. A reduction to a few European warehouses was considered and about 10 new potential locations were selected.

Given the complexity of the design problem and the fact that the board had to react soon to the investment request of one of the plants, we discussed with the board the possibility to divide the design process into two phases:

Phase 1: Focusing on cooperation between the three plants with respect to their production and the distribution tasks, while also considering (globally) the deliveries to the customers in the 500 market areas. This should result in a so called `top structure' of the LND.
Phase 2: On the basis of the selected top structure, in this phase the detailed redesign of the distribution structure was done.

The division in two phases proved to be very useful, because the top structure turned out to be independent of the alternative distribution structures we considered. Figure 7 shows how the framework was worked out for these two phases of the project.

Number of loops and cycles
After the request of one of the plants for investments in production capacity, the board started the process of designing a new LND. It selected some external trends that should be considered, such as internationalization in transport, increasing transport rates, etc. The board also set the objectives for the total LND in terms of cost reductions and customer service levels. The ELD was asked to work out an LND. They started by designing a `top structure'. For this part, the framework shows a simple process. In about four loops between development, evaluation and selection a few scenarios were selected. Finally, the board decided on one of these proposals and the implementation started. There was virtually no involvement of the field in this process, except in some questionnaire surveys to gather data on sales forecasts and cost rates.
The implementation of the `top structure' and the designing of the distribution structure were started simultaneously. The decision process for the distribution structure was similar to the decision process for the design of the `top structure', although at the end of the third cycle of designing alternatives by the ELD, extensive checks for feasibility and suggestions for improvement were developed in cooperation with the field.
After one more loop between development and selection by the ELD, in which the field suggestions were combined, the final selection and authorization by the board took place.
The `top structure' and the resulting distribution structure were reviewed in one review process.

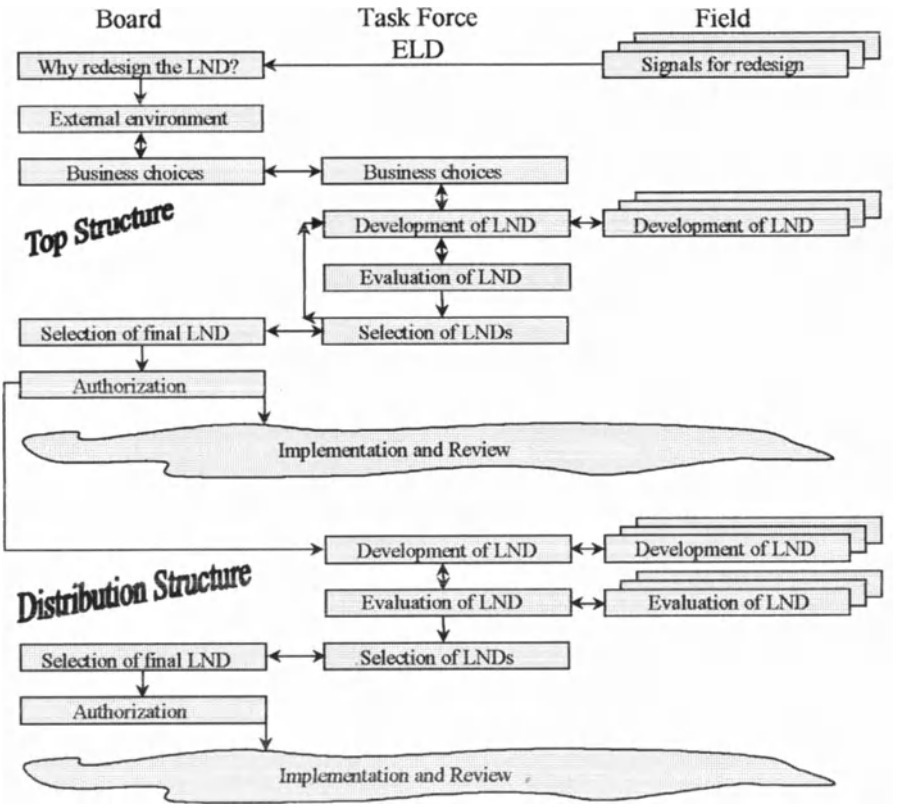


Fig. 7 Framework for an application in the business electronics industry.

Time needed

The total process, from the initiation by the board, which formed the basis for the design of the top structure, to the authorization of the design of the distribution structure took about nine months. The design of the `top structure' took about three months; the design of the distribution structure about six months. The amount of time needed for data gathering was one month for the global data for the top structure and another two months for the more detailed data for the distribution structure. The development of scenarios took two weeks for each of the eight cycles. The checks for feasibility in the field took about one month.

Results

The result of this project was a plan for a rationalized production and distribution structure: product flows between the three plants were introduced, plants were given responsibility for distribution in a specific geographical area and the

number of warehouses was reduced from 16 national to five European warehouses. The original request for a large investment, needed to double the capacity of one of the plants, was not granted. In fact, it was decided to reduce this plant by half!

The final result of these changes was a reduction of the order-to -install lead time from 72 to 24 hours and a variable logistics cost reduction of 10% per year. The other 20% cost reduction that was needed, was achieved through a strong reduction of fixed costs, realized mainly by reducing the number of warehouses.

8.3 Evaluation of the applications

In both applications, the framework was utilized, although in widely different ways. In the food company, the top management and the field were strongly involved. In the consumer electronics company, board involvement was low, although it was the board that took the main decisions, without checking in advance the ideas of the field (in the case of the top structure). In this company, the ELD was responsible for the operations (unlike the task force in the first case) and therefore took responsibility for the scenarios they developed. In the first case, there was no centralized operational responsibility and therefore the decision process was much more complicated than in the second case.

In both applications, the data gathering process was time-consuming: it took about one month to gather the raw data for the identification and some initial analyses and another two months to collect detailed data for the most important elements of the new logistics network.

9. Concluding remarks

In this paper we developed a framework to support the complex process of designing a competitive logistics network for a specific industrial company.
In the practical situations we were involved in, it turned out that the redesign of a logistics network could result in savings in the total variable logistics costs amounting to about 10% to 15% annually, while improving customer service. This results are in line with the results Geoffrion and Powers (1995).

In this paper we assumed that the information processing that is needed to plan and control the logistics activities from the ordering stage through production and distribution to delivery, ran smoothly, without causing any delays. This starts with the orders placed by customers, for instance in a traditional shop, through tele-shopping, at a call center or by electronic commerce. This information should be distributed efficiently among the participants in the logistics network, in order to ensure that the right activities are initiated at the right moments. Finally the product should be delivered in time according to the customers' specifications.
Although we did not put explicit attention to the information processing problems, we implicitly dealt with these problems. For example, one criterion

used in the evaluation of an LND was its operational feasibility; if problems, for instance in information processing, were causing delays, the LND could be adjusted. The proposed lead times incorporated in the DSS SLAM and its MILP model exploited the possibilities of electronic data exchange, and were therefore smaller than might be needed for classical information processing.
So, the framework we developed provides a structure for the design of logistics networks in which information aspects are incorporated, although refinements are needed. Extensions of the framework with respect to the integration of goods flows and information flows may lead to new challenges and requirements for LNDs - a promising area for further research.

References

Aaker, D.A. (1984): Developing Business Strategies. (John Wiley & Sons) New York

Ackoff, R.A. / Gharajedeghi, J. / Finnel, E.V. (1984): A Guide to Controlling your Corporation's Future. (John Wiley & Sons) New York

Anthony, R.N. (1992): The Management Control Function. The Harvard Business School Press, Boston, Massachusetts

Bender, P.S. (1985): Logistic System Design in: Robeson, J.F. and House, R.G. (eds), The Distribution Handbook. (Free Press) New York, 143-256

Bertazzi, L. / Speranze, G. (1998): The minimization of the logistic costs on sequences of links with given shipping frequencies. In: Advances in Distribution Logistics. (Springer Verlag)

Boutellier, R. / Kobler, R.A. (1998): Strategic eurologistics design. In: Advances in Distribution Logistics. (Springer Verlag)

Bowersox, D.J. (1992): Logistical Excellence, it's not business as usual. (Digital Equipment Corporation) Burlington

Bowersox, D.J. et al. (1995): World Class Logistics: the challenge of managing continuous change. (Council of Logistics Management) Oak Brook, Illinois

Broekstra, G. (1984): MAMA: Management by Matching, a Consistency Model for Organizational Assessment and Change in Trappl, R. (ed). in: Cybernetics and Systems Research 2 (Elsevier Science Publishers B.V.) North Holland

Broekstra, G. (1989): Het creëren van intelligente organisaties. (Eburon) Delft, the Netherlands (in Dutch)

Bruns, A. (1998): A local search heuristic for the two-stage capacitated facility location problem. In: Advances in Distribution Logistics. (Springer Verlag)

Cohen, M.A. / Fisher, M. / Jaikumar, R. (1989): International Manufacturing and Distribution Networks: a Normative Model Framework. in: Ferdows, K. (ed), Managing International Manufacturing (Elsevier Science Publishers B.V.) North-Holland, 67-93

Cook , R.L. / Burley , J.R. (1985): A Framework for Evaluating International Physical Distribution Strategies. in: International Journal of Physical Distribution and Materials Management 15, 4, 26-38

Corbett, Ch. / Blackburn, J.D. / Van Wassenhove, L. (1998): Partnerships or tug of war? In: Advances in Distribution Logistics. (Springer Verlag)

Daduna, J. (1998): Modeling the distribution processes of tour operator catalogues. In: Advances in Distribution Logistics. (Springer Verlag)

Diks, E.B. / Kok, A.G. de (1998): Transshipments in a divergent 2-echelon distribution system. In: Advances in Distribution Logistics. (Springer Verlag)

Fine, C.H. / Hax, A.C. (1985): Manufacturing Strategy: A Methodology and an Illustration. in: Interfaces. 15, 6, Nov.-Dec., 28-46

Fleischmann, B. (1998): Design of freight traffic networks. In: Advances in Distribution Logistics. (Springer Verlag)

Geoffrion, A.M. / Powers, F.P. (1995): Twenty years of strategic distribution design: an evolutionary perspective. in Interfaces 25, Sept.-Oct., 105-127

Hagdorn, L. (1996): Decision Support for Strategic Planning in Logistics. Ph.D. Thesis, Erasmus University Rotterdam, Rotterdam School of Management, The Netherlands

Hagdorn, L. / Warffemius, P. (1995): Decision Support Systems for Strategic Planning in Logistics -an overview-. Working paper in Management Report Series of the Rotterdam School of Management, Erasmus University Rotterdam, The Netherlands

Klose, A. (1998): Obtaining sharp lower and upper bounds for two-stage capacitated facility location problems. In: Advances in Distribution Logistics. (Springer Verlag)

Korhonen, P. / Moskowitz, H. / Wallenius, J. (1992): Multiple criteria decision support - A review. In: European Journal of Operational Research 63, 361-375

Kraus, S. (1998): Analysis of the environmental impact of distribution systems estimating the length of trunk tours. In: Advances in Distribution Logistics. (Springer Verlag)

Maljers, F.A. (1995): Strategische Allianties. Erasmus University Rotterdam, Rotterdam School of Management, the Netherlands, (in Dutch)

Mintzberg, H. (1994): The rise and fall of strategic planning. (Prentice Hall International)

Mintzberg, H. / Raisinghani, D. / Thèorêt, A. (1976): The structure of ``Unstructured'' Decision Processes. in: Administrative Science Quarterly 21, June, 246-275

Mourits, M. (1995): Design of a Distribution Planning Support System. Ph.D. Thesis, Technical University Delft, The Netherlands

Porter, M.E. (1985): Competitive Advantage. (The Free Press)

Rushton, A. / Saw, R. (1992): A Methodology for Logistics Strategy Planning. in: The International Journal of Logistics Management 3, 1, 46-62

Tüshaus, U. / Wittmann, S. (1998): Strategic logistic planning by means of simple plant location: a case study. In: Advances in Distribution Logistics. (Springer Verlag)

Van der Laan, E. / Salomon, M. / Van Nunen, J.A.E.E. (1998): Reverse logistics and inventory control with product remanufacturing. In: Advances in Distribution Logistics. (Springer Verlag)

Van de Ven, A.D.M. / Ribbers, A.M.A. (1993): International Logistics: A Diagnostic Method for the Allocation of Production and Distribution Facilities. in: The International Journal of Logistics Management 4, 1, 67-81

Van Nunen, J.A.E.E. / Beulens, A.J.M. / Benders, J.F. (1984): On solving assignment type mixed integer linear programming problems within decision support systems. in: Wissenschaftliche Zeit\-schrift der Technische Hochschule Leipzig 8, 2, 89-95

Vos, G.C.J.M. (1993): International Manufacturing and Logistics, A Design Method. Ph.D. Thesis, Eindhoven University of Technology, The Netherlands

Wheelen, T.L. / Hunger, J.D. (1995): Strategic Management and Business Policy. (Addison Wesley)

Wlcek, H. (1998): Local-search heuristics for the design of freight carrier networks. In: Advances in Distribution Logistics. (Springer Verlag)

Delivery service: expectation, performances and costs for a distributor[1]

C. Henaux[2] and P. Semal[3]

Institut d'Administration et de Gestion, Université Catholique de Louvain, Place des Doyens 1, B-1348 Louvain-la-Neuve

Abstract. The basic service offered by a distribution center consists of providing the customer with what he ordered at the place he specified and at the time he wanted. However, since the expectations of the customers vary, this service is generally offered under various forms that have names such as emergency orders, express orders, replenishment orders, stock orders, scheduled orders, etc.

The evaluation of one such delivery service by the distribution center can be performed at three main levels. At the level of the service mix, the complementarity of the offered services must be evaluated. At the level of the service itself, the fit between its performance and the market segment it is supposed to serve should be checked. At the level of the implementation of the service, the performance of the different operations that constitute the service must be jointly evaluated and optimised.

In this work, we propose a framework for the assessment of the service and for its implementation. This framework does not give a recipe for the selection and the implementation of an adequate service. It only lists the main questions whose answers -- which are specific to each company -- will help in selecting and implementing an adequate delivery service.

Introduction

Customer service for a distributor is a crucial source of added value. It aims at satisfying the customer requirements and expectations in a global way. It is the space where marketing and logistics are linked: products take value when they are delivered from the manufacturer to the consumer in the proper quantity at the time and the location specified. Thus, customer service includes a set of factors that affect the process of delivering a product according to the customer's individual requirements. In this paper, we will focus only on one component of the customer service: the delivery service, that is the part of the logistic process between the moment an order is placed up to the time the customer receives his order.

If we put ourselves in the place of a distribution center (DC) -- what we will do throughout this paper -- we can provide this delivery service under various forms

1This work was done under the sponsorship of DHL, Belgium.

2email: henaux@prod.ucl.ac.be

3email: semal@prod.ucl.ac.be

and with various characteristics. For the delivery time, for example, shall we aim at serving our customers within 24 hours, within 48 or 72 hours or within one week? There is no absolutely correct answer. The only reasonable answer is that it depends on the customers and it depends on the costs.

This is why, in this paper, we try to define the complete frame in which our delivery service should be analysed (see Cohen / Lee (1990), Christopher (1992a)). This frame includes our customers and their expectations, our competitors with their offer and performances, our long term views and the strategy of our group and finally, our operations with our costs. In the middle of this picture is our delivery service or our set of delivery services that should be checked against each element. Do we meet the requirements of our customers? How does our service compare with that of others? Does our service fit the strategy of our company? Do we provide our service at the right cost?

All these questions must be answered at the level of each distributor. We can only propose guidelines for addressing these questions by trying to point out the most relevant actors for each decision. This will be a step-by-step program, subdivided in different phases. Each company has then to define its own priorities and, as observed during many interviews, they can be very different from company to company and from sector to sector.

This paper is organised as follows. Section 1 and 2 introduce the notions of market segment and of service mix and highlight the key factors for defining an adequate service. Section 3 focuses on indicators and points out several associated problems. Finally, Section 4 proposes a systematic analysis of the time and cost performances of a service.

1. Customer service

"Customer service" can be defined in various ways (see Bowersox / Closs (1996), Christopher (1992a), Christopher (1992b), European Logistics Association (1994)). Here is one definition.

> *Customer service is the general term used to describe the ability of an organisation to address the needs of the customers at the time, at the place and in the way required by them*

In the logistics sense, it is a service organised to provide a continuing link between the time that the order is requested and the goods are received with the objective of satisfying present customer needs and criteria and anticipating and satisfying expectations. If we look for an operational definition of customer service, the following questions need first to be addressed: who are the customers of my distribution centre, what are their needs, and what is our service? Let us try here to quickly answer these questions.

1.1. The customers and their needs

My customers are very different. They have different functions (local distributors, field engineers, consumers, OEM...), order very different products and are located

in different parts of the world. The distinction between the different types of customers, what we will call *market segments* (MS) throughout the paper, is necessary as long as their needs or expectations concerning the logistic function are different. A local independent distributor has different expectations than a field service engineer performing some maintenance or a final customer requiring an essential spare part to repair a down unit. Different customers, having different needs, can be grouped in different segments (MS) and be offered different services.

1.2. Basic service and service mix

Our basic service consists of providing the customer with what he ordered at the place he specified at the time he wanted. This means we should be able to process an order, pick the items from the warehouse, pack them and deliver them on time to our customer. However, because of the different expectations of our customers, we will provide this service in various forms, i.e., with different characteristics. Customers may emphasise speed, cost or reliability. A technician who has to repair a down unit wants the spare part to be delivered soon. A distributor who wants to replenish his inventory for the coming high-demand season requires cheap bulk deliveries. A pharmaceutical factory requiring some active product wants to get it on time in an 100% reliable way.

Ideally, we should offer a service that is cheap, fast and highly reliable. In practice, however, this is not possible and we will propose for our customers a *service mix*, a set of possible services that differ in their characteristics and aim to cover all the different needs or expectations of our customers. Table 1.1 shows a generic service mix we have observed under various forms in many different companies. This generic service mix consists of three types of orders, called stock, scheduled and emergency orders, respectively.

Table 1.1. Generic service mix

	Stock Order	Scheduled Order	Emergency Order
Frequency	weekly	Twice a week	daily
Cutoff	Friday 12.00	Mon., Wed., 7:00 p.m.	12:30
Delivery	next week	Thursday and Monday	overnight
Stock availability	98%	95%	95%
Quality	99.9%	99.9%	99.9%
Cost	X	Y	Z
Cost paid by:	us (DC)	us (DC)	the customer

In this example, an emergency order can be placed any day at any time before 12:30 and the order will be delivered overnight. The other characteristics of the emergency order are also specified: 95% stock availability (at the line level) and less than one quality problem (wrong or damaged shipment) over 1000 shipments.

The party who is charged for the service is also defined in this example. Similarly, stock orders and scheduled orders (the other services) have clear specifications.

A complete service mix, such as the one above, can be offered to a same market segment. However, in most cases, only parts of the mix are accessible to a given segment or some characteristics change from segment to segment. For example, a scheduled order is delivered from Belgium to the close European Community within 48 hours, but after three days to Scandinavian countries or in Portugal.

In practice, the notion of customer service goes beyond the questions related to delivery performance and includes many other elements before (written customer service policy and strategy, accessibility, organisation structure, system flexibility), during (orders fill rate, order status information) and after the transaction (after-sales support and call-out time, product tracing and warranty, customer complaints, non-conforming product). However, in this study we will focus on the delivery performances of the service only.

1.3. Service Mix and Continuous Improvement

The notions of market segments and of service mix immediately raise several fundamental questions:

- how to select the market segments?
- how to select an adequate service mix?
- how to measure and to evaluate it?
- how to improve it?

These questions could be seen as the main loop of a continuous improvement approach, the PDCA (Plan-Do-Check-Act) loop well known in the field of total quality management (see Shiba / Graham / Walden (1993)). These questions are successively addressed in the following sections.

2. Choosing a Service Mix (SM)

Before choosing a service mix, we will first try to understand the relationship between the expectations of the customers (market segments) and the characteristics of the services we offer. Then, benchmarking will allow us to compare the position of our company with that of our competitors. From these two analyses and our long term strategy, we should be able to select a precise service mix. All of these steps are analysed here.

2.1. Service characteristics

Our basic service consists of the delivery of products. Table 2.1 lists the main characteristics of the service.

Each customer could specify precise requirements for each of these performances. Similarly, each service we offer will be characterised by precise values for each of these performances. At this point, this list of performances calls for several comments.

- By *product palette* we mean the range of products we distribute. If we distribute bikes, our customers most likely will also sell helmets and perhaps

biking maps. They could, therefore, be interested in being supplied by a reduced number of distribution centres offering a broader product palette.

- The *price of the products* you distribute can also be a characteristic of your service. This is only relevant when these products (basic products like oil, filters, etc.) can be supplied by competitors.
- The *cost of providing a delivery service* is a major characteristic of the service. As a practical measure, we will keep this characteristic separate because it acts as a counterweight to all the other characteristics.
- The *stock availability* could be added to the list. However, we could also say that an order is delivered within 24 hours in 95 % of the cases and within 5 days for the remaining 5 %. We prefer to keep the stock availability out of the list at this point since it is not a characteristic the customer is directly interested in. If the customer requires receipt of all items of an order within five days, then he does not want to know from where the items have been shipped.

Table 2.1. Main characteristics of the service

Time aspects:	length of the order cycle, frequency of delivery, delivery reliability and consistency;
Quality aspects:	order completeness, damaged shipments, accuracy of shipments and invoices;
Other services aspects:	availability of information on order status, after-sales support;
Product aspects:	palette of distributed product, competitive pricing of distributed products;
Cost aspects:	cost of providing the service.

2.2. Market segments

Different customers have different requirements in terms of the above characteristics. In order to list these requirements, we first need to categorise the customers. Below, we propose four customer attributes: their functions, the nature of the products they order, their geographical location and our relationship with them. This is just an example based on our interviews. Other customer attributes could be more appropriate.

- *By their function.* They may be field service engineers requiring our parts for maintenance; they may use our products as raw material in their own production or service process; they may simply distribute our products further down the chain or they may just be normal end consumers.
- *By the nature of the products ordered.* These products may be accessories or essential spare parts; products from our parent company or other products we deliver; products with or without warranty; high or low density value products, etc.

- *By their geographical location.* Belgian customers most surely have different needs and expectations than French, Finnish or African customers.
- *By our relationship with them.* They may be dealers or plants belonging to our group or they can be completely independent distributors.

These customer attributes are useful in the sense that a difference in any attribute most often leads to a difference in requirements in the delivery service. A dealer has different requirements than an independent distributor. The criticality of essential spare parts and of accessories is different and leads to different requirements.

Grouping customers according to their requirements or needs in terms of the above list of performances will lead to the notion of market segments. The service requirements (concerning cost, order cycle time, quality, ...) should be homogeneous within each market segment and different between the market segments. This operation is called "*market segmentation*" (see Lambin (1993), Ch. 6). It requires a systematic review of the different types of customers. Besides the precise needs of each market segment, the absolute and relative importance of each segment in terms of sales, both today and in the future, must be recorded. This is usually referred to as the *attractiveness* of a market segment.

An important question when identifying the market segments is how broad is our product palette. This question depends on the freedom of the DC.

Note that a customer can belong to several segments according to his needs. For example, a field service engineer requires in a very predictable way basic products (oil, filters) for maintenance purposes. On the other hand, his requirements for spare parts are not as predictable. He will therefore have different requirements and belong to different segments.

At this point, we should be aware of the various market segments, their sizes and their requirements. Let us now systematically address the following questions:

- how are we perceived by each market segment?
- what are the market segment responses to changes in performance?
- how do our competitors do?
- what is our strategy?

2.3. Global perception by each market segment

The goal here is to determine the perception of our customers. Are they satisfied with our services or do they complain about some services? Are we seen as fast, reliable and high quality or as slow, inconsistent and poor quality?

2.4. Analyse market responses

The relative impact of service parameters on the market segments must be evaluated. For example, how much would we benefit:

- from an increase of the stock availability for essential spare parts,
- from an increase (or decrease) of the frequency of the stock orders,
- from a broadening of the product palette,
- from a decrease of the delivery cost for some orders, etc.

Estimates can be obtained from marketing, the customer service department or the order entry service. Many distribution centres rely on customer surveys as a fundamental source of information. The goal is to determine which service should be modified and in which direction. The attractiveness of each market segment constitutes essential data at this point.

2.5. Benchmark (see Christopher (1992a))

Our present performance must be compared with that of our competitors. Today, the need to examine the service criteria of the competition is vital.

The key areas for benchmarking the internal or external distributor in the supply chain are value-added services, customer satisfaction and delivery performance. It is important to note that it is not just distributor performance that should be monitored in best-in-class companies (although this is our main concern here) but also the management of contact points

2.6. Company strategy

Strategic decisions must be made (see Christopher (1985), Christopher (1992b)). They define how the company wants to be seen. For example, will the distribution service be cheap, fast or reliable? These decisions must agree with the general marketing strategy of your company. You cannot offer a cheap, low quality delivery service if you distribute high quality products.

2.7. Selection of a service mix

At this point, we (distribution centre) should have all the elements for choosing a service mix adequately. These are:

- the different market segments we want to serve, with their identity and attractiveness;
- the characteristics they require and their sensitivity to these characteristics;
- the position of our competitors, our own position and our long term strategy.

Assume, as an example, that the three main attributes that differentiate our market segments are: the cost of the delivery, the time (order cycle time) and the quality. Each of the market segments can be represented in a three dimensional space[4] as on Figure 2.1.

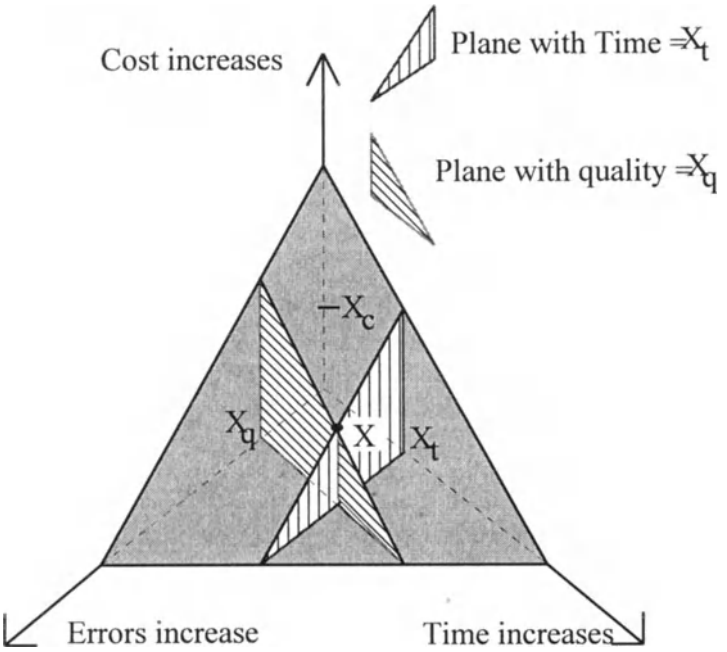
Not all the points of this space are feasible requirements. If they were, everybody would require the origin, that is a delivery service in no time with no errors and at no cost! The set of feasible requirements can be represented by a surface such as the depicted shaded triangle of 2.1. If a market segment X requires a service within X_t days and with a quality X_q, it will cost[5] X_c. If another market segment requires a quality higher than X_q, while keeping the same time performance X_t, we will move along the north-west line and increase costs. Similarly, if a market

[4]This kind of representation is classical. It was used in (A.T.-Kearney (1993)) as a tool for comparing companies.

[5]We only keep the cheapest solution for a given set of performances (X_t, X_q). This leads to the discussed surface which is, in reality, the hull of the set of all solutions.

segment requires a lower cost, some performance requirement (time or quality) must be relaxed.

Figure 2.1. Market Segment Characterization



In the same three-dimensional space, we can represent the requirements of the different market segments and the performance of the services we offer. On Figure 2.2, we give as an example[6] the position of three different market segments:

- the distributors interested in cheap service,
- the field service engineers who require a quick service and
- the OEMs (Original Equipment Manufacturer) who need a cheap and reliable service;

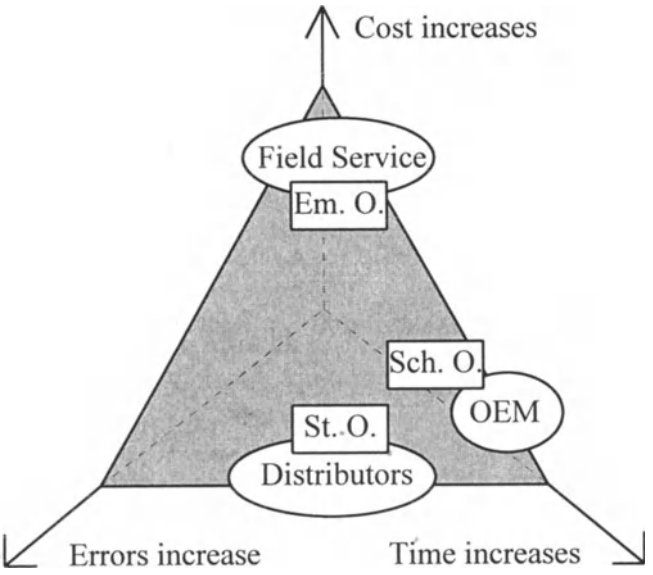
and the position of our three types of orders (services) which are supposed to serve them: the emergency (Em. O.), the scheduled (Sch. O.) and the stock orders (St. O.), respectively.

[6]This is just an example. Each company should perform its own study to identify precisely its market segments.

Our service mix in this example is thus composed of these three kinds of orders. The evaluation of the service mix is then threefold.

1. First, we should check that the different services do not overlap each other. This is condition for having clearly differentiated services.

Figure 2.2. Example of Market Segments and Associeted Services



2. Second, each service should correspond to the requirements of the market segment it is supposed to cover. We do not want to provide an average service for an average customer. Each service has to fit one or more market segments.
3. Third, the performances of each service should be optimised. This means moving the shaded surface as close to the origin as possible.

This threefold evaluation could lead to strategic decisions. Below are some examples related to the evaluation of the mix (Point 1 above).

We could decide, in this example, to abandon the stock orders in favour of scheduled orders, if their characteristics are too close. Also, if the scheduled orders are not sufficiently cheaper than emergency orders, the market segment the scheduled orders are supposed to serve could be cannibalised by the emergency orders. Service pricing could help maintain this distinction.

Another strategy could be to try to serve different market segments with the same service. The goal then would be rationalisation and economies of scales. We

could, for example, offer only the emergency order services at an intermediate price. However, the gap between the market requirements and the offer must be carefully reviewed.

The selection of this service mix is both difficult and vital. Whatever our choice is, we must continuously check whether it remains adequate. This means first a measurement process and then a feedback process. Both processes are studied in the following sections.

3. Measurement

Several years ago, service was measured with a subjective appraisal of how well and how quickly the supplier responded to inquiries and provided assistance. Consequently, many firms still do not have internal targets for performance or any means to measure and evaluate their performance in this field, despite the critical importance of customer service.

Measurement is an important part of the management process. It is essential to verify that all implemented decisions have really reached their goals. It is the third step of the PDCA cycle (see Shiba / Graham / Walden (1993). Once you have planned (P) to improve something and have done (D) something about it, you must check (C) whether the solution really caused improvements. The fourth step will then be to carry out the final actions (A).

Defining measurements means answering the following questions: what, how and when to measure? Let us first see what should be measured. We need measurements:

1. to define the service characteristics that differentiate the various market segments. Clearly, time and cost are key elements of our service. Is that all? In Section 2, we represented the various segments by a point in a three-dimensional space. These dimensions were quality, time and cost. The question is whether there are other key elements.
2. to select precise indicators, that is, precise characteristics of the service according to each of these dimensions. These are variables indicating the effectiveness of a part or of the whole process. For example, time is a key element for my customers. Are they interested in a short order cycle, in a stable one or in a later cutoff time?
3. to determine the exact values each market segment expects us to reach for these indicators. For example, field service engineers want to get the spare part overnight while distributors would be satisfied with receiving their orders within 7 days.
4. to determine the values we reach with the services we offer. These values will be checked against strategic stated goals, against market segment goals or against results obtained in previous periods. They will indicate how quickly improvement is being made and will allow us to check the fit between the demand and our service. For example, we could observe that a

given distributor is served 10 days after he places an order when he required a seven day service.

5. to determine the different contributions of our operations. These contributions will allow us to determine where to react if needed. For example, we could discover that the carrier delivering to this unhappy distributor takes 5 days to consolidate our shipments.

Steps 1 to 3 aim to define the precise objectives our customers expect us to reach. These are often called external indicators. Step 4 corresponds to the evaluation of our current services with respect to these objectives. Step 5 should provide us with the information necessary to enable improvement. These are often referred to as internal indicators.

3.1. Key elements and indicators

In Tables 3.1 to 3.3, we present different key elements and for each a list of common indicators, each with its strengths, weaknesses, and applications.

Table 3.1. Indicators for quality

Quality system	
What	How
Telephone traces or unanswered calls	% answers
Letters' demands answered during the day	% answers
Customer complaints reopened	level and intensity
Non-conforming product (trend - causes - quick response - % orders)	any product with one or more characteristics that depart from the specifications, drawing, or other approved product description
Incorrect shipment of goods and invoices	% orders
Documentation quality	

Table 3.2. Indicators for delivery

Rapid and consistent delivery	
What	How
On-Time Delivery (OTD)	% of orders, line items, total units or money volume delivered on the date the customer requested
X-day delivery	% of orders delivered within the X days
Frequency of delivery	number of deliveries per period of time
Order cycle time	elapsed time (days) from order to delivery
Order processing time	period of time required to administratively process a customer's order and make it ready for shipment
Order shipment time	days - weeks

To reach overall quality requires quality programs. This means a set of procedures, controls and built-in mechanisms aiming at solving the detected quality problems. This is why quality indicators as those listed above have to be recorded and archived. They will provide the information for corrective actions like complaint resolution.
The overall quality should also be reached at an adequate cost (fitness of cost). This means that inspections have to be balanced with built-in quality process. For example, with non-conforming products, customers may be lost. If the error can be corrected, it will generally be at a higher cost than if the checks had been built in. In the same way, wrong shipments can occur. Without sufficient checks, such a mistake could result in payment for undesired items or bad customer relationships. The problem may be corrected by instituting a check by a second material handling person. However, although the incorrect orders are detected and corrected, the process itself is faulty.

Time indicators will be discussed in details in Section 4. Many of the time indicators indirectly depend on the availability of the items that are ordered. It is therefore crucial to measure this availability for management purposes.

Table 3.3. Indicators for availability

Availability (% orders, % order lines, % items)	
What	How
Stock availability (%)	percentage of demand for a given line item that can be met from available inventory
Back orders by age	an aging of backorders (unfilled customer orders or commitments)
Product substitutions	
After-sales support	

Companies need to develop a set of service indicators appropriate to their own products and meaningful to their customers. The controlling factors in establishing the customer service objective for an item are the cost of carrying the item and the cost of a stock-out. These costs depend on different features such as unitary cost, sales volume, rotation time, criticality, supply lead time, etc. For example, inexpensive, easily stored products whose absence results in relatively high costs should have high customer service performance objectives. This means that measuring availability figures at the aggregate level is inadequate. This is illustrated by the example described in Section 4.3.

Cost indicators, both external and internal, are also needed. The external indicators refer to the cost of the service as seen by the customer. It can be the price he pays for the service or the conditions under which the service is free. The internal cost indicators are needed for budget control and for management purposes. They are addressed in detail in Section 4.2.

3.2. One significant element: On-time Delivery (OTD)

While all elements of delivery service are of potential importance, one particular element is increasingly seen as being a crucial source of differential advantage: On-Time Delivery (OTD). In this section and in the following section, we will use OTD to illustrate all the difficulties that occur when trying to define, measure and improve a precise service.

OTD is a measure of the success in meeting the customer request date. Performance is usually based on the percentage of orders, line items, total units or money volume that were actually delivered on the date requested.

Let us first consider OTD measured at the order level. An order is counted as being delivered on time if and only if *all the items* of that order are delivered on the date requested by the customer. The probability that it happens decreases as the number of lines on the order increases. This way of measuring OTD is appropriate for a field service engineer who places an order with all the items he requires to perform a repair. Indeed, if 20 different items are needed for the repair, the order should be counted as on time only if all 20 units are delivered on time. In this case, it makes sense to delay the whole shipment until all the items are available. Incentives for the distribution centre staff to improve OTD measured at the order level are, in this case, adequate. It is indeed better to serve 95 percent of such orders fully and to delay completely the other 5 percent rather than sending each customer 95 percent of what he ordered.

However, OTD measured at the order level may have drawbacks. First, the temptation to never serve an order that has become overdue is high. Such an order can no longer contribute positively to the OTD performance. Second, delivering a small or a big order on time has the same importance for the global OTD performance. It becomes easy to improve the OTD performance by systematically delaying the few big orders to satisfy the many small ones. These two drawbacks are classical pitfalls related to the use of a single performance measure. They can be alleviated by additional performance indicators. For the overdue orders, the following measures can be used.

1. The percentage of backorders shipped within different time periods.
2. The average time and the standard deviation of the time it takes to ship a backorder.
3. An aging of backorders with limits established as strategic customer service goals on the various brackets.

For differentiating big from small orders, OTD measured at the order level could be complemented with a measure at the line or at the item level. However, there are cases where OTD measured at the order level is simply inadequate.

When our customers are local distributors who place big replenishment orders, OTD measured at the order level is not appropriate. Indeed, if only 95 percent of the lines of an order can be delivered on time, the distributor is interested in getting them. He generally does not want the whole shipment to be delayed because of the missing items. In this case, it seems more appropriate to measure

OTD at the line level, that is as the number of *lines* that could be delivered on time. The objective should then be to deliver 95 percent of each order rather than 95 percent of the orders. This would introduce a greater equality among the customers.

Detailed analysis of OTD at the order level has shown two drawbacks: any two orders are identical and an overdue order could never be served. The same apply to OTD at the line level. For overdue lines, the use of the above additional aging measures can be adequate. If two order lines can have vastly different levels of importance, OTD could be measured at the item level or at the value (money) level.

These different levels could be interpreted as different weighting schemes. At the money level, we weight each item by its value. At the item level, we weight each line by the number of items, etc. Other weighting schemes could be designed. An example would be to weight each item according to its criticality. Note, however, that any weighting scheme introduces an additional complexity.

None of these methods for measuring OTD is right or wrong. Individual circumstances, or more precisely, individual customers will determine which is the most appropriate. Finally, to measure delivery performance against promised delivery dates without establishing why the problem exists is to understand only half of the picture. To establish the origin of an OTD problem, it is necessary to measure internal standards of performance, such as order processing, cycle, and shipment times. These times have to be accurately recorded and used to identify when changes in the process are required.

3.3. Gaps (see Kumar, Anil / Sharman, Graham (1992))

Zeithalm, Parasuraman and Berry cited gaps as potential reasons of service shortfalls (see Zeithaml, Valarie A. / Parasuraman, A. / Berry, Leonard L. (1990)). They identify four gaps:

- unknown customers' expectations
- inappropriate internal standards for service performance
- service performance gap
- unreliable deliveries.

All customer service measures are surrogates for how the customer rates the organisation's service. Gap problems may be defined as a difference between consumer expectations and his perceptions concerning customer service.

To avoid these problems, it is recommended that all managers visit customers to acquire this information firsthand, as it will be illustrated in the following point.

3.3.1. External causes

It is essential to first define OTD from the customer's viewpoint. After having segmented customers into broad groups with different distribution strategies, it may be useful to know each of the main customers' expectations to manage OTD individually for them.

The customer sometimes measures OTD differently from the supplier. It may be a question of reference base (total units shipped versus orders shipped) or of information (the same date is not used by the two parties). In this case, the first objective is to ensure that everybody has the same strategic OTD performance goals. Measure OTD and not a proxy evaluation. Be sure you measure the same standard in the same way as your vendors and your customers. This may be the first step to re-evaluate items in your inventory. For example, if the gap is due to the reference base, such as customers measuring the complete order and you measuring units shipped, it may be necessary to evaluate the proportion of less-frequently-ordered but typically out-of-stock items and keep an ample inventory for these items.

3.3.2. Internal causes

Variable supply lead times can decrease the OTD performance of the distribution centre. For example, some companies have a transport time, from the manufacturer to the European distribution centre, of about 4 weeks by ocean. The orders from the distribution centre must become definite at least 4 weeks before the boat leaves. Depending on the supplier, the lead time is between 8 and 12 weeks. The same is true for some sales zones.

Too many departments involved in the processing of an order or an unbalanced workload between the different operations (warehousing, shipping and customer service) may also cause an OTD gap. A pharmaceutical company solved the problem by placing all the operations of a particular order under the responsibility of one person. One person combines the complete process from the origin -- the placement of an order from the customer -- to the shipment of the products. This simplifies and speeds up information flow and co-ordination of activities.

Expectations of the customers in terms of OTD can vary with time and with the items ordered.

For seasonal products, in the automotive sector, such as air-conditioning systems in the summer season and body parts (fenders, bumpers, doors) in the autumn and winter season, you need to have accurate forecasts if your deliveries are to be as reliable as for standard products. A tightening up on reliability might reduce holding costs and still provide an acceptable level of customer service. In this case, our customer service department should be part of our forecasting process.

Similarly, OTD need not be uniform among all products. For example, high-volume products can be offered with greater certainty than low-volume products, as it is probably better to focus on the products accounting for a larger percentage of revenues.

4. Improving and (re-)evaluating the service mix

In this section, the selected service mix will be analysed in detail. In other words, it will be checked whether each service we offer really meets the corresponding market segments needs in terms of the various requirements. For practical reasons, we will focus on the time and the cost dimensions only. This analysis should also show where improvements are possible for each service.

Section 4.1 analyses the service structure. This allows a general cost analysis to be performed (Section 4.2). Section 4.3 checks how the emergency service agrees with its market segment. Section 4.4 deals similarly with the stock order service.

4.1. Decompose the physical distribution flow into activities

The physical distribution includes a set of activities related with the movement and storage of products, usually finished goods, merchandise or spare parts, from manufacturing to the end customer. In many cases, this movement is made through one or more levels of field warehouses. Refer to (see Bowersox / Closs (1996)) for a good reference on physical distribution operations.

The order preparation includes a set of activities relating to paperwork, picking, and packaging. Usually, this process includes several operations, done by different persons in different services. A customer order is accepted by the order taking department. This means accepting and translating what a customer wants into terms used by the distributor. The order then goes to the financial department where another person checks the credit allocated to the customer. Then the sales department determines the price to charge. The order is then transmitted to inventory control, which checks the product's availability. Then the warehouse defines a shipment program. The logistic department determines the mode of transport -- rail, road, air or water -- and chooses a carrier. In the warehouse, the items to be shipped are picked, packed and checked. The shipment is finally given to the carrier who executes the delivery. Obviously, some of these steps can be skipped or combined.

All the operations involved in the physical distribution can be organised into successive groups or activity centres:

1. Order processing
   The administrative activity required to process a customer's order and make it ready for shipment. This includes all the paper work associated with an order (receipt, customs, invoice...).
2. Inventory control and holding
   Inventory control encompasses the set of activities and techniques needed to maintain the desired levels of items.
3. Warehousing
   These are the activities, automated or manual, related to the storage of the goods in the warehouse. The typical sequence of operations is: unloading, unpacking, inspection and storage. Additional work like product relocation, scrapping and repackaging is also considered as part of warehousing operations.

4. Picking
   This is the process by which the ordered items are taken from their storage area and brought to the packing area.
5. Packing
   The packing function comprises all the operations needed to build a "shippable" parcel. It includes checking, wrapping and labelling of the box and printing the necessary transport documents.
6. Transport
   This is the set of operations a parcel undergoes from the supplier's location to the customer's receiving location.
   An operational unit can be a plant, a freight consolidation or "unconsolidation" point, a transhipment location or a store.
   Consolidation means that packages and lots that move from suppliers to a carrier terminal are sorted and then combined with similar shipments from other suppliers for travel to their final destination, the external customer. Freight consolidation is done to reduce transportation costs by means of a better utilisation of the transportation resources.
7. Delivery
   The process of delivering the consignment to the consignee at the agreed time and place is the last operation of the delivery service.

Each of the steps in the chain contributes to the global performance of the service. For example, transport induces delay in the order cycle, is a potential source of quality problems, and generates costs. The exact contribution of each activity centre in terms of time, money and quality must be clearly identified for management purposes.

The contribution of the different operations in terms of time and quality is rather straightforward. This is not the case with costs, as Subsection 4.2 addresses this question in detail.

To evaluate the service globally, the contribution of the different operations must be mutually compared. However, several options exist for each operation. Each option has its advantages and drawbacks. Therefore, whether an option is adequate or not can only be assessed in the frame of precise service objectives. In Subsection 4.3 and 4.4, we propose a rough scheme for the assessment of emergency orders and stock orders respectively.

4.2. Determine the cost contribution of each activity

A possible reason why distribution costs were neglected for decades is that distribution was considered as a cost centre and not a centre adding value. Across European and North American industry, it has been estimated that distribution costs typically vary between 5 and 10 per cent of sales (see Christopher (1992a), p.61). Often, the distribution remains susceptible to important cost reductions.
Although the importance of price declines in favour of service, the customer is not ready to pay an unlimited price in exchange for a top quality physical distribution.

The objective of controlling costs is to quickly inform the responsible managers when distribution strays from strategic decisions taken and from budget forecasts.

Thus, a good knowledge of the logistics related cost structure is of first importance. The introduction of a cost control function begins (see Cooper / Kaplan (1988), Johnson / Kaplan (1987)) with identification and accurate definition of all activity centres, which was done in the previous subsection. Let us review these centres and see how the costs evolve.

1. Order processing
   The administrative operations are either automated (on-line check/allocation/order) or performed by people. This could be considered a fixed cost per order or a fixed cost for the distribution centre.
2. Inventory holding
   Stored goods generate two kinds of costs: holding cost (deterioration, obsolescence, insurance) on the one hand and opportunity cost of the capital tied-up in inventory on the other hand. Both costs are generally expressed as some percentage of the stored value. For example, a 20 percent rate means that storing goods for one million Belgian Francs during one year generates a cost of two hundred thousand Belgian Francs.
   The investment in stored goods is mainly influenced by the procurement lead time and by the quality of the demand forecast.
   The performance of a company, in reference to inventory, is usually measured by the turnover ratio, which is calculated by dividing the cost of goods sold each year by the average value of the inventory. The larger the turnover ratio is, the smaller the holding cost will be. However, this ratio is often determined globally, hiding differences between parts.
3. Warehousing
   Warehousing costs include fixed costs such as warehouse and maintenance facilities, material handling and other equipments. They also incorporate all the in-bound operations (unloading, unpacking, inspection and shelving) which are proportional to the number of processed items.
4. Picking
   The picking operations are labour intensive. The amount of work is mainly proportional to the number of lines processed.
5. Packing
   Packing is proportional to the number of orders processed and to a smaller extent to the number of lines.
6. Transport
   The transport costs are mainly a function of the transportation mode or combinations of modes (air, truck, rail or sea). They also depend on the volume or weight carried and the final destination.

These activity centres are not equally demanding for money. Three main areas can be identified globally: the inventory costs; the labour costs (mainly related to the in- and out-bound operations) and the transport costs. Cost reductions can be

obtained by optimising each area separately or together. Tables 4.1 and 4.2 consider these alternatives.

Table 4.1. Individual cost optimisations

Optimise separately:	By:
the inventory costs	using better forecasts, better communication of information, shorter lead times, shorter inbound operations.
the handling costs (see Askin / Standridge (1993))	using high tech equipment (economy of scale), optimised order sequences (by sequencing and grouping the orders to be picked in order to optimise handling), optimised warehouse organisation (by modifying the item locations according to the picking frequency), optimised order composition (by informing the customer of the consequences of the way he orders).
the transport costs	using the right transport mode.

Table 4.2. Concurrent cost optimisations

Optimise jointly	How:
inventory costs and transport costs	This is the classical case (see Cohen / Kleindorfer / Lee (1988)) where transport costs are saved by storing at lower levels the items that have the largest demand. A European DC will typically ask its national distributors to store essential A parts and to serve the "down unit" segment from its inventory. Similarly, if there is a higher level, parts that are very expensive and slow moving could be stored at that level only.
inventory costs and handling costs	Shouldn't we systematically pick, pack and send a complete box? The idea is to avoid unpacking and repacking as much as possible. If some parts are subject to different demands, we could have several storage locations for these parts to improve the handling of these orders.
handling costs and transport costs	Is it cheaper to store the units where they are manufactured, allowing them be sent from there to the customer? Can we avoid unnecessary handling in central warehouses at the cost of less optimised transport?

All of these cost optimisations can be considered independently of the market segment that is served as long as the cost reductions do not alter the performance of the service. However, not all of the options meet the service requirements of a market segment. This means that the exact alternatives that need to be compared

depend on the kind of service that is considered. Furthermore, these options do often alter other service performances. That is why most optimisations have to be considered in the frame of a specific service. This is illustrated in the following sections by means of two examples.

4.3. Example 1: the "down unit" segment

As a first example, we chose the market segment composed of customers requiring essential parts for their equipment to work. We called this segment the "down unit" segment. These customers require direct deliveries overnight or at least within 24 hours. Their main concern is about the speed of the delivery. By definition, this segment focuses on the parts that are essential for the units to work. We should therefore keep in mind that this segment relates to a subset of the products that are distributed. Table 4.3 outlines the characteristics of this segment.

Table 4.3. Characteristics of the "down unit" segment

Size of market segment:	in orders and in sales
Major requirements:	daily orders
	direct delivery
	next day delivery
Minor requirements	cutoff time: as late as possible
	stock availability
	cost
Parts:	only essential parts (no accessories)
Served by:	emergency orders

The analysis will now aim to study the performance of the service offered to this market segment (here, the emergency orders). For conciseness, we will mainly focus on two characteristics: the time and the cost dimension.

4.3.1. Time performance

This market segment is supposed to be served by means of emergency orders. Table 4.4 gives an example of how this service could be implemented.

Table 4.4. Time decomposition in the "down unit" segment

Activities	duration	schedule
Cutoff time		12.30
Paperwork	1/2 hour	13.00
Planning	1/2 hour	13.30
Picking and packing	3 hours	16.30
Loading:	1/2 hour	17.00
Departure time of express carrier		17.00
Delivery to the customer	overnight	9.00

The cutoff time is usually determined by rolling all the scheduled operations backwards from the transport departure time. Then, the question is whether or not the expectations of the MS for time performance are met.

If the cutoff time (12:30 in this example) is too early, changes in our process become necessary. Here are some possibilities.

- *reduction of the paperwork time* by automating the process completely from the placement of the order through to the printing of the picking order, for example.
- *reduction of the picking time.* Several ways are possible. The sequence of the items to be picked can be optimised. The physical location of the parts susceptible to emergency orders could also be optimised. Eventually, a separate inventory could be created for these parts.
- *reduction of the total flow time by overlapping the operations.* Instead of performing in batch the administration, planning, picking, packing and the loading operations, an overlapped schedule can be used. This means that the emergency orders would be processed at the time they are placed and stored in a separate area. Of course, this overlap reduces the opportunities for optimising picking and packing. On the other hand, the last emergency order could be placed much later.
- *delaying the departure time.* In some cases, the time limit for transport departure has been set at the end of the last working shift when the real time limit is much later. Planning a shift that takes advantage of the real limit would be a simple solution in such cases. A much more difficult solution consists of moving closer to the express carrier hub.

Cutoff time is not the only important timing aspect. Because of bottlenecks, inefficient processes, and fluctuations in the volume of orders handled, there can be considerable variations in the time taken for the set of activities to be completed. The overall effect can lead to a substantial reduction in delivery reliability. Some emergency orders could miss the transport departure.

4.3.2. Stock availability performance

Here, we wish to emphasise that the expectations of this segment in terms of stock availability could be very different from other segments. If all the parts that are distributed are required by all the segments, the toughest expectations should be met. However, in our example, the "down unit" segment focuses on essential parts only. One could therefore imagine different objectives for stock availability according to the criticality of the parts.

4.3.3. Cost performance

All the cost optimisations mentioned in section 4.2 are of course applicable. However, within this emergency service, the transport operation is implemented by means of an express carrier that guarantees overnight delivery. The cost contribution of the transport operation will therefore be the dominant factor of the service cost.

As long as the transport cost remains rather insensitive to the distance, more centralised solutions can be considered to save handling and/or inventory cost (refer to the last two opportunities for reducing cost listed in table Y).

The question here is whether the expectations in terms of costs of the "down units" customers (or mine, if I pay the costs) are satisfied. Do my customers agree with the costs they pay or, if I pay the costs, is my service globally profitable? Are our customers prepared to give up some requirements for some other advantages? For example, is overnight delivery a need or can the customer in this MS wait for delivery until the next day at noon? Most often, the "down unit" customers cannot.

However, if they can, their order could be piggybacked on a truck that would reach its destination only a couple of hours (or one day) later depending on the necessity for "unconsolidation" (do not forget that express carrier guarantees direct delivery while the truck will most likely travel to a break bulk point). This piggybacking could be performed by different means:

- On our own trucks;
- On the trucks of neighbouring companies with whom we could collaborate;
- On the trucks of specialised companies to whom we could subcontract the whole market segment.

In many companies, this kind of solution is actually implemented. It is called the "express order" service. However, it is not really meant for the "down units" segment.

4.4. Example 2: the "distributor" segment

Table 4.5 outlines the characteristics of this segment.

Table 4.5. Characteristics of the "distributor" segment

Size of market segment:	in orders and in sales
Major requirements:	low cost
	frequency: daily if possible
Minor requirements	cutoff time: as late as possible
	delivery time: fixed known delay
	fill rate
Parts:	all parts
Served by:	stock orders

4.4.1. Time performance

This market segment is supposed to be served by stock orders. Table 4.6 shows a generic example of how this service could be implemented.

Again, the time performance must be reviewed by comparison with the expectations of the MS. The two main time aspects here are the length of the order cycle and its variability.

Table 4.6. Time decomposition in the "distributor" segment

Activities	duration
Paperwork	(hours) negligible
Planning	(hours) negligible
Picking and packing	1 day
Loading	(hours) negligible
Consolidation	1, 2 or 3 days
International transport	1 day
Local Distribution	1 or 2 days
Total (Order cycle time)	between 4 and 7 days

Here is a rough list of possibilities for shortening the order cycle:

- reduce picking time or do not perform the picking in batch;
- do not consolidate;
- reduce consolidation time;
- speed up transport time (by a better transport synchronisation).

Note that a reduction of the consolidation time would also reduce the variability of the length of the service.

However, since the main requirement of the segment is low cost, we should consider time reductions that do not increase the total cost. In other words, we should find a solution that reduces the time by relaxing other requirements.

In this case, we could reduce the consolidation time by reducing the frequency of the deliveries. By allowing the trucks to leave for example on Tuesday and on Friday only, we implicitly define an internal consolidation time. However, the consolidation occurs within the DC with our products. This means that the customers can still place orders during this internal consolidation time. Eventually, the customers can only place their orders on the last day. This leads to the notion of scheduled orders.

The selection of the precise days for the scheduled deliveries will most often be a result of general considerations such as warehouse workload balancing and/or "commonality" of destinations (e.g., serve all the Nordic countries together).

If you cannot consolidate internally, external consolidation is also possible. Eventually, the whole market could be subcontracted to a specialised company that is able to perform this consolidation.

4.4.2. Cost performance

We assumed that low cost was the main requirement of this market segment. Let us see how and to what extent this objective can be achieved. Here below we review the different cost areas and comment on them.

The holding costs are relatively small for this segment. Indeed, the number of transactions in this segment is rather high and inventory value increases at a slower rate relative to volume. Only expensive, slow moving parts could be a problem.

The picking and packing costs are proportional to the number of order lines and orders, respectively

The picking costs are proportional to the average time for picking. This time can be optimised by means of equipment, warehouse organisation and picking order optimisation. This last point favours a batch organisation of the picking in order to allow this picking to be optimised.

The picking costs are also proportional to the number of order lines. Picking a can twice takes more time than picking two cans once. The customers must be aware of this fact. If they really want a cheap service, they must be aware of the extra processing time they generate by ordering small quantities. In any case, they should be followed and measured according to the way they order. Different counter-measures are possible. Bulk units that cannot be split could be defined. Small orders could be systematically rounded up. Feedback on the package size could be given.

Packing costs and, to a smaller extent, transportation costs are proportional to the number of orders placed. Splitting an order generates additional packing and transport costs. Again, the customers must be aware of this fact and some incentive should exist for grouping the orders.

Compared to the "down units" segment, the transport costs for the "distributor" segment get relatively small since bulk transportation means are used. However, this remains valid only as long as complete truck loads (TL) are used. This requires internal or external consolidation.

To summarise, we point out the following tendencies that aim to keep the service (stock order) in line with expectations (low cost):

- keep transport costs low by using internal or external consolidation. Internal consolidation leads to the notion of scheduled order which globally reduces the order cycle time at the cost of reduced delivery frequency;
- keep the labour costs low by providing incentives for orders with few lines and high quantity per line (complete pallets, for examples). Some companies do offer free or discounted stock orders if its value exceeds some threshold. However, a high value order does not always correspond to an order with few lines and high quantity per line.

5. Reconsider the whole process

Up to now, we considered our service as the centre of a big picture composed of our customers, our competitors, our strategy and our operations. Inside this picture, we described how a continuous improvement loop could be initiated and maintained.

The scope of this loop must be extended beyond this picture if it is to remain credible. The organisations on both sides of the distribution function must be integrated in the loop: the companies that manufacture the products we deliver on one side and the customers' organisations on the other. Such considerations invoke the notion of integrated supply chain (see Bowersox / Closs (1996), Christopher (1992a), Christopher (1992b), European Logistics Association (1994)).

5.1. The company's side

Two departments require especially tight contact with the logistics function: the manufacturing department and the marketing department.

Our influence on both results directly from our contact with customers. We know what is being sold and we know our customers' concern. A close collaboration with the production department can lead to the following advantages:

- Permanently informing the production department of the amount of sales helps planning for future production schedules. This avoids the potential negative consequences of the lot sizing performed by the distribution centre. Finally, it will improve the fill rate of replenishment orders we place to the production plants.
- A good knowledge of the production plans allows us to inform customers of the exact delivery dates in case of stock-out.
- The direct impact of the production cycle time on safety stock is an opportunity for joint cost optimisation. Indeed, any reduction of the production cycle time often generates extra costs. However, these costs can be at least partly compensated by a reduction of the safety stock in the warehouse.

The marketing department should also be informed about sales and our customers' concerns. This information could be used to launch appropriate promotional campaigns. Programs for the introduction of new products also require this information.

Symmetrically, an information flow from the marketing department to the distribution function must also exist. Distribution must be aware of future promotional campaigns and new product launches. This will help to plan future demand and avoid excessive shortages. On the other hand, safety stock for products that will soon be obsolete can be reduced in order to reduce scrap.

Beside the exchange of information, joint cost optimisations are also possible. These exist at different levels.

- At the level of the product palette definition, any standardisation of the products lead to a reduction in the number of parts to be stored and therefore in a reduction of inventory costs.
- At the level of the service performance selection, marketing must be clearly aware of the strengths and weaknesses of the distribution function. A complete geographical market segment could be abandoned if the logistics is not able to offer a decent service at a decent price.
- At the level of the product, packaging aspects could help or plague the handling of the products.

5.2. The customer side

Throughout this paper, the customers' requirements in terms of service performance were accepted without discussion. The question we raise now is whether these requirements are always justified or not. In other words, are there economical opportunities of relaxing these requirements? A typical example is the introduction of a common information system that allows an order to be made earlier and a cheaper delivery process to be used.

5.3. Information systems

All of these examples began by showing that adequate communication of information is crucial. Another example is the information related to the state of the warehouse. A physically or administratively overloaded system most often reduces customer service. For example, exceeding storage capacity may result in some of the following: weather damage due to outside storage; misplacement of lots due to the impossibility of using the normal storage locations; item damage due to aisle storage.

The use of advanced computer-to-computer order entry may provide the required integration between marketing, production, the distribution functions and the customers. It enables the customer to order directly and to obtain the price, the date it will be shipped and the expected date he will receive it.

Nevertheless, increased communication is not appropriate in all cases. Where time, service and distribution costs are critical, the potential is great for increased use of telephone and other on-line systems.

6. Conclusions

In this paper, we focused on the delivery service distribution centres offer. We did not try to determine whether a given service was good or bad, but rather we try to define a frame and a set of systematic questions that should be raised in order to evaluate whether the service is adequate or not.
This systematic evaluation could be roughly summarised by considering Table 6.1 that puts the delivery service in its environment.

Table 6.1. Delivery service and other functions

Function:	Distribution	↔	Production	↔	Marketing	↔	...
	⇩ ⇘						
Service:	**Delivery**	↔	After sales	↔	Information	↔	...
	⇩ ⇘						
Service element:	Emergency order	↔	Scheduled order	↔	Stock order	↔	...
	⇩ ⇘						
Operations:	Handling	↔	Inventory	↔	Transport	↔	...

In this table, the delivery service is depicted as one of the services offered by the distribution centre which in turn is one department or function of a company. In the other direction, the delivery service can be decomposed into different service elements (orders) which in turn rely on different operations. If you consider all the elements of this table, a systematic evaluation of the delivery service requires the raising of all the questions of the form:

"Is each element optimised by itself and are the elements of a same level balanced and co-ordinated?"

We also try to show in this paper that the selection of the goals for each element and for each level is a complex process that can only be addressed by reference to the real market of the company.

Acknowledgement

We would like first to thank the managers of the different companies we visited. They all opened their doors to us and gave us part of their valuable time. Without their patience, this work would not have been possible.

We are also deeply grateful to several colleagues, especially C. Delporte, Y. de Rongé and M.-P. Kestemont for their help and comments on different parts and versions of this work, as well as P. Hyden who corrected our English.

References

Askin, G. / Standridge, Charles R. (1993): Modeling and Analysis of Manufacturing Systems. John Wiley & Sons, Inc.

A.T.-Kearney (1993): Logistics Excellence in Europe. A. T. Kearney, Inc.

Bowersox, Donald J. / Closs, David J. (1996): Logistical Management -- The Integrated Supply Chain Process. The McGraw-Hill Companies, Inc.

Cohen, M.A. / Kleindorfer, P.R. / Lee, H.L. (1988): Service constrained (s,S) inventory systems with priority demand classes and lost sales. in Management Sci. Vol. 34 No. 4, 482-499

Cohen, M.A. / Lee, H.L. (1990): Out of touch with customer needs ? Spare parts and after sales service. Sloan Management Review, 55 Winter, 55-66

Christopher, M. e.a. (1979): Customer service and distribution strategy. Associated Business Press, London

Christopher, M. (1985): The Strategy of Distribution Management. Gower Press, London

(1992a): Logistics and Supply Chain Management -- Strategies for Reducing Costs and Improving Services. Pitman Publishing, London

(1992b): Logistics - The strategic issues. Chapman & Hall, London

Cooper, R. / Kaplan, R.S. (1988): Measure Costs: Make the Rights Decisions. Harvard Business Review, septembre-octobre, 96-103

European Logistics Association (ELA) (1994): Terminology in Logistics - Terms and Definitions. 2d edition, Brussels

Johnson, H.T. / Kaplan R.S. (1987): Relevance Lost: The Rise and Fall of Management Accounting. Harvard Business School Press, Boston

Kumar, Anil / Sharman, Graham (1992): We love your product, but where is it? The McKinsey Quarterly, Number 1, 24-44

Lambin, Jean-Jacques (1993): Strategic Marketing -- A European Perspective. McGraw-Hill Book Company

Pfohl, H.-Chr. (Ed.) (1994): Future Developments in Logistics and the Resultant Consequences for Logistics Education and Training in Europe. Logistics Educators Conference 1994. European Logistics Association, Brusssels

Ploos van Amstel, M.J. / Ploos van Amstel, W. (1987): Economic Trade-Offs in Physical Distribution - A pragmatic Approach. International Journal of Physical Distribution and Materials Management, Volume 17, n° 7

Robeson, James F. (1985): The distribution handbook. The free press, Macmillan, Inc., New York

Sharman, G. (1989): What 1992 means for logistics. The Mc Kinsey Quarterly, Mc Kinsey & Company, New York, Spring

Shiba, Shoji / Graham, Alan / Walden, David (1993): A New American TQM - - Four Practical Revolutions In Management. Productivity Press/The Center For Quality Management

Simmons, G. / D. Steeple (1991): Overhead Recovery - It's Easy as ABC. Focus: on Logistics and Distribution Management, October 1991, Volume 10 number 8, 24-27

Touche, Ross / Institute of Logistics (1995): European logistics comparative costs and practices. European Logistics Association. Brussels

Zeithaml, Valarie A. / Parasuraman, A. / Berry, Leonard L. (1990): Delivering Quality Service. The Free Press

Chapter 2

Warehouse Location and Network Design

A Local Search Heuristic for the Two-Stage Capacitated Facility Location Problem

Arno Bruns

Universität St. Gallen, 9000 St. Gallen, Switzerland

Abstract. The *Two Stage Capacitated Facility Location Problem* is the problem of finding the locations of depots from a set of potential depot sites to satisfy given customer demands, the assignment of customers to the selected depots and the product flow from the plants to the depots such that the total system cost is minimized. The total demand of all customers assigned to any depot can not exceed the capacity of that depot and the shipments from any plant to a depot must not exceed a given supply capacity of that plant.

A new local search heuristic based on well-known drop and interchange techniques is presented. A first feasible solution is constructed using a modified version of the "drop"-approach. This solution is improved by a descent type local search on the set of open depots. The procedure has been tested on a set of randomly generated problems. Compared to lower bounds computed by a general mixed-integer programming solver, the results obtained were less than [%2.4] from these lower bounds.

1 Introduction

Distribution logistics is concerned with the provision of goods from one or more supply points to a number of demand points which are separated by some distance. The design of the distribution network determines the flow of these goods from the supply to the demand points. The selection of a distribution system is one of the major entrepreneurial decisions and plays a key role for the success of an enterprise. An inadequate distribution system causes high costs, and changes to the system are often costly and time consuming.

In general, a group of decision makers is involved in designing or redesigning a distribution system. Because of the inherent political nature, it is necessary to substantiate the arguments for one or the other location by objective decision rules. Modelling the problem and performing an analysis based on such a model helps to reduce the subjectivity of the decisions.

A model which can be applied to a wide range of problems arising in the design or redesign of a distribution system is the so-called *Two Stage Capacitated Facility Location Problem* (TSCFLP). The new heuristic procedure presented in this paper aims to render good solutions for such type of problems with short computational time.

This paper is organized as follows. First, the problem and its mathematical formulation is described. Then, a detailed description of a new heuristic solution procedure is given. Finally, some computational results are reported.

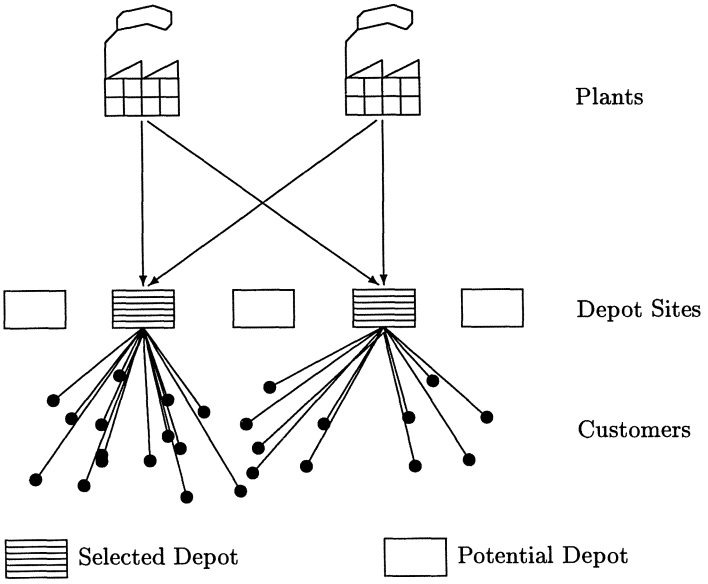


Fig. 1. Typical two stage distribution system

2 Problem Formulation

A number of plants supply goods in order to satisfy the demand of a certain number of customers located in a geographical area. The goods are shipped from the plants to the customers through some intermediary facilities. These facilities, here called depots, break the inflow into smaller consignments, which are shipped directly to the customers. The underlying problem is shown in Fig. 1 and consists of three layers and two stages.

Solving the TSCFLP consists of finding the number and the locations of depots from a set of potential depot sites, determining the allocation of the customers to the depots and the amounts transported from the plants to the depots, such that the total system cost is minimized.

Furthermore, the following assumptions are made:

- Every customer can only be served by exactly one depot.
- All time dependent parameters refer to the same time period.
- The locations of the plants, the potential depot sites and the customers are known.
- The supply capacities of the plants, the transshipment capacities of the depot sites and the demand of the customers are also known.
- The sum of the customer demands does not exceed the sum of the plant capacities.
- The sum of the throughput capacities of all the potential depot sites is not less than the total demand.

- The demand of the customers is produced (resp. supplied) and transported in the same period.
- Transportation costs on the first distribution stage, between plants and depots, are linear functions of the amount transported.
- Depot costs are linear functions of the throughput per period.
- Products can be aggregated to one product group.

The cost components included in this model are:

- the manufacturing costs at the plants as long as they can be assigned directly to the product group,
- the transportation costs between plants and depot sites (if the manufacturing costs are included in the model, they are added directly to the transportation costs),
- the fixed costs of operating a depot at the potential depot sites,
- the throughput cost per unit for every depot site,
- the costs of satisfying the demand of the customers from the potential depot sites; these costs must not be linear in distances or transportation quantities.

Mathematically, the problem can be formulated as follows:

$$v_{\mathrm{TSCFLP}} = \min \left(\sum_{p=1}^{P} \sum_{j=1}^{J} t_{pj} \cdot z_{pj} + \sum_{i=1}^{I} \sum_{j=1}^{J} (c_{ij} + \rho_j \cdot b_i) \cdot x_{ij} + \sum_{j=1}^{J} f_j \cdot y_j \right) \tag{1}$$

subject to:

$$\sum_{j=1}^{J} x_{ij} = 1 \qquad \forall\, i \tag{2}$$

$$\sum_{i=1}^{I} b_i \cdot x_{ij} \leq s_j \cdot y_j \qquad \forall\, j \tag{3}$$

$$\sum_{j=1}^{J} z_{pj} \leq \sigma_p \qquad \forall\, p \tag{4}$$

$$\sum_{p=1}^{P} z_{pj} = \sum_{i=1}^{I} b_i \cdot x_{ij} \qquad \forall\, j \tag{5}$$

$$x_{ij} \leq y_j \qquad \forall\, i,j \tag{6}$$

$$z_{pj} \geq 0 \qquad \forall\, p,j \tag{7}$$

$$y_j,\ x_{ij} \in \{0, 1\} \qquad \forall\, i,j \tag{8}$$

where:

b_i	is the demand of customer i,
c_{ij}	is the cost of delivering b_i units from depot site j to customer i,
f_j	denotes the fixed operating cost of a depot at site j,
p, j, i	are the subscripts of the plants, the potential depot sites and the customers, resp.,

P, J, I	are the number of plants, potential depot sites and customers, resp.,
s_j	is the throughput capacity of depot site j,
t_{pj}	is the cost of transporting one unit from plant p to depot site j,
v_{TSCFLP}	is the minimum total cost of the distribution system,
ρ_j	is the throughput cost per unit at site j,
σ_p	is the capacity of plant p,
x_{ij}	is a binary variable indicating if customer i is assigned to depot j or not,
y_j	is a binary variable indicating if the depot at site j is operated or not, and
z_{pj}	is a variable indicating the amount transported from plant p to depot j.

The objective function is composed of three cost terms. The first one is the total cost of the first transportation stage. The second term represents the costs associated with the assignments of the customers to the depots and the third term is the sum of the fixed operating costs.

The first set of constraints ensures that the demand of every customer is satisfied by the depots. Restrictions (3) limit the throughput at each depot site. If the depot site is not operated, the throughput has to be zero. The amount supplied by each plant may not exceed the capacity of that plant. This is specified by (4). Restrictions (5) are the so called "flow conservation" constraints, i.e. the inflow into each depot must equal the outflow. Since the condition, that each customer can only be assigned to an operating depot, can be deduced from, constraints (6) are actually superfluous. But, in order to reduce the solution space of the relaxed problem during a branch-and-bound search, these restrictions are introduced. Constraints (7) and (8) specify the nature of the variables.

There is a vast literature on discrete facility location problems, and various heuristic and exact solution methods have been proposed. The different approaches can be divided into three classes: primal heuristics, dual-based or Lagrangean heuristics, and exact methods.

Primal heuristics aim at computing an upper bound by first constructing a feasible solution, mostly in a greedy-like manner, and applying eventually local search for improvement. Greedy heuristics for the "Capacitated Facility Location Problem (CFLP)" (the TSCFLP reduces to an CFLP if $\sigma_p \geq \sum_i b_i \quad \forall p$) have been used since the 60ies. Kuehn and Hamburger started in 1963 with the proposal of a heuristic called "add procedure". Soon after, Feldman et al. (1966) presented the "drop heuristic", which is based on a similar idea. A summary of heuristics for the CFLP is given by Jacobsen (1983) and Domschke and Drexl (1985). The worst case behaviour of these heuristics with respect to the "Uncapacitated Facility Location Problem (UFLP)" has been analysed by Cornuejols et al. (1997).

Dual-based or Lagrangean heuristics use the information of a relaxation to obtain lower bounds and to construct feasible solutions from a solution

of the relaxation. Lagrangean relaxation heuristics to the CFLP have been proposed e.g. by Beasley (1988), Sridharan (1991, 1993) and by Klincewicz and Luss (1986). Beasley (1993) gives an overview on these approaches. Cornuejols et al. (1991) compare different Lagrangean relaxations for the CFLP. Klose (1997) presents a Lagrangean heuristic for the TSCFLP, which renders solutions close to optimality and is able to generate sharp lower bounds.

Exact solution procedures are based on branch-and-bound techniques, decomposition and/or branch-and-cut methods. Geoffrion and Graves (1974) use Benders decomposition for the exact solution of a multicommodity version of the TSCFLP. They use a "flow formulation" by introducing variables x_{pjil}, which denote the amount of commodity l shipped from plant p through a depot at site j to customer i. Hindi and Basta (1994) present a branch-and-bound solution technique for the TSCFLP without single sourcing. A branch-and-bound algorithm for a two-stage problem (with plant location) has been proposed by Kaufman et al. (1977). Van Roy (1986) uses cross decomposition to solve the CFLP to optimality. Van Roy (1989) also solves a relatively large fixed charge network flow problem – a generalization of the TSCFLP – using a general purpose MIP-solver based on automatic reformulation and branch-and-cut. Facets and valid inequalities, which can be the basis of a branch-and-cut algorithm for the CFLP and the TSCFLP, have been found by Aardal et al. (1995).

Exact methods and Lagrangean heuristics have been applied successfully to "easy" facility location problems like the uncapacitated or the capacitated facility location problem. The added complexity of the TSCFLP raises the need for other heuristics. Local search based primal heuristics have been successfull in solving other "difficult" combinatorial problems like the vehicle routing problem and therefore it was expected that they perform well on the TSCFLP too.

3 The Local Search Heuristic

In this section, the basic idea and the different phases of the algorithm are presented.

3.1 Basic Idea

A common way to solve complex problems heuristically is to decompose them into subproblems and to solve those repeatedly. This strategy is also followed here. Thereby, the core of the heuristic is the fast solution of these subproblems. At the end of the algorithm, a feasible solution to the TSCFLP is constructed by combining the results of the subproblems. New in this heuristic approach is the sequencing, adaptation, simplification and the combination of solution concepts that are already available in literature.

First, the TSCFLP is reduced to a CFLP with single sourcing by eliminating the first distribution stage and adding an estimate $\tilde{t}_{ij}$ of the cost to

transport customer i's demand from the plants to depot j to the costs of serving customer i from depot j:

$$v_{\text{CFLP}} = \min \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \check{c}_{ij} \cdot x_{ij} + \sum_{j=1}^{J} f_j \cdot y_j \right) \tag{9}$$

subject to constraints (2), (3), (6) and (8), where $\check{c}_{ij} = \tilde{t}_{ij} + c_{ij} + \rho_j \cdot b_i$.

Then, the CFLP is split into two subproblems. Solving the first subproblem, the selection of the number and location of the open depots, is performed with the help of a new drop-approach and a local search technique. The drop-approach is used in a first phase – called construction phase – to obtain an estimate of the number of open depots in good solutions. The local search part – called improvement phase – increases or decreases the number of open depots and/or interchanges depot sites.

Once the set M of open depots is selected, the CFLP can be reduced to a "Generalized Assignment Problem (GAP)":

$$v_{\text{GAP}} = \min \left(\sum_{i=1}^{I} \sum_{j \in M} \tilde{c}_{ij} \cdot x_{ij} \right) \tag{10}$$

subject to

$$\sum_{j \in M} x_{ij} = 1 \qquad \forall\, i \tag{11}$$

$$\sum_{i=1}^{I} b_i \cdot x_{ij} \leq s_j \qquad \forall\, j \in M \tag{12}$$

$$x_{ij} \in \{0, 1\} \quad \forall\, i \text{ and } j \in M \tag{13}$$

where $\tilde{c}_{ij}$ is composed of $\check{c}_{ij}$ and an estimate $\tilde{f}_{ij}$ of customers i's share of the fixed operating cost of depot j.

This second subproblem is solved heuristically by assigning the customers to the open depots at lowest cost, regardless of any capacity constraint, and then rearrange them in order to get a feasible solution for the GAP and for the CFLP. Feasibility was always obtained but the solution quality for the GAP was not monitored. This process of solving the two subproblems and combining the solutions is repeated for every set of open depots obtained by the drop-approach or the local search part.

In the last step of the algorithm, the best solution of the CFLP is used to construct a solution to the TSCFLP: First the throughputs

$$\beta_j = \sum_{i=1}^{I} b_i \cdot x_{ij}$$

at the depot sites are computed. Then the solution of the first stage "Transportation Problem (TP)":

$$v_{\mathrm{TP}} = \min \left(\sum_{p=1}^{P} \sum_{j \in M} t_{pj} \cdot z_{pj} \right) \tag{14}$$

subject to:

$$\sum_{p=1}^{P} z_{pj} = \beta_j \quad \forall\, j \tag{15}$$

$$\sum_{j \in M} z_{pj} \leq \sigma_p \quad \forall\, j \tag{16}$$

$$z_{pj} \geq 0 \quad \forall\, p, j \tag{17}$$

is combined with the solution of the CFLP to form a feasible solution to the TSCFLP. An overview about the decomposition and the general proceding is given in Fig. 2.

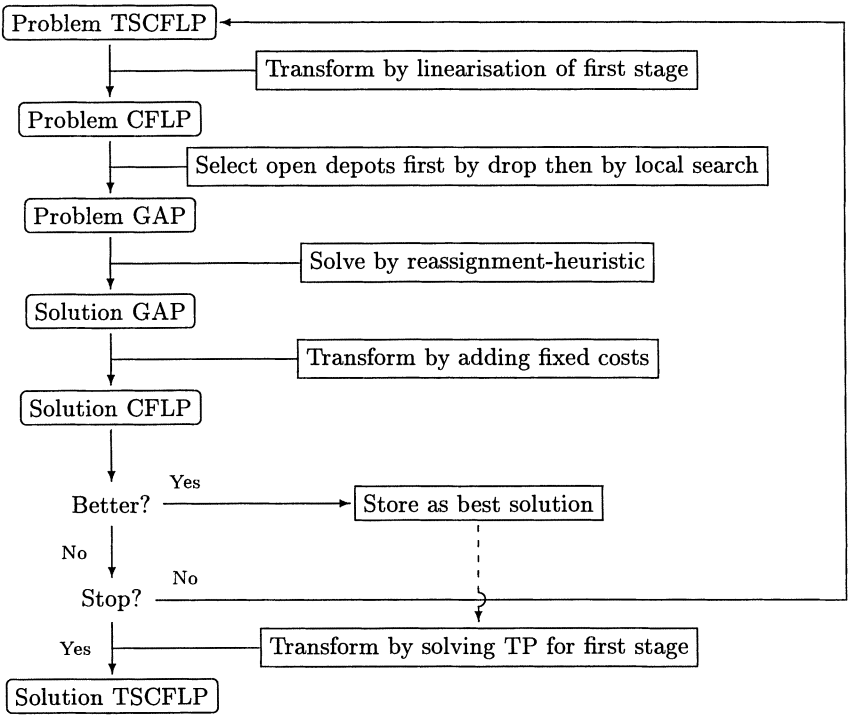


Fig. 2. Solution process of the new heuristic

The heuristic procedure presented here is refered to as "Drop-Interchange-Heuristic (DIH)". Taking into account some preliminary calculations, the DIH

can be divided into four parts: initialisation, construction and improvement phase as well as combining results.

3.2 Initialisation Phase

All calculations of the Drop-Interchange-Heuristic are based on a single cost matrix $\tilde{c}_{ij}$. The coefficient $\tilde{c}_{ij}$ estimates the total cost of serving customer i from a not specified plant through depot site j. It includes an estimate $\tilde{t}_{ij}$ of the first stage transportation cost, a share $\tilde{f}_{ij}$ of the fixed operating cost f_j of that depot, the assignment cost c_{ij} and the corresponding throughput cost $(\rho_j \cdot b_i)$. The cost c_{ij} of serving the customer i from the depot j and the throughput cost $(\rho_j \cdot b_i)$ are known, but a customer's share of the fixed costs and the transportation costs of the first stage has to be estimated.

These two cost terms are estimated as follows:

1. Customers i's share $\tilde{f}_{ij}$ of the fixed cost of depot j is estimated as:

$$\tilde{f}_{ij} = \frac{f_j}{s_j} \cdot b_i \qquad \forall i,\, j.$$

   This estimation is adequate if the depots are operated close to their capacity limits, which should be valid in most cases for near optimal solutions.

2. The estimate $\tilde{t}_{ij}$ of the first stage transportation cost of serving customer i by depot j is more difficult, because the plant serving the depot is not known in advance.
   For each depot, the minimum and the maximum transportation cost $t_j^{\min}$ and $t_j^{\max}$ per unit,

$$t_j^{\min} = \min_p t_{pj} \quad \text{and} \quad t_j^{\max} = \max_p t_{pj} \qquad \forall j$$

   are computed first.
   The estimation of $\tilde{t}_{ij}$ is then performed by adding a correction term t_j^{cor} to the minimum transportation cost per unit and multiplying the sum by the demand b_i of customer i:

$$\tilde{t}_{ij} = \left(t_j^{\min} + t_j^{\text{cor}}\right) \cdot b_i \qquad \forall i,\, j,$$

   where:

$$t_j^{\text{cor}} = \frac{t_j^{\max} - t_j^{\min}}{2} \cdot \frac{\sum_{i=1}^{I} b_i}{\sum_{p=1}^{P} \sigma_p} \cdot \frac{\min\{P,\, J\}}{J} \cdot \left(1 - \frac{s_j}{\sum_{j=1}^{J} s_j}\right) \qquad \forall j.$$

The correction term t_j^{cor} is influenced by different factors:

- An upper bound for this correction term is set to half the difference of the maximum and minimum transportation cost per unit. This value is a rule of thumb which was obtained by testing.
- If the total plant capacity is large (relative to the total demand) then the correction term is decreased. Large plant capacities imply that the depots can be assigned easier to the plants at lowest transportation cost.
- The production capacity constraints are more likely to be restrictive if the number of plants is large relative to the number of depot sites.
- If the distribution of the depot capacity is skewed then the depots with a high capacity are more likely to be assigned to plants at low transportation cost per unit than depots with a very small capacity.

While this type of estimation is quite arbitrary, testing has shown that the total estimated transportation cost on the first stage was always close (less than [%5]) to the one computed by an optimal algorithm.

With the estimate of the share of each customer on the fixed costs and the estimate of the first stage transportation costs, the elements of the single cost matrix can now be calculated as follows:

$$\tilde{c}_{ij} = \tilde{t}_{ij} + c_{ij} + \rho_j \cdot b_i + \tilde{f}_{ij} \qquad \forall i, j.$$

3.3 Construction Phase

The aim of the construction phase is, as mentioned earlier, to obtain an estimate of the number of depots used in good solutions of the CFLP and the TSCFLP. In the first step of this phase, customers can be assigned to any of the depot sites. This is equivalent to operating a depot at all possible depot sites. In each of the following steps of the construction phase, one depot site is excluded and assignments to this site are forbidden. The process of excluding depot sites is continued until the sum of the capacity of the remaining depots is lower than the total demand of the customers. Furthermore, in each step a feasible solution for the CFLP is generated and compared to the best solution obtained so far. If the new solution improves the best upper bound on the estimate of the total cost of the TSCFLP, it is saved as the new best one. In which order depots should be dropped is determined at the outset of the procedure and not, as in the well-known drop-algorithm, troughout the procedure. With this modification, the computing time is reduced substantially.

Determining the Drop Order. To determine in which order depot sites are excluded, the following steps are performed:

1. Calculate the average cost $\bar{c}_j$ of supplying one unit to any customer from depot j:

$$\bar{c}_j = \sum_{i=1}^{I} \frac{c_{ij}}{b_i} \qquad \forall j.$$

2. Estimate the average cost $\tilde{o}_j$ of shipping one unit from any plant to a customer through depot j by:

$$\tilde{o}_j = \bar{c}_j + \rho_j + \sum_{i=1}^{I} \frac{\tilde{f}_{ij} + \tilde{t}_{ij}}{b_i} \qquad \forall j.$$

   The values of $\tilde{f}_{ij}$ and $\tilde{t}_{ij}$ correspond to those of the initialisation phase.
3. Sort the depots according to nonincreasing values of $\tilde{o}_j$. The resulting ranking is the order in which the depots have to be excluded in the construction phase.

Constructing a Solution for the CFLP. As mentioned, the rule of excluding depot sites leads in every step to a set M of open depots. This set determines the possible assignments for the customers. For a given set M, the construction of a solution to the GAP and the CFLP is performed as follows:

1. Assign the customers at least cost to one of the depots of M regardless of the capacity constraints of the depots:

$$x_{ij} = \begin{cases} 1 \text{ if } \tilde{c}_{ij} = \min\limits_{k \in M} \tilde{c}_{ik} \wedge j \in M \\ 0 \text{ otherwise} \end{cases} \qquad \forall i, j.$$

2. Calculate the sum β_j of the demands assigned to depot j:

$$\beta_j = \sum_{i=1}^{I} b_i \cdot x_{ij} \qquad \forall j.$$

3. Determine the set $\hat{M}$ of depots whose capacity is violated:

$$j \in \hat{M} \iff \beta_j > s_j.$$

4. Compute the minimum additional cost $\hat{c}_i$ to reassign customer i, which is currently assigned to a depot $j^* \in \hat{M}$, to a depot $j \in M$, which has enough remaining capacity:

$$\hat{c}_i = \min_{\substack{j \in M \\ s_j - \beta_j - b_i \geq 0}} (\tilde{c}_{ij} - \tilde{c}_{ij^*}) \qquad \forall i.$$

   Sort these customers according to nondecreasing values of $\hat{c}_i$.

5. Perform the reassignments according to the ranking as long as the receiving depot has sufficient capacity. If β_j decreases to $\beta_j^{\text{new}} \leq s_j$ by such a reassignment, then j is taken out of $\hat{M}$:

$$\hat{M} = \hat{M} \setminus \{j\}.$$

6. Now the set of depots whose capacity is violated is examined:
   - If $\hat{M}$ is empty, a feasible solution is found and this part of the algorithm is finished.
   - If $\hat{M}$ is not empty and no further reassignment is possible, restart with step 2. The receiving depot may not have sufficient capacity anymore due to earlier reassignments.
   - In the case that $\hat{M}$ is not empty and no reassignment has been performed, start with the next part of the algorithm.

Similar reassignment procedures have been used by Barcelo et al. (1991), Barcelo and Casanova (1984), and Klincewicz and Luss (1986) in Lagrangean heuristics for the CFLP with single sourcing. They differ mostly in the way the reassignment costs are computed and in the selection scheme for the reassignments.

Reducing Infeasibility. The following procedure is equal to the one proposed by Klose (1995). The heuristic procedure tries to reduce an infeasibility measure u,

$$u = \sum_{j \in M} \max\{0, \beta_j - s_j\},$$

of the solution with respect to the depot capacity constraints by interchanging customers between depots.

A subset I_1 of the customers assigned to depot j_1 is assigned to depot j_2, and a subset I_2 of the customers of depot j_2 is assigned to depot j_1. The cardinality ($|I_1|, |I_2|$) of the sets I_1 and I_2 is set to (0,1), (1,0), and (1,1). This is done as follows:

1. Calculate the infeasibility u of the solution. Stop, if the infeasibility u is zero or no exchange of customers can reduce the infeasibility.
2. Take a combination of two open depots j_1 and j_2 and select subsets I_1 and I_2 of the customers assigned to each depot.
3. Compute the change $u_{I_1 I_2}$ in the infeasibility defined as:

$$\begin{aligned} u_{I_1 I_2} = {} & \max\{0, \beta_{j_1} - \sum_{i \in I_1} b_i + \sum_{i \in I_2} b_i - s_{j_1}\} + \\ & \max\{0, \beta_{j_2} + \sum_{i \in I_1} b_i - \sum_{i \in I_2} b_i - s_{j_2}\} - \\ & \max\{0, \beta_{j_1} - s_{j_1}\} + \max\{0, \beta_{j_2} - s_{j_2}\}. \end{aligned}$$

4. If the infeasibility is reduced by exchanging the subsets, $u_{I_1 I_2} \leq 0$, then the next step is performed. Otherwise, go to step 2.

5. Calculate the change $\hat{c}_{I_1 I_2}$ in cost caused by such an exchange:

$$\hat{c}_{I_1 I_2} = \sum_{i \in I_1} (-\tilde{c}_{ij_1} + \tilde{c}_{ij_2}) + \sum_{i \in I_2} (\tilde{c}_{ij_1} - \tilde{c}_{ij_2}) .$$

6. Determine the ratio

$$r_{I_1 I_2} = \frac{\hat{c}_{I_1 I_2}}{-u_{I_1 I_2}}$$

between the cost difference and the reduction of infeasibility.

7. Select the exchange, which has the minimum ratio $r_{I_1 I_2}$ for a given combination of depots. Perform the exchange of the subsets and go to step 1.

Estimating the Total Cost of the TSCFLP. The solution produced by the two preceding procedures consists in the set $M^\star$ of open depots, the throughput $\beta_j^\star$ at the depots and the allocation $x_{ij}^\star$ of the customers to the depots. This information is used to calculate the values of the variables $y_j^\star$:

$$y_j^\star = \begin{cases} 1 \text{ if } \beta_j > 0 \\ 0 \text{ otherwise} \end{cases} \quad \forall j.$$

Note that, especially at the beginning of the algorithm, it is possible that some of the depots of $M^\star$ are not used according to this calculation.

The estimation of the total cost $\tilde{v}$ is now computed by

$$\tilde{v} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\tilde{t}_{ij} + \rho_j \cdot b_i + c_{ij}\right) \cdot x_{ij}^\star + \sum_{j=1}^{J} f_j \cdot y_j^\star$$

which is equal to (9).

If $\tilde{v}$ improves the best upper bound found so far, the solution is saved as the new best one. After this comparison, the next depot is excluded and the next feasible solution constructed, etc., until the total capacity of the remaining depots is insufficient.

3.4 Improvement Phase

Once a feasible solution has been obtained, the estimates of the first stage transportation costs $\tilde{t}_{ij}$ and consequently the elements $\tilde{c}_{ij}$ of the single cost matrix can be improved. Performed after completion of the construction phase and then each time an improved solution of the CFLP has been found, the improvement of the estimate is conducted in two steps.

First, the transportation problem of the first stage (14) to (17) is solved and the present estimate

$$\tilde{v}(\text{TP}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \tilde{t}_{ij} \cdot x_{ij}^\star$$

of this cost is computed. Then, the correction term t_j^{cor} is adjusted by:

$$t_j^{\text{cor, new}} = t_j^{\text{cor}} \cdot \frac{\tilde{v}(\text{TP}) - \sum_{i=1}^{I} t_j^{\min} \cdot b_i}{v(\text{TP}) - \sum_{i=1}^{I} t_j^{\min} \cdot b_i} \quad \forall j,$$

and the new estimates $\tilde{t}_{ij}^{\text{new}}$ $\forall i, j$ are calculated as mentioned in Sec. 3.2.

The Local Search. The local search consists of five different methods to explore the neighborhood of the best solution of the CFLP found so far. These are called "One Out", "Two Out - One In", "One Out - One In", "One In" and "One Out - Two In" and are executed in this order.

1. One Out:
   In every step force one depot $j_1 \in M$ out of the solution. If the total capacity of the remaining depots is sufficient, i.e.
   $$\sum_{\substack{j \in M \\ j \neq j_1}} s_j \geq \sum_{i=1}^{I} b_i,$$
   go to step 6. Otherwise select the next depot. There are up to $|M|$ steps.
2. Two Out - One In:
   In each step take a combination of two depots currently in the solution, $j_1, j_2 \in M$, and exclude them. Take one closed depot $j_3 \in \bar{M}$ and include it. The total depot capacity of the new solution is compared with the total demand of the customers. If
   $$\sum_{j \in (M \setminus \{j_1, j_2\}) \cup \{j_3\}} s_j < \sum_{i=1}^{I} b_i \,,$$
   then the solution is discarded and the next step is performed. Otherwise go to step 6. This method has up to $0.5 \cdot |M| \cdot (|M| - 1) \cdot |\bar{M}|$ steps.
3. One Out - One In:
   A step of this method consists of excluding one depot $j_1 \in M$ from the solution and replacing it with one closed depot $j_2 \in \bar{M}$. Again, if the total depot capacity is larger or equal to the total demand of the customers
   $$\sum_{j \in (M \setminus \{j_1\}) \cup \{j_2\}} s_j \geq \sum_{i=1}^{I} b_i$$
   then step 6 is applied. Otherwise the next combination is selected. The number of steps in this part is up to $|M| \cdot |\bar{M}|$.

4. One In:

   This method is simply the reverse of the first method. In each step take a depot $j \in \bar{M}$ not currently in the solution and include it in the solution. Go to step 6. The method consists of up to $|\bar{M}|$ steps.
5. One Out - Two In:

   In contrast to the second method, exclude only one depot $j_1 \in M$ of the solution and include two unused depots j_2, $j_3 \in \bar{M}$. If the exchange is infeasible, i.e.

$$\sum_{j \in (M \setminus \{j_1\}) \cup \{j_2, j_3\}} s_j < \sum_{i=1}^{I} b_i,$$

   then take the next combination. Otherwise go to step 6. This method has a up to $0.5 \cdot |M| \cdot |\bar{M}| \cdot (|\bar{M}| - 1)$ steps.
6. A feasible solution is constructed by the same procedure as in the construction phase. If a solution of lower total cost with respect to the CFLP (9) is found, it is saved as the new best solution and the improvement phase is restarted. Otherwise, the next method out of the five is applied.

If all the methods above do not lead to an improved solution value, the local search is terminated, and the following heuristic is applied to improve the customer assignments.

Improving Customer Allocation. The aim of solving the GAP was up to this point only to find a feasible solution. Now, a local search heuristic for the GAP is applied.

The procedure tries to improve the customer assignments by interchanging customers between depots. A subset I_1 of the customers assigned to depot j_1 is assigned to depot j_2 and a subset I_2 of the customers of depot j_2 is assigned to depot j_1. The cardinality $(|I_1|, |I_2|)$ of the sets I_1 and I_2 is set to (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1) and (2,2). Such neighborhoods are used by Osman (1995) for the GAP. He also extended the presented local search descent method to simulated annealing and tabu search approaches.

For each combination of open depots j_1 and j_2 and each combination of subsets I_1 and I_2 the following procedure is executed:

1. Denote by

$$b_{i_1 j}^{\min} = \min_i \{b_i | x_{ij} = 1\} \quad \text{and} \quad b_{i_2 j}^{\min 2} = \min_{i \neq i_1} \{b_i | x_{ij} = 1\}$$

   the smallest and the second smallest demand currently assigned to depot j and by

$$b_{i_3 j}^{\max} = \max_i \{b_i | x_{ij} = 1\} \quad \text{and} \quad b_{i_4 j}^{\max 2} = \max_{i \neq i_3} \{b_i | x_{ij} = 1\}$$

   the largest and second largest demand assigned to depot j.

Perform some simple checks:
- If $c_{ij_1} \leq c_{ij2}\ \forall i \in I_1$ and $c_{ij_1} \geq c_{ij2}\ \forall i \in I_2$ then no improvement is possible by reassigning customers between depots j_1 and j_2.
- If $s_{j_1} < \beta_{j_1} + b^{\min}_{i_1 j_2}$ then no improvement is possible for $(|I_1|, |I_2|) = (0, 1)$.
- If $s_{j_1} < \beta_{j_1} + b^{\min}_{i_1 j_2} + b^{\min 2}_{i_2 j_2}$ then no improvement is possible for $(|I_1|, |I_2|) = (0, 2)$.
- If $s_{j_1} < \beta_{j_1} - b^{\max}_{i_3 j_1} + b^{\min}_{i_1 j_2}$ then no improvement is possible for $(|I_1|, |I_2|) = (1, 1)$.
- If $s_{j_1} < \beta_{j_1} - b^{\max}_{i_3 j_1} + b^{\min}_{i_1 j_2} + b^{\min 2}_{i_2 j_2}$ then no improvement is possible for $(|I_1|, |I_2|) = (1, 2)$.
- If $s_{j_1} < \beta_{j_1} - b^{\max}_{i_3 j_1} - b^{\max 2}_{i_4 j_1} + b^{\min}_{i_1 j_2} + b^{\min 2}_{i_2 j_2}$ then no improvement is possible for $(|I_1|, |I_2|) = (2, 2)$.

If an exchange is necessarily senseless then select the next combination. Otherwise continue with the next step.

2. Calculate the throughput at the depots, if the subsets are exchanged:

$$\beta^{\text{new}}_{j_1} = \beta_{j_1} - \sum_{i \in I_1} b_i + \sum_{i \in I_2} b_i$$

and

$$\beta^{\text{new}}_{j_2} = \beta_{j_2} + \sum_{i \in I_1} b_i - \sum_{i \in I_2} b_i.$$

3. If $\beta^{\text{new}}_{j_1} > s_{j_1}$ or $\beta^{\text{new}}_{j_2} > s_{j_2}$ then select the next combination of customers and depots. Otherwise continue with the next step.
4. Determine the cost difference $\hat{c}_{I_1 I_2}$ if the exchange can be performed:

$$\hat{c}_{I_1 I_2} = \sum_{i \in I_1} (-\tilde{c}_{ij_1} + \tilde{c}_{ij_2}) + \sum_{i \in I_2} (\tilde{c}_{ij_1} - \tilde{c}_{ij_2}).$$

5. If $\hat{c}_{I_1 I_2} < 0$ then perform the exchange of the subsets and store the new best solution. Restart the procedure with the first combination. Otherwise, continue with the next combination.

3.5 Combining Results

Up to this part of the algorithm, the total cost of a solution is only an estimate. The last task for the DIH is to calculate the exact cost of the solution found and to combine the results of the subproblems. This is done in two steps.

First, a transportation problem for the first stage analogous to (14) – (17) is solved and an optimal solution $z^\star_{pj}$ to the TP is obtained.

This gives, together with the best solution $x^\star_{ij}$, $y^\star_j$ to the CFLP, the following cost with respect to the TSCFLP (1):

$$v_{\text{DIH}} = \sum_{p=1}^{P} \sum_{j=1}^{J} t_{pj} \cdot z^\star_{pj} + \sum_{i=1}^{I} \sum_{j=1}^{J} (c_{ij} + \rho_j \cdot b_i) \cdot x^\star_{ij} + \sum_{j=1}^{J} f_j \cdot y^\star_j$$

4 Test problems

Contrary to the vehicle routing problem there are no "classical" test problems for the TSCFLP available. Thus, in order to obtain numerical results, a set of problems had to be generated. The parameters were chosen with the aim of generating realistic problems.

The geographical locations of the 100 largest cities of Switzerland were used as basis for all the generated problems. In each problem instance the locations of the plants, the depots and the customers were randomly chosen from these 100 points. Six problem classes, which differ in the number of plants, the number of depots and the number of customers, were generated as follows:

Class	a	b	c	d	e	f
# plants, P	10	10	10	20	20	20
# depots, J	25	25	50	25	25	50
# customers, I	50	100	100	50	100	100

In each problem class, a total of eight problem instances were generated with different depot capacities, throughput costs and fixed operating costs. Each of the three parameters was either set to a "small" (s) or a "large" (l) level.

Problem	1	2	3	4	5	6	7	8
Capacity s_j	s	s	s	s	l	l	l	l
Operating cost ρ_j	s	s	l	l	s	s	l	l
Fixed cost f_j	s	l	s	l	s	l	s	l

The actual problem instances were generated according to the following rules:

- The demands b_i of the customers are uniformly distributed in the interval $[1, 100]$.
- The cost c_{ij} of serving customer i from depot j was calculated by

$$c_{ij} = d_{ij} \cdot b_i \cdot 0.2 \qquad \forall i, j\,,$$

  where d_{ij} is the distance from customer i to the depot j in kilometers.
- The transportation cost t_{pj} per unit from plant p to depot j was calculated according to

$$t_{pj} = d'_{pj} \cdot 0.1 \qquad \forall p, j\,,$$

  where d'_{pj} is the distance from the plant to the depot in kilometers reflecting the economies of scale on the first stage transportation cost.
- The plant capacities σ_j are uniformly distributed in the interval

$$\left[0.5 \cdot \frac{B}{P}, 1.5 \cdot \frac{B}{P}\right]$$

  where B is the total demand ($B = \sum_{i=1}^{I} b_i$).

- Depot capacities s_j:
  - For problem instances with large depot capacities, these capacities are uniformly distributed in the interval

$$\left[5 \cdot \frac{B}{J}, 15 \cdot \frac{B}{J}\right]$$

    where B is the total demand and J the number of depots.
  - Analogously for problem instances with small depot capacities the interval

$$\left[2.5 \cdot \frac{B}{J}, 7.5 \cdot \frac{B}{J}\right]$$

    is used.
- Fixed operating cost f_j:
  - In the case of a problem instance with large fixed operating costs, the cost f_j was calculated by

$$f_j = s_j \cdot \bar{d} \cdot 0.04 \qquad \forall j$$

    where $\bar{d}$ is the average distance between depots and customers.
  - Analogously in the case of a problem instance with small fixed operating costs,

$$f_j = s_j \cdot \bar{d} \cdot 0.02 \qquad \forall j$$

    was used.
- Throughput cost ρ_j:
  - For problems with large throughput cost, these are uniformly distributed in $[0, \bar{d} \cdot 0.04]$.
  - Analogously for problems with small throughput cost the interval $[0, \bar{d} \cdot 0.02]$ is used.

The test problems are named according to the following convention: The first letter is always "t" to indicate the test problem. The second letter specifies the problem class "a" to "f". The next three letters specify the variation of the parameters, "s" for small and "l" for large values, where the first letter denotes the depot capacity, the second the throughput cost and the third the fixed operating cost. The problem name "talsl" specifies the problem instance in class "a" with large depot capacity, small throughput cost and large fixed operating cost.

5 Numerical Results

The DIH algorithm was implemented in SunSoft PASCAL 4.0 on a SUN ULTRA workstation (167 MHz). The lower bounds and the optimal solutions were obtained on the same workstation by the general purpose LP/MIP-Solver CPLEX 3.0, with a given limit of three hours of CPU-time.

In Tab. 1, the second column gives the total CPU-time of the Drop-Interchange-Heuristic in seconds. The significant differences in computing time are due to the fact, that the improvement phase restarts every time a better solution is found. Column three compares the objective value of the heuristic with the optimal solution or with the lower bound, resp. It indicates the percentage by which the Drop-Interchange-Heuristic exceeds the optimal solution or the lower bound, resp. In case the comparison was made with respect to the optimal solution, the corresponding percentage is shown in bold face numbers and the CPU-time of CPLEX is given in the last column.

As can be seen from Tab. 1, the solutions found by the DIH exceed the lower bounds by 2.37% in average over all problem classes and instances. With respect to the optimal solutions the performance is probably even better. An estimate for that performance can be derived by taking the average of the nine test problems which have been solved to optimality by CPLEX. Comparing only those solutions, the values found by DIH are only 1.82% higher on average.

With a maximum observed deviation of 8.04% from the optimum, a minimum deviation less than 0.01% and low system requirements, the heuristic has to be considered as an effective one.

Unfortunately, it is not possible to compare the heuristic to the optimal solution for larger problem instances anymore. The only way to evaluate the performance of the heuristic is to compare it with other heuristics. For this propose the Drop-Interchange-Heuristic is compared to the Lagrangean heuristic presented by Klose (1997). As the Lagrangean heuristic was tested with a number of different parameters and configurations yielding a slightly different behavior, the comparison is made with the best solution, the average solution and the worst solution produced by the Lagrangean heuristic, given in Tab. 2.

Comparing the solutions of the Drop-Interchange-Heuristic with the best solutions found by the Lagrangean heuristic, it can be seen that the values of the DIH were 1.53% higher on average. The CPU-time needed by the DIH was a fraction, a factor of 1/83.35 on average, of the time needed by the Lagrangean heuristic. Therefore, the Drop-Interchange-Heuristic has to be considered as not dominated in performance. In three of the 48 test problems even a slightly better solution was found by the Drop-Interchange-Heuristic reassuring the previous statement. The DIH however lacks the ability to generate a lower bound for a problem instance at hand, but this is the case for most of the heuristic procedures.

Comparing the Drop-Interchange-Heuristic to the average performance of the Lagrangean heuristic shows, that there is still a slight advantage (0.76%) for the Lagrangean heuristic in terms of the average solution value. A superiority of 1.56% in the average solution value is the result when comparing the Drop-Interchange-Heuristic to the worst-case behavior of the Lagrangean heuristic.

Table 1. Results of Class "a" to "f" Problems

Problem	sec DIH	% to Bound	sec CPLEX	Problem	sec DIH	% to Bound	sec CPLEX
tasss	1.15	**1.20**	543.70	tbsss	14.37	0.41	
tassl	1.28	2.69		tbssl	30.53	4.13	
tasls	1.33	4.54		tbsls	31.50	1.23	
tasll	2.30	1.49		tbsll	14.83	0.57	
talss	1.55	**0.17**	104.35	tblss	4.75	5.94	
talsl	1.28	0.75		tblsl	4.65	1.47	
talls	0.83	7.03		tblls	6.13	**0.15**	66.26
talll	0.98	**3.30**	4224.76	tblll	43.83	**0.13**	980.88
Average	1.34	2.65		Average	18.82	1.75	
Min	0.83	0.17		Min	4.65	0.13	
Max	2.30	7.03		Max	43.83	5.94	
tcsss	59.67	2.76		tdsss	2.10	0.73	
tcssl	55.85	3.92		tdssl	1.38	1.97	
tcsls	61.73	2.40		tdsls	1.70	0.59	
tcsll	93.70	1.40		tdsll	1.47	0.34	
tclss	17.68	1.50		tdlss	2.88	0.28	
tclsl	29.13	2.87		tdlsl	0.78	**8.04**	73.35
tclls	19.28	3.86		tdlls	1.13	0.99	
tclll	30.77	3.04		tdlll	1.18	**0.01**	141.35
Average	45.98	2.72		Average	1.58	1.62	
Min	17.68	1.40		Min	0.78	0.01	
Max	93.70	3.92		Max	2.88	8.04	
tesss	6.80	3.11		tfsss	101.7	2.16	
tessl	3.60	7.01		tfssl	69.53	3.47	
tesls	6.40	0.81		tfsls	80.63	4.50	
tesll	3.72	0.22		tfsll	67.55	2.76	
telss	43.63	**0.01**	464.89	tflss	22.18	6.77	
telsl	7.70	1.06		tflsl	28.70	0.72	
tells	7.60	**3.41**	2693.92	tflls	18.43	3.14	
telll	7.75	3.15		tflll	27.30	1.60	
Average	10.90	2.35		Average	52.01	3.14	
Min	3.60	0.01		Min	18.43	0.72	
Max	43.63	7.01		Max	101.78	6.77	

6 Conclusions and Outlook

The algorithm presented in this paper can be seen as a typical heuristic. The basic idea of the algorithm is rather simple, and it is capable of solving medium-sized problems with more than sufficient exactness in reasonable time. As with most heuristics, it is unknown by which amount the computed solution value exceeds the optimal one. The overall performance of this heuristic has to be considered as very satisfactory.

Table 2. Comparison of DIH and Lagrangean Heuristic by Klose

	class	best LgH solution		av LgH solution		worst LgH solution	
		100% Value	Time-factor	100% Value	Time-factor	1.00 Value	Time-factor
av	a	1.80	151.74	1.12	110.61	-1.99	181.87
	b	1.13	25.02	0.79	28.52	-0.91	53.83
	c	1.48	23.90	0.36	16.56	-2.69	32.35
	d	1.35	171.98	0.56	107.58	-1.52	207.21
	e	1.39	105.19	0.99	61.00	-0.80	120.17
	f	2.06	22.26	0.77	13.44	-1.48	29.73
min	a	0.01	35.60	-1.59	33.38	-10.10	52.30
	b	0.00	4.48	-0.34	4.53	-2.40	8.24
	c	0.00	2.69	-1.50	2.87	-8.27	4.32
	d	-0.02	24.69	-0.57	23.72	-3.32	33.05
	e	0.00	6.44	-1.33	4.03	-8.61	6.44
	f	-0.24	2.62	-1.97	1.91	-5.15	3.04
max	a	5.16	488.05	4.38	256.37	0.25	488.05
	b	4.42	46.66	3.39	78.73	0.01	178.17
	c	2.70	65.51	1.38	40.51	-0.07	88.10
	d	8.04	632.53	5.79	313.10	1.05	632.53
	e	6.86	524.10	6.58	246.55	5.56	524.10
	f	5.89	71.54	4.21	39.58	2.39	98.03
av	all	1.53	83.35	0.76	56.28	-1.56	104.19
min	all	-0.24	2.62	-1.97	1.91	-10.10	3.04
max	all	8.04	632.53	6.58	313.10	5.56	632.53

The heuristic was only compared to the Lagrangean heuristic mentioned. Therefore, it would be interesting to compare it to other solution methods for the TSCFLP. A further research area is the performance of the heuristic on problem instances of realistic size or on real problems of enterprises.

The research can be continued by evaluating other methods of exploring the neighborhood of a specific solution. It would be interesting to use local search techniques as for example "simulated annealing" or "tabu search" in the improvement phase in order to investigate if the performance of the heuristic can be increased.

Acknowledgement: This research was founded by the "Swiss Federal Commission for Technology and Innovation (KTI)".

References

Aardal, K. / Pochet, Y. / Wolsey, L. A. (1995): Capacitated facility location: Valid inequalities and facets. *Mathematics of Operations Research*, 20:552–582.

Barcelo, J. / Casanovas, J. (1984): A heuristic Lagrangean algorithm for the capacitated plant location problem. *European Journal of Operational Research*, 15:212–226.

Barcelo, J. / Fernandez, E. / Jörnsten, K. O. (1991): Computational results from a new Lagrangean relaxation algorithm for the capacitated plant location problem. *European Journal of Operational Research*, 52:38–45.

Beasley, J. E. (1988): An algorithm for solving large capacitated warehouse location problems. *European Journal of Operational Research*, 33:314–325.

Beasley, J. E. (1993): Lagrangean heuristics for location problems. *European Journal of Operational Research*, 65:383–399.

Cornuejols, G. / Fisher, M. L. / Nemhauser, G. L. (1977): Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23:789–810.

Cornuejols, G. / Sridharan, R. / Thizy, J. M. (1991): A comparison of heuristics and relaxations for the capacitated plant location problem. *European Journal of Operational Research*, 50:280–297.

Domschke, W. / Drexl, A. (1985): ADD-heuristics' starting procedures for capacitated plant location models. *European Journal of Operational Research*, 21:47–53.

Feldman, E. / Lehrer, F. A. / Ray, T. L. (1966): Warehouse location under continuous economies of scale. *Management Science*, 12:670–684.

Geoffrion, A. M. / Graves, G. W. (1974): Multicommodity distribution system design by Benders decomposition. *Management Science*, 20:822–844.

Hindi, K. S. / Basta, T. (1994): Computationally efficient solution of a multi-product, two-stage distribution-location problem. *The Journal of the Operational Research Society*, 45:1316–1323.

Jacobsen, S. K. (1983): Heuristics for the capacitated plant location model. *European Journal of Operational Research*, 12:253–261.

Kaufman, L. / Eede, M. V. / Hansen, P. (1977): A plant and warehouse location problem. *Operational Research Quarterly*, 28:547–554.

Klincewicz, J. G. / Luss, H. (1986): A Lagrangian relaxation heuristic for capacitated facility location with single source constraints. *The Journal of the Operational Research Society*, 37:495–500.

Klose, A. (1995): A lagrangean heuristic to solve the two-stage capacitated facility location problem. Working paper, Institut für Unternehmensforschung (Operations Research), Hochschule St. Gallen, Bodanstrasse 6, CH–9000 St. Gallen, October 1995.

Klose, A. (1997): Obtaining sharp lower and upper bounds for two-stage capacitated facility location problems. This volume.

Kuehn, A. A. / Hamburger, M. J. (1963): A heuristic program for locating warehouses. *Management Science*, 9:643–666.

Osman, I. H. (1995): Heuristics for the generalized assignment problem: Simulated annealing and tabu search approaches. *OR-Spektrum*, 17(4):211–225, October 1995.

Sridharan, R. (1991): A lagrangian heuristic for the capacitated plant location problem with side constraints. *The Journal of the Operational Research Society*, 42:579–585.

Sridharan, R. (1993): A lagrangian heuristic for the capacitated plant location problem with single source constraints. *European Journal of Operational Research*, 66:305–312.

Van Roy, T. J. (1986): A cross decomposition algorithm for capacitated facility location. *Operations Research*, 34:145–163.
Van Roy, T. J. (1989): Multi-level production and distribution planning with transportation fleet optimization. *Management Science*, 35:1443–1453.

Modelling the Distribution Processes of Tour Operator Catalogues

Joachim R. Daduna

Fachhochschule für Wirtschaft Berlin, Badensche Straße 50 - 51, D - 10825 Berlin

Abstract: In the German tourism branch, the distribution processes of tour operator catalogues at present are not efficiently organized. This situation necessitates fundamental improvements from an economic as well as an ecological point of view. Starting from the actual operational processes an, efficient distribution structure is developed which is based on a concept of logistic consolidators. For this solution, two different models are presented in detail and analyzed with respect to their possible applications. Taking the actual surroundings as well as future trends in information and communication technologies into consideration, the concluding statements give a rough insight into the current state, including further steps which are necessary to bring this concept to a successful completion.

Keywords: Distribution, information management, tour operator, travel agency, tour operator catalogue

1 Introduction

In the German tourism branch, tour operator catalogues are at the moment the most important medium to present detailed information to (potential) clients. This situation is illustrated in particular by the fact, that the tour operators spend about 50% of their advertising budgets for this specific information media. Taking into consideration that these budgets amount to 4% of returns, it refers to an extent of DM 420 Mio, based on total sales in the tourism branch of about DM 21.000 Mio for the season 1994/95. Starting from the fact that the average cost to produce a catalogue comes to DM 2,80 (see Lettl-Schröder (1995b)), it can be assumed that about 150 Mio catalogues have been printed for this season. Based on an average weight of 250 g per catalogue, the total weight amounts to 37.000 t.

The structure in this market segment shows (in 1995) a number of about 700 to 800 tour operators which can be clustered into five greater, 20 to 30 middle-sized and 700 to 750 small-sized companies (see Ammann / Illing / Sinning (1995), p 46) These companies offer their products to approximately 17.500 travel

agencies (see DRV - Deutscher Reisebüroverband (1995)) which show at present an extreme spectrum in their structure and size. To give information to these distributors (and also to the potential clients), different catalogue titles are produced by the tour operators (or rather by specific services providers) and made available free of charge. Each of the travel agencies has on an average the products of 90 tour operators for sale.

The organisation of the supply chain to serve the travel agencies with tour operator catalogues shows distinct deficits for the first deliveries at the beginning of a (seasonal) selling period as well as for the subsequent deliveries for replenishment of stocks. These results are stating, among others, a market study carried out on behalf of a German tourism journal[1]. The focal points, criticized in this study refer to the situation concerning economic and also ecological aspects (see Daduna / Koch / Maciejewski (1997)) with tour operators on one hand and travel agencies on the other hand.

For the travel agencies the main problems depend on the calculation of the demand of catalogue titles and on the distribution concepts which are not efficient and not unique among the tour operators, too. In most cases the calculations are based on unsuitable data, e.g. on the sales of the previous selling period(s), which normally leads to an insufficient estimation of demand. These problems especially appear because of short-termed alterations in clients' behaviour. Beside this, a strong seasonality in demand must be taken into consideration. At the beginning of a selling period, a large share of the produced catalogues must be delivered, while the subsequent deliveries during the period are distributed extremely uneven.

Furthermore, if the occasion arises, the necessary stock replenishments are not efficiently organized. By reason of inefficient information and distribution structures, ordering times up to two or three weeks may happen. Thus, at the beginning of a selling period the travel agencies are supplied with a too large number of catalogues, so that at present first deliveries run up to a share of 50 to 70% of the produced volume. This situation results in expansive inventories at the travel agencies and necessitates an according storage capacity.

In addition, it must be taken into consideration that nearly each tour operator has an own order processing (see Leitermann (1996), pp 82), so that it is impossible for the travel agencies to organize these processes efficiently from their point of view. For the tour operators the inefficiency concerning subsequent deliveries also becomes apparent as a problem in sales promotion, because without the catalogues the travel agencies get into difficulties when giving their clients an extensive information about the offered products.

Besides, there are other problems depending on the insufficient estimation of demand. The calculation of the size of printing runs for the different catalogue titles, which is based on those estimations, normally leads to volumes of production which do not fit the real demand. The production surplus which is

---

[1] cf. details of this study in the FVW Fremdenverkehrswirtschaft International (Anonymous (1995)).

estimated to up to 20% of the produced number of copies (see Lettl-Schröder (1995a)), causes unnecessary cost of about DM 84 Mio.

These expenses depend only on the production cost of not distributed catalogues, but in this case further cost to dispose this surplus has to be taken into consideration, too. This obligation results from legal conditions (in Germany), as with the *Kreislaufwirtschafts- und Abfallgesetz* (KrW/AbfG) coming into force in fall 1996, the tour operators are responsible to dispose their remainders satisfying the specific requirements of this law. But, to calculate the complete disposal cost, it must be taken into consideration that the share of 20% of not used catalogues represents only a lower bound, disregarding the remainders of the travel agencies. Therefore, the total cost of inefficient planning and distribution processes mount up to far more than the above mentioned sum.

The rough overview illustrates the existing problems in the supply of the travel agencies with tour operator catalogues, which at present are not solved with respect to an operational efficiency. Therefore, in the following a solution concept is presented, showing a distribution procedure which is more efficient not only from an economic point of view, but also concerning ecological aspects.

Starting in the first step with a short description of the actual distribution procedures, it leads in the second step to a definition of essential requirements concerning an appropriate logistic structure. Based on these conditions, two different models are developed which are suitable to solve the given distribution problem. Mathematical formulations of both models are shown and their practical application are discussed, taking the actual environment as well as future trends into consideration, especially in information technologies. The concluding statements give information about the current state, including further steps which are necessary to bring such a concept to success.

2 Actual distribution procedures

The actual distribution procedures for tour operator catalogues differ largely between the tour operators concerning the determination of the estimated demand for the various titles and especially for carrying out the distribution. A study of the business processes for this branch has shown that overall a company-specific procedure is preferred (see, e.g., Leitermann (1996), pp 82). The employment of external logistic services is in most cases common. These are, for example, haulage contractors which take charge of specific parts of the distribution processes, first of all concerning the (physical) operations which include all shipments from the printing plants to the travel agencies. A basic structure of these processes is shown in Fig. 1.

From these facts, it becomes clear that the basic idea of a disembodiment of logistic tasks, in the meaning of an *outsourcing strategy*[2], is applied to a certain extend. In spite of this step, which is realized by many tour operators to attain

[2] For a definiton of the *outsourcing conception* in the field of logistics cf. e.g. Broggi (1994), Wißkirchen (1995), Fischer (1996) or Uhlig (1996a).

more efficiency in logistic operations, the main problems within this distribution structure result in the missing of a company overlapping coordination of the processes and also, as mentioned above, in the insufficient information about the existing demand structures. In addition, Fig. 2 and 3 give a rough overview on the basic structure of the information processes at present concerning the estimation of demands and also the operational runs within the distribution chain.

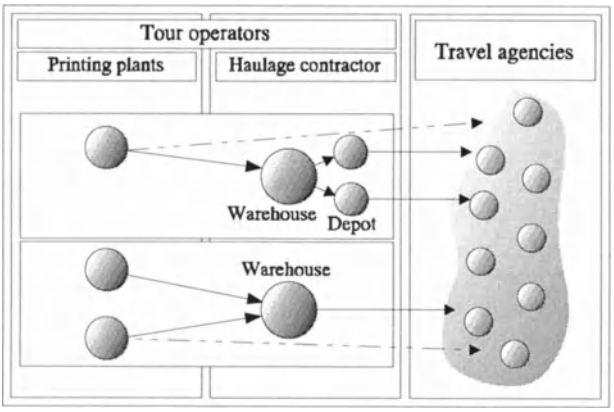


Fig. 1: Basic structure of actual distribution procedures

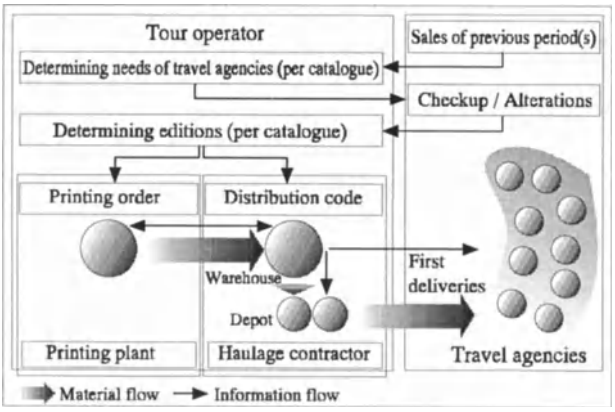


Fig. 2: Basic structure of first deliveries

The above mentioned deficiencies become obvious as shown in these two figures, especially concerning the structure of the subsequent deliveries. Thus the processing of repeat orders for stock replenishments take an extended time, which depends on one hand on collecting orders by the tour operators during a week and on the other hand on the partial separation in the organisation of information flow and material flow. Within this structure the haulage contractors are only re-

sponsible for carrying out the physical distribution, while essential parts of planning and control remain in the sphere of influence of the respective tour operators.

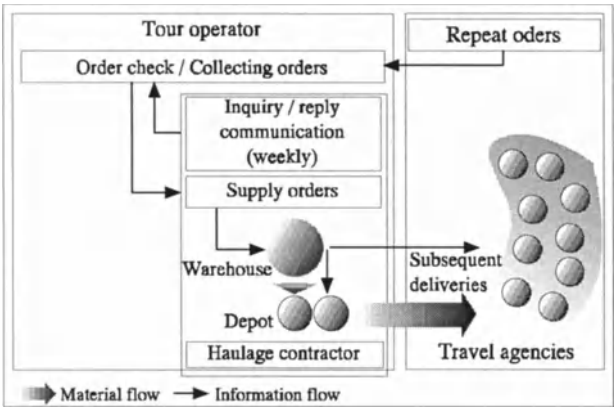


Fig. 3: Basic structure of subsequent deliveries

Therefore, to attain more efficient operations and a demand-oriented supply, it is of prime necessity to synchronize information and material flows. In connection with this step, the responsibility of the distribution processes must be concentrated in the same place, not with respect to only one tour operator but rather to a great number of them or, if possible, by including all.

3 Requirements for an efficient distribution structure

Defining the requirements of an efficient distribution structure, the strategic and operational aims of the tour operators as well as of the travel agencies must be taken into consideration. The tour operators are mainly interested in avoiding the production of those shares of catalogues which are not handed out to (potential) clients. But, in connection with this aim, they need an allocation of the catalogues according to the real demand at the travel agencies, too, which differ subject to seasonal influences on one hand and to the socio-economic structure of the population within the region served by a travel agency on the other hand. The travel agencies first of all expect simplified order processes in connection with significant shortenings in delivery times and an inventory reduction (see *Anonymous* (1995)).

Beside these requests of tour operators and travel agencies, some global aspects relating to economics and ecology have to be included in the discussion. In this context especially a reduction of the expenditures on transportation must be mentioned as well as the disposal of not used catalogues. Looking at these aspects, an interesting point of view comes to the fore, as this situation shows, that there

does not exist a contradiction between the individual (economic) aims of tour operators and travel agencies and public aims but rather a congruence.

These different aspects and the difficulties shown for the actual situation lead to a number of criteria for the design of an efficient distribution structure:

- Development of a centralized distribution system including (most of) the tour operators
- Bundling transport operations
- Designing a process-oriented and company-independent information and communication structure
- Reducing inventory by a demand-oriented supply of needed catalogue titles
- Possibility to include the disposal of not used catalogues
- Adaptability to alterations in the specific surroundings

A centralized distribution system necessitates a concept that is based on *logistic consolidators*[3] which provide an overall logistic management, integrating physical transportation and warehousing in connection with information systems. This approach proceeds from the basic reflection to concentrate the business activities of the companies on their core competence (see, e.g., Prahalad / Hamel (1990) and Suter (1995)) All other activities which can not be performed effciently must be acquired, except those which are of strategic importance concerning the respective business purposes. This includes, e.g., the supply of primary products as well as different manners of service. For tour operators, the distribution of catalogues can definitely not looked upon as a core activity so that it seems to be useful to acquire complete logistic services to reduce operational cost (see Bliesener (1994)).

A practicable structure for the distribution of tour operator catalogues based on logistic consolidators is shown in Fig. 4. Within the scope of this modified distribution process, the largest amount of produced catalogues is shipped in the first step from the printing plants to the distribution centers which are operated by logistic consolidators. Only some travel agencies with a greater demand for specified catalogue titles are served directly in the first deliveries, while this part of the distribution process is of course organized by the logistic consolidators, too. For all other cases in the first and especially in the subsequent deliveries, the supply of travel agencies is carried out solely with incorporation of a distribution center. In comparison to actual distribution procedures, it is possible to bundle the supply of the travel agencies in the first deliveries and also in the subsequent deliveries.

---

[3] For a definiton of a *logistic consolidator* cf. e.g. Wißkirchen (1995) and Fischer (1996).

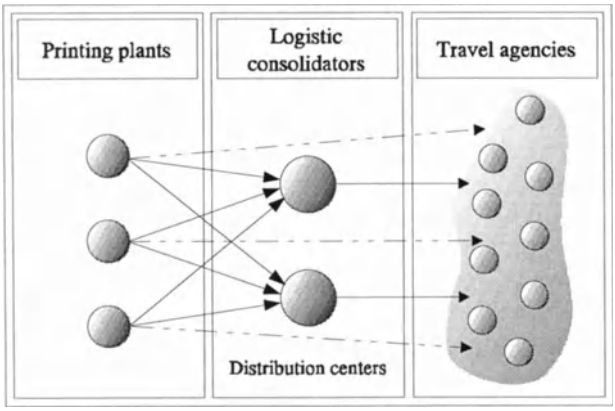


Fig. 4: Basic model for a modified distribution structure

As mentioned above, a logistic consolidator will be responsible not only for the (physical) operations of the distribution processes but for an interorganizational information management, too. To assign these competences to such an external service company is a preemptory necessity, because a complete coordination must be available to attain an efficient carrying out of the shipments and also for monitoring and control of material and information flow. The basis of the aimed up information management consists of a virtual structure (see, e.g., Herget (1995) and Reekers / Smithson (1996)) connecting the different participants within the distribution chain. Fig. 5 gives an overview of the information flow concerning the distribution process and some of the (additional) tasks of information management which are assumed by a logistic consolidator.

From the travel agencies' point of view, this information concept leads to a significant simplification for ordering additional catalogues during the course of a selling period and in the delivery processes, too, as they have only one counterpart for all transactions. Based on this concept, they receive only one consolidated sending which normally contains catalogue titles from different tour operators. Moreover, by supplying in accordance to due date and also to order volume the inventory can be reduced to an evident extent.

As mentioned above, the tour operators became responsible by law for the disposal of remainders. Therefore, a solution must be developed to integrate these additional tasks of *reverse logistics* into a joined distribution system. In this context the disposal has to include the remainders in the distribution centers as well as those at the travel agencies. To reduce the logistic efforts, the collection of the unused catalogues on the second stage should be connected with the deliveries of the catalogues for the new selling period. Since the selling periods show an overlapping phase, the collecting process can be combined only with the subsequent deliveries.

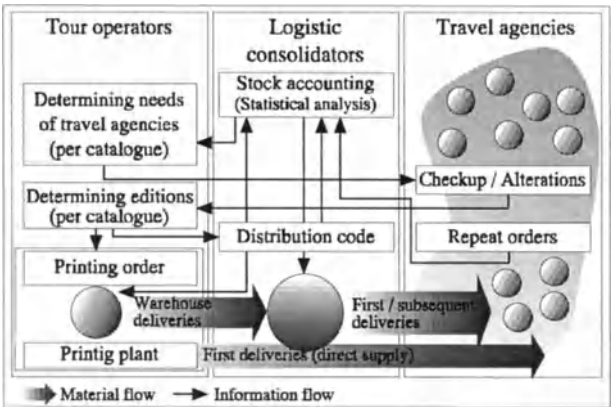


Fig. 5: Basic structure of information flows

Beside these aspects which concern the actual surroundings, some future developments have to be taken into consideration. At present the tour operator catalogues have a decisive influence within advertising concepts and sales promotion for the tour operators. With the extreme increase in the spreading of high-performance PC's, especially for private applications, more and more the employment of CD-ROM techniques is discussed in this field. One of the advantages of such techniques exists in the opportunity to apply multimedia concepts in advertising. In comparison to the print media, these techniques allow a more attractive presentation of offered products by combining, e.g.; printed information with voice or/and video sequences. Another strategy in advertising and sales promotion depends on the availability of network providers like the Internet. Making use of this technology, an additional distribution channel becomes available, and it will be possible to serve directly the end-users without the incorporation of travel agencies, and especially without physical distribution procedures.

It is expected that these various information techniques become more and more applied in the tourism branch (see, e.g., Ernst (1994), Rohte (1994) or Schertler (1994)), so that the share of tour operator catalogues in advertising may go down extremely during the next years. Proceeding from these trends, the amount of produced catalogues will become reduced within the next decade to about 40% of todays amount. Therefore, it must be possible to adapt the capability of the distribution system to the modified requirements. To attain a continuous utilization rate of the distribution centers for the logistic consolidators, it is necessary to make efforts to take control of additional tasks, which arise from the alterations of the employed techniques, e.g. the copying of CD-ROMs and their distribution together with the catalogues. Beside these required adaptions, of course the field of activities should be extended to neighbouring tasks for diversifying, so that the dependence of only one branch becomes compensated.

4 Distribution models

To model the problem of distributing tour operator catalogues, different approaches are possible. As there do not exist any experiences with such a distribution system or a similar application[4], a basic structure has to be defined. The main criteria coming up for discussion in this respect, are the number of distribution centers (respectively the number of logistic consolidators) which should be realized and the responsibility for the accomplishment of the (physical) distribution on the last step in the transportation chain.

The first model (*Model 1*) which has been designed as a specialized distribution system to guarantee short response times for subsequent deliveries. It is based on a decentralized structure with about 100 distribution centers and a separate car fleet at each center. Discussing the realization of Model 1, a second model (*Model 2*) with a complete different approach comes to the fore. This model shows a more centralized structure with only a small number of distribution centers, and contrary to Model 1, the distribution on the last step is performed by a parcel service. These two models for solving the underlying strategic planning problems are presented in detail in the following sections, and their advantages and disadvantages are discussed in detail.

4.1 Model 1

Model 1, which is shown in its basic structure in Fig. 6, consists of *a 2-stage multicommodity distribution process* in which the different catalogues represent the commodities. The first stage shows the transportation of the catalogues from the printing plants s_k ($k = 1, .., q$) to the possible locations w_i ($i = 1,.., m$) of the distribution centers. The available amount in s_k is a_k^h ($k = 1, .., q$; $h = 1, .., \eta$) where η is the number of different titles of catalogues included in the given planning process. Here it is possible, that on one hand a catalogue title h is produced in only one of the printing plants as well as in two or more, and on the other hand each printing plant can produce more than one catalogue title. On the second stage the supply b_j^h ($j = 1, .., n$; $h = 1, .., \eta$) of the travel agencies v_j ($j = 1, .., n$) has to be performed. This process contains the first deliveries of new catalogues as well as the subsequent deliveries within a selling period.

The distances (average transportation cost per unit) between the printing plants s_k and the possible locations of the distribution centers w_i are given by the matrix $D^1 = (d_{ki}^1)$ ($C^1 = (c_{ki}^1)$) and the distances between the w_i and the travel agencies v_j by the matrix $D^2 = (d_{ij}^2)$ ($C^2 = (c_{ij}^2)$). The shipments between s_k and w_i for catalogue type h are x_{ki}^{1h} and correspondingly x_{ij}^{2h} for the shipments between w_i and v_j.

---

[4] The distribution structures to supply pharmacies or bookshops are similar to some respects, but they can not be applied, because distributors in these cases are traders, not logistic consolidators. Therefore, the material flows on the two stages must be looked at as independent processes.

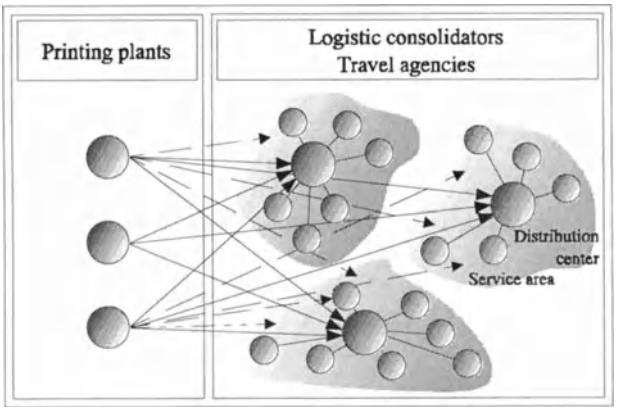


Fig. 6: Basic structure of Model 1

To solve this planning problem, the direct supply of travel agencies within the first deliveries can be disregarded, because the location of all s_k and all directly served v_j are known as well as their demand at that time. This subproblem can be formulated as a (*multicommodity*) *transportation problem*, which must be solved independently from the other steps of the distribution problem.

Therefore, from this point of view the basic planning process consists of two different problems, which must be solved simultanously (see Daduna (1985), pp 141). The first stage represents an (*uncapacitated*) *multicommodity warehouse location problem* (WLP) (see, Klincewicz / Luss / Rosenberg (1986), Klincewicz / Luss (1987), Crainic / Dejax / Delorme (1989), Crainic / Delorme (1993), or Gao / Robinson (1994)) with the printing plants as the one node set $S = \{s_k\}$ and the possible locations of distribution centers as the other set $W = \{w_i\}$. All shipments on this stage are carried out only at a particular time before a selling period starts. In contrast to this fixed dates a quasi-continuous problem exists on the second stage, because the quantities b_j^h are determined (for the whole selling period) but the dates of delivery during the period are unknown (except the first deliveries). The underlying task is in this case to define service areas which become assigned to the (established) distribution centers. Such a problem can be formulated as a *set covering problem* (SCP) (see Domschke / Drexl (1996), pp 148). The planning of the *daily tours* to serve the travel agencies, which must be solved as a *vehicle scheduling problem* (see, e.g. Golden / Assad (1988), Laporte (1992), and Domschke (1997), pp 204), is not included in this discussion, because these problems are part of the operational and not the strategic planning.

To solve such a complex structured mathematical model, which can be characterized as a *multicommodity location-allocation problem* for *many-to-many distribution*[5], there does not exist an appropriate algorithm. Therefore, an efficient

[5] For a (*single commodity*) *location-allocation problem* for a *many-to-many distribution* see e.g. Campbell (1992).

procedure must be found to determine fitting solutions for this model. Three different approaches in this case are possible to solve a (*uncapacitated*) *2-stage multicommodity WLP*[6] with a given radius $\tilde{r}$ and binary assignments of the travel agencies to the distribution centers, to solve in an iterative procedure a sequence of *partial single arc* (*multi-commodity*) *bottleneck transshipment problems* (see Daduna (1985), pp 138) with the bottleneck objective function on the second stage, and to solve, also in an iterative procedure, a *2-stage multicommodity WLP* determining *pareto-optimal* solutions.

For the first approach the following mathematical formulation is given[7]:

$$\text{Minimize } Z(x,y) = \sum_{k=1}^{q}\sum_{h=1}^{\eta}\sum_{i=1}^{m} c_{ki}^{1} x_{ki}^{1h} + \sum_{i=1}^{m}\sum_{h=1}^{\eta}\sum_{j=1}^{n} c_{ij}^{2} x_{ij}^{2h} + \sum_{i=1}^{m} f_i y_i \qquad (1)$$

subject to:

$$\sum_{i=1}^{m} x_{ki}^{1h} = d_k^h \qquad \forall\ k = 1, .., q;\ h = 1, .., \eta \qquad (2)$$

$$\sum_{k=1}^{q} x_{ki}^{1h} = \sum_{j=1}^{n} x_{ij}^{2h} \qquad \forall\ i = 1, .., m;\ h = 1, .., \eta \qquad (3)$$

$$\sum_{i=1}^{m} x_{ij}^{2h} = b_j^h \qquad \forall\ j = 1, .., n;\ h = 1, .., \eta \qquad (4)$$

$$x_{ij}^{2h} y_i \le b_j^h \qquad \forall\ j = 1, .., n;\ h = 1, .., \eta;\ i = 1, ..., m \qquad (5)$$

$$z_{ij} \le y_i \qquad \forall\ i = 1, .., m; j = 1, ..., n \qquad (6)$$

$$\sum_{i=1}^{m} z_{ij} = 1 \qquad \forall\ j = 1, ..., n \qquad (7)$$

$$z_{ij} d_{ij}^2 \le \tilde{r} \qquad \forall\ i = 1, .., m; j = 1, ..., n \qquad (8)$$

$$z_{ij} \in \{0,1\} \qquad \forall\ i = 1, .., m; j = 1, ..., n \qquad (9)$$

$$y_i \in \{0,1\} \qquad \forall\ i = 1, .., m \qquad (10)$$

$$x_{ki}^{1h},\ x_{ij}^{2h} \ge 0 \qquad \forall\ h = 1, .., \eta;\ i = 1, ..., m;\ j = 1, .., n;\ k = 1, ..., q \qquad (11)$$

---

[6] For similar structured WLP-problem formulations cf. e.g. Domschke / Drexl (1996), pp 57 (*capacitated 2-stage WLP*) and Puerto / Hinojosa (1997) (*general multicommodity, multilevel, multietapic capacitated plants location problem*).

[7] The formulation of Geoffrion / Graves (1974) can not be used in this case, because the *flow* of the different catalogues between the printing plants and the travel agencies must be taken into consideration. Therefore, aggregated cost coefficients are unsuited to model such structures.

with:

$$z_{ij} = \begin{cases} 1 & \text{if travel agency } j \text{ is served by distribution center } i \\ 0 & \text{otherwise} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{if distribution center } i \text{ becomes realized} \\ 0 & \text{otherwise} \end{cases}$$

The main difficulties within this solution procedure are to calculate the value for $\tilde{r}$, because the given radius determines the number of locations to be realized. To come to an appropriate decision, the problem can be solved with different values for $\tilde{r}$ to attain alternative solutions having an enlarged freedom of choice.

The second approach is based, for example, on an add- or drop-strategy (see Domschke / Drexl (1996), pp 60) with a *partial single arc* (*multicommodity*) *bottleneck transshipment problem* as the underlying relaxation. These optimization problems show the following mathematical formulation:

$$\text{Minimize } BT(x) = \left\{ Max \left\{ d_{ij}^2 \mid x_{ij}^{2h} > 0 \right\} \right. \tag{12}$$

subject to:

(2) - (5), (7), and (9) - (11)

To include the sum minimization objective function in this solution procedure, especially concerning the first stage of the distribution process, a (*multicommodity*) *transshipment problem* (see, e.g., Kennington / Helgason (1980) or Domschke (1995), pp 154) with $BT(x)$ as an additional constraint on the second stage must be solved in a closing step for each iteration.

The third approach is based on a *2-stage multicommodity WLP*, too. In opposition to the above presented procedure the radius becomes not a fixed value, but alters step by step within the iterations of the solution procedure. The following mathematical formulation describes the problem to be solved at each iteration t:

$$\text{Minimize } Z(x,y,t) = \sum_{k=1}^{q} \sum_{h=1}^{\eta} \sum_{i=1}^{m} c_{ki}^1 x_{ki}^{1h}(t) + \sum_{i=1}^{m} \sum_{h=1}^{\eta} \sum_{j=1}^{n} c_{ij}^2 x_{ij}^{2h}(t) + \sum_{i=1}^{m} f_i y_i(t) \tag{13}$$

subject to:

(2) - (5), (9) - (11), and

$$r_t = r_{t-1} - 1 = \left\{ Max \left\{ d_{ij}^2 \mid x_{ij}^{2h}(t-1) > 0 \right\} \right\} - 1 \tag{14}$$

In this formulation r_t represents a *generated variable*, which depends of the length of the arcs in the solution for iteration t-1. The procedure starts with $r_t = \infty$ and ends if no feasible solution for the last generated r_t can be determined.

Making use of pareto-optimal solutions, as in this case, seems to be a concept of great interest to handle real-world problems. Starting from different structured solutions the decision-maker becomes responsible to find a compromise between the two conflicting objectives to attain a cost minimal solution on one hand from solving a 2-stage multicommodity WLP and a high service level (on the second stage) on the other hand, integrating a *min-max solution strategy*.

4.2 Model 2

The concept of Model 2 shows a more centralized structure in comparison to Model 1, as it is based on a smaller number of realized distribution centers. The most important modification, however, consists of the fact that an additional step is included into the distribution chain (see Fig. 7). In this case the transportation of the catalogues from the distribution centers to the travel agencies are not directly carried out but it is made use of a *parcel service*. This company organizes the shipments within a seperate distribution system, independent of the locations of the realized distribution centers operated by logistic consolidators.

The remaining planning problem becomes reduced to a(n *uncapacitated*) *multicommodity WLP* (see, e.g., Domschke / Drexl (1996), pp 54) in its basic structure. But by solving this problem some difficulties appear, as the "demands" in the distribution centers are unknown. Therefore, for all types of catalogues an artificial demand l_i^h for each possible distribution center must be defined. So the solution procedure consists of three steps. At first, an *uncapacitated p-center problem* (see, e.g., Domschke / Drexl (1996), pp 137) has to be solved to aggregate the demand of the travel agencies to p nodes. In addition this approach is suitable to include regional deviations of clients' behaviour in the demand structures.

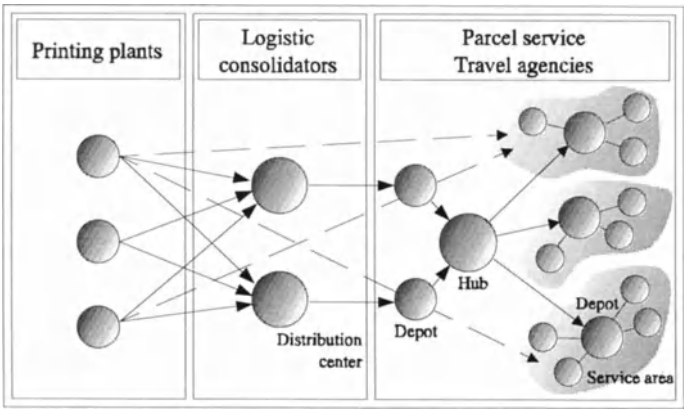


Fig. 6: Basic structure of Model 1

The next step is to define a set W of possible locations for the distribution centers and to assign an artificial demand for the different types of catalogues to each of them. Based on these data a(n *uncapacitated*) *multicommodity WLP* can be solved, which is shown by the following mathematical formulation:

$$\text{Minimize } Z(x,y) = \sum_{k=1}^{q}\sum_{h=1}^{\eta}\sum_{i=1}^{m} c_{ki}^{1} x_{ki}^{1h} + \sum_{i=1}^{m} f_i y_i \qquad (15)$$

subject to:

(2),

$$\sum_{i=1}^{m} x_{ki}^{1h} = l_i^h y_i \qquad \forall \; i = 1, .., m; h = 1, .., \eta \qquad (16)$$

(10) and

$$x_{ki}^{1h}, \geq 0 \qquad \forall \; h = 1, .., \eta; i = 1, ..., m; k = 1, ..., q \qquad (17)$$

Although the location of the distribution centers can basically be determined without including the depot structure of the parcel service, it nevertheless seems to be useful to be taken into consideration. An appropriate structure should not present an efficient distribution system only for the first stage but should also show short distances between a distribution center and its next depot, as such solution leads to a reduction in distribution efforts on the second stage. Therefore, a possible set of locations W can be fixed, mainly based on the locations of the existing depots.

4.3 Comparison of the models

From the basic idea, Model 1 as well as Model 2 satisfy all requirements to attain a distribution structure for the given problem, which is more efficient than the different procedures used at present. It is also assumed for both models that an appropriate information management system is available, especially concerning the interorganisational information flows (see, e.g., Reekers / Smithson (1996)). Comparing these two different models, the advantages and disadvantages of the distribution processes have to be analyzed in detail. Beside the criteria for the design of an efficient distribution structure, which are given in Section 3, economic and ecologic aspects have to be taken into consideration.

Model 1 shows distinct advantages in the organization of the supply on the second stage of the distribution chain. The great number of distribution centers (up to about 100 centers) leads to a clear view on the respective service area and the logistic consolidators are able to organize the transportation of the catalogues autonomously, as they operate separate car fleets. Therefore, it is possible to guarantee a short time of delivery supplying the asigned travel agencies and based on this distribution (and information) structure, a repeat order can be carried out

within a few hours if necessary. It must be seen as an advantage, too, that the disposal of the remainders can be included in the distribution processes without additional cost. The original distribution problem in this case becomes enlarged to a *pick-up-and-delivery problem* (see, e.g., Fisher / Tang / Zhen (1995)).

From the economic point of view, the most important disadvantages of Model 1 depends on the extensive fixed cost which are the consequence of the great number of realized distribution centers and the necessary car fleets. Therefore, the expenditures for such a distribution structure are comperatively on a high level, so that it is impossible in this case to realize the aimed economies-of-scale in a sufficient extend. Furthermore, it must be seen in this context that the original aims of an outsourcing strategy can not be attained which request significant cost reductions in the logistic processes.

Another important aspect which shows distinct disadvantages concerning Model 1 exists in the considerable absence of a short-termed or medium-termed adaptability to alterations in the specific surroundings. As mentioned in Section 3, it must be included in such a concept that the actual volume of transport, which appear from todays function of tour operator catalogues as the main advertising material in the tourism branch, may alter within the forthcoming years. If such a situation evolves the warehouse capacities are under-utilized and the operational efficiency goes down. Moreover, on the basis of the great number of distribution centers it necessitates capital expenditure on a large extend to take charge of additional tasks for diversifying in offered services.

Model 2 shows in comparison to Model 1 lower expenditures to establish the needed infrastructure, because in this case only a small number of (up to 10) distribution centers are realized. Besides this, those fixed cost which depend on operating a separate car fleet at each center become converted to variable cost by employing a parcel service to carry out the supply of the travel agencies. Based on this structure the necessary adaptions, which are set off by a decreasing volume of catalogues, refer to the warehouse capacities only, whereas the reduced transportation volume has no direct consequential effects. For this part of the distribution processes only the shipping contracts must be transformed to the change of frameworks.

However, there are some disadvantages concerning Model 2, which depend on the integration of a parcel service on the last step of the distribution chain. Making use of an external service to carry out the shipments in this step of the distribution processes, the logistic consolidators are unable to influence the operations, because these are part of a general distribution system. Moreover, it must be taken into consideration that the disposal of the remainders can not be handled in the same way as in Model 1. In this case, the collecting procedures represent supplementary shipments which generate separate shipping contracts and therefore additional carrier's charges.

The comparison of these two models shows that the essential advantages follow from Model 2, especially by taking the future developments of the demand structure into consideration which are effected by making use of capable information technologies. When realizing this model the risks appearing from alterations in the demand of tour operator catalogues may be reduced on a large scale.

5 Concluding remarks

The existing structures of the logistic procedures for the distribution of tour operator catalogues show an urgent need for extensive alterations, not only from an economic point of view but also concerning different ecological aspects. These requirements are known to the tour operators as well as to the the travel agencies and they are not in controversy. Nevertheless, at present especially the tour operators make no recognizable efforts to solve these essential problems within a concerted action. On the contrary, some of them approve to these reflections on one hand, but argue on the other hand that it would be better to wait for a running distribution concept which can be joined later. The travel agencies are interested in such a solution considerably, but outgoing from their position they have no opportunity to undertake a leading part to enforce these development processes.

The wait-and-see attitude of tour operators is based on different reasons. At first, many of them have no detailed information about the cost turning up from their logistic operations and also from the obligation to dispose the remainders in accordance with the above mentioned KrW/AbfG. Therefore, they can not calculate the retrenching of expenditure, which depends on alterations in the distribution processes.

Another aspect is the development of an interorganisational information structure which becomes necessary to attain efficient planning procedures and to make use of capable tools for monitoring and control, too. In many cases these structures lead to the fear that competitive operators could be able to receive information about in-company data, especially concerning their clients and their planned strategies.

In spite of all that, at present it is planned to analyze Model 2 in detail and to work out an appropriate concept for a distribution system, which is more efficient in comparison to the existing operational processes and shows cost savings, too. The main steps which have to be dealt with for realizing such a logistic concept are the following:

- *Determination of (formal) structures for the logistic consolidators*

  For the basic structure there are two different solutions. On one hand a cooperative with the tour operators as members can be established or an analogous form of an interorganisational cooperation (see, e.g., Uhlig (1996b)). On the other hand it is also possible to entrust an external service company with all logistic tasks concerning the catalogue distribution. Besides this, travel agencies or their trade association(s) can be also taken into consideration.

- *Selection of a parcel service*

  A parcel service which operates in the whole area of the German federal republic has to be selected. The main criteria are the density of the distribution network and the offered terms of shipment cost.

- *Determination of possible locations and planning data*

  The artificial demands of the distribution centers have to be defined. Based on these results a number of possible locations must be found out and analysed in detail. The existing locations of the hub(s) and depots of the selected parcel service will be used in this case as an appropriate basic structure. Finally the distance and cost matrices between the printing houses and the possible locations has to be calculated.

- *Determination of the distribution centers*

  Based on the established data set an uncapacitated multicommodity WLP has to be solved to set up the locations of the distribution centers.

- *Capital expenditure and operating cost*

  The capital expenditure and the operating cost for the determined distribution structure must be calculated to submit a tender to the tour operators.

The discussions which took place up to now show that a modified structure in the distribution of tour operator catalogues leads to distinct savings on operational cost. Therefore, it seems to be useful to speed up the necessary efforts to alter the existing distribution processes in this branch and to establish an attractive service system. Beside this it must be pointed out that the solution which has been presented for these specific problems allows a transfer to similar logistic problems, too. So it could be of interest to find out additional fields of application to employ these basic ideas.

References:

Ammann, S. / Illing, P. / Sinning, M. (1995): Die Tourismusbranche - Eine segmentspezifische Strukturanalyse: Charakteristika - Erfolgsfaktoren - strategische Herausforderungen. (Forschungskreis Tourismus Management Trier)

***Anonymous* (1995):** Bessere Steuerung würde Probleme lösen. in: FVW Fremdenverkehrswirtschaft International 23/95, 143-145

Broggi, M.K. (1994): Outsourcing von Logistikdienstleistungen. in: Zeitschrift für Logistik 4-5/94, 34-36

Bliesener, M.-M. (1994): Outsourcing als mögliche Strategie zur Kostensenkung. in: BFuP 4/94, 277-290

Campbell, J.F. (1992): Location and allocation for distribution systems with transshipments and transportation economies of scale. in: Annals of Operations Research 40, 77-99

Crainic, T.G. / Dejax, P . / Delorme, L. (1989): Models for multimode multicommodity location problems with interdepot balancing requirements. in: Annals of Operations Research 18, 279-302

Crainic, T.G. / Delorme, L. (1993): Dual-ascent procedures for multicommodity location-allocation problems with balancing requirements. in:TS 27, 90-101

Daduna, J.R. (1985): Engpaßzeitminimierung bei Transport-, Umlade- und Standortproblemen. (Peter Lang) Frankfurt am Main, Bern, New York

Daduna, J.R. / Koch, R. / Maciejewski, P. (1997): Entwicklung eines logistischen Dienstleistungkonzepts für die Belieferung von Reisebüros mit Veranstalterkatalogen. in: Zimmermann, U. / Derigs, U. / Gaul, W. / Möhring, R.H. / Schuster, K.-P. (eds.): Operations Research Proceedings 1996. (Springer) Berlin, Heidelberg, New York, 295-300

Domschke, W. (1995): Logistik: Transport. 4. Aufl. (Oldenbourg) München, Wien

Domschke, W. (1997): Logistik: Rundreisen und Touren. 4. Aufl. (Oldenbourg) München, Wien

Domschke, W. / Drexl, A. (1996): Logistik: Standorte. 4. Aufl. (Oldenbourg) München, Wien

DRV - Deutscher Reisebüro Verband e.V. (Hrsg.) (1995): Tourismusmarkt der Zukunft - Die Entwicklung des Reiseveranstalter- und Reisevermittlermarktes in der Bundesrepublik Deutschland. Frankfurt am Main

Ernst, M. (1994): Zur Bedeutung der Telekommunikation für die Institutionalisierung von Informationsmärkten und deren Wirkung für den Wettbewerb auf internationalen Verkehrsmärkten am Beispiel Tourismus. in: Zeitschrift für Verkehrswissenschaft 65, 89-114

Fischer, E. (1996): Outsourcing von Logistik - Reduzierung der Logistiktiefe zum Aufbau von Kompetenzen. in: Schuh, G. / Weber, H. / Kajüter, P. (Hrsg.): Logistik- Management: Strategische Wettbewerbsvorteile durch Logistik. (Schäffer-Poeschel) Stuttgart 1996, 227-239

Fisher, M.L. / Tang, B. / Zhen, Z. (1995): A network-flow based heuristic for bulk pickup and delivery routing. in: TS 29, 45-61

Gao,L.-L. / Robinson Jr., E.P. (1994): Uncapacitated facility location: General solution procedure and computational experiences. in: EJOR 76, 410-427

Geoffrion, A.M. / Graves, G.W. (1974): Multicommodity distribution system design by Benders decomposition. in: MS 20, 822-844

Golden, B.L. / Assad, A.A. (ed.) (1988): Vehicle routing: Methods and studies. (North-Holland) Amsterdam

Herget, J. (1995): Das betriebliche Informationsmanagement vor einer Neuorientierung. in: NfD 46, 25-32

Kennington, J.L. / Helgason, R.V. (1980): Algorithms for network programming. (Wiley) New York, Chichester, Brisbane, Toronto

Klincewicz, J.G. / Luss, H. (1987): A dual-based algorithm for multiproduct uncapacitated facility location. in: TS 21, 198-206

Klincewicz, J.G. / Luss, H. / Rosenberg, E. (1986): Optimal and heuristic algorithms for multiproduct uncapacitated facility location. in: EJOR 26, 251-258

Laporte, G. (1992): The vehicle routing problem: An overview of exact and approximate algorithms. in: EJOR 59, 345-358

Leitermann, T. (1996): Möglichkeiten bei der Gestaltung des Geschäftsprozesses "Erstellung und Distribution von Reisekatalogen". Diplomarbeit Fachhochschule Konstanz

Lettl-Schröder, M. (1995a): Kosten- und Umweltgründe zwingen zum Handeln. in: FVW Fremdenverkehrswirtschaft International 23/95, 142-143

Lettl-Schröder, M. (1995b): Viele Kataloge pro Buchung. in: FVW Fremdenverkehrswirtschaft International 23/95, 144

Prahalad, C.K. / Hamel, G. (1990): The core competence of the corporation. in: Harvard Business Review 68, 79-91

Puerto, J. / Hinojosa, Y. (1997): The general multicommodity, multilevel, multietapic capacitated plants location problem: Analysis and resolution. Paper presented at the 3rd Operations Research Conference, Havanna, March 10 - 14, 1997

Reekers, N. / Smithson, S. (1996): The role of EDI in inter-organizational coordination in the European automotive industry. in: European Journal of Information Systems 5, 120-130

Rothe, S. (1994): Neue Distributionsstrategien im Tourismus am Beispiel der Reiseveranstalter und Reisebüros. in: Schertler, W. (ed..): Tourismus als Informationsgeschäft. (Wirtschaftsverlag Ueberreuter) Wien, 89-121

Schertler, W. (1994): Informationssystemtechnologie und strategisches Tourismusmanagement. in: Schertler, W. (ed..): Tourismus als Informationsgeschäft. (Wirtschaftsverlag Ueberreuter) Wien, 525-585

Suter, A. (1995): Kernfähigkeiten aktiv managen - strategisch und operativ. in: io Management Zeitschrift 64, Nr. 4, 92-95

Uhlig, T. (1996a): Logistik-Outsourcing: Eine lukrative Tätigkeit für Verlader und Dienstleister. in: Internationales Verkehrswesen 48 / Heft 4, 39-43

Uhlig, T. (1996b): Konzeptioneller Ansatz zur Gestaltung von unternehemensübergreifenden Logistik-Kooperationen. in: Distribution 4/96, 8-11

Wißkirchen, F. (1995): Outsourcing der Distributions-Logistik - Konzeption und Realisierung der Ausgliederung von Logistikleistungen. in: zfo 4/95, 231-236

Obtaining Sharp Lower and Upper Bounds for Two-Stage Capacitated Facility Location Problems

Andreas Klose

Universität St. Gallen, 9000 St. Gallen, Switzerland

Abstract. The *Two-Stage Capacitated Facility Location Problem* (TSCFLP) is to find the optimal locations of depots to serve customers with a given demand, the optimal assignment of customers to depots and the optimal product flow from plants to depots. To compute an optimal solution to the problem, *Benders' decomposition* has been the preferred technique. In this paper, a *Lagrangean heuristic* is proposed to produce good suboptimal solutions together with a lower bound. Lower bounds are computed from the *Lagrangean relaxation* of the capacity constraints. The Lagrangean subproblem is an *Uncapacitated Facility Location Problem* (UFLP) with an additional knapsack constraint. From an optimal solution of this subproblem, a heuristic solution to the TSCFLP is computed by reassigning customers until the capacity constraints are met and by solving the transportation problem for the first distribution stage. The Lagrangean dual is solved by a variant of *Dantzig-Wolfe decomposition*, and elements of *cross decomposition* are used to get a good initial set of dual cuts.

1 Introduction and Problem Formulation

The *Two-Stage Capacitated Facility Location Problem* (TSCFLP) consists in finding the optimal locations of depots to serve customers with a given demand, to optimally assign customers to depots and to determine the product flow from plants to depots. Total costs are made up of the costs to ship the commodity from the plants to the depots, the costs to serve the customers from the depots, the costs of throughput at each depot and the fixed costs of maintaining the depots. To formulate the model mathematically the following notation is used:

I	set of plants i
J	set of potential depot sites j
K	set of customers k
p_i	production capacity of plant i
s_j	capacity of a depot at site j
d_k	demand of customer k
D	total demand ($D = \sum_k d_k$)
t_{ij}	cost of shipping one unit from plant i to a depot at site j
c_{kj}	cost of supplying customer k by a depot at site j
f_j	fixed costs of maintaining a depot at site j

x_{ij} a decision variable denoting the amount shipped from plant i to facility j

z_{kj} a 0-1 variable that takes on the value 1 if customer k is supplied by facility j

y_j a 0-1 variable that will be 1 if a depot at site j is chosen

The problem can then be stated as the following mixed-integer program:

$$v(\text{TSCFLP}) = \min \sum_{i\in I}\sum_{j\in J} t_{ij}x_{ij} + \sum_{k\in K}\sum_{j\in J} c_{kj}z_{kj} + \sum_{j\in J} f_j y_j \tag{1}$$

$$\sum_{j\in J} z_{kj} = 1, \quad \forall\, k \in K \tag{2}$$

$$\sum_{k\in K} d_k z_{kj} \le s_j y_j, \quad \forall\, j \in J \tag{3}$$

$$z_{kj} - y_j \le 0, \quad \forall\, k \in K,\, j \in J \tag{4}$$

$$\sum_{j\in J} s_j y_j \ge D, \tag{5}$$

$$\sum_{j\in J} x_{ij} \le p_i, \quad \forall\, i \in I \tag{6}$$

$$\sum_{i\in I} x_{ij} = \sum_{k\in K} d_k z_{kj}, \quad \forall\, j \in J \tag{7}$$

$$x_{ij} \ge 0, \quad \forall\, i \in I,\, j \in J \tag{8}$$

$$z_{kj} \in \{0,1\}, \quad \forall\, k \in K,\, j \in J \tag{9}$$

$$y_j \in \{0,1\}, \quad \forall\, j \in J\,. \tag{10}$$

The first summation in the objective function (1) gives the total cost of shipping the commodity from the plants to the depots. The second summation corresponds to the costs of supplying the customers from the depots. The coefficients c_{kj} are the costs of transporting an amount of d_k from the depot at site j to the customer site k as well as the costs of handling this amount at the depot j. The last sum in the objective function gives the fixed costs of maintaining the depots. The constraints (2) are the demand constraints. The constraints (3) force z_{kj} to be 0 for all k if $y_j = 0$, and limit the throughput at each depot j not to be greater than its capacity s_j. The constraints (6) are the supply constraints and (7) are the "flow conservation constraints". From the viewpoint of modelling, the constraints (4) and (5) are superfluous. But, the incorporation of these constraints tightens the bound if Lagrangean relaxation of the capacity constraints is applied. Furthermore, the constraint (5) ensures that in a solution of the Lagrangean subproblem, the selected depots have enough capacity to meet all the demand. This is necessary to be able to construct a feasible solution to the original problem from such a solution of the Lagrangean subproblem.

Without loss of generality, it is assumed that $p_i \leq D$ for all $i \in I$, $\sum_i p_i \geq D$, $d_k \leq s_j$ for all $k \in K$ and $j \in J$, $\sum_k d_k \geq s_j$ for all $j \in J$ and $\sum_j s_j \geq D$.

The problem can be formulated in different ways. One alternative formulation is obtained if the capacity constraint (3) is substituted by

$$\sum_{i \in I} x_{ij} \leq s_j y_j \quad \forall\, j \in J\,. \tag{11}$$

On the other hand, the problem can be viewed as a kind of fixed-charge network flow problem with capacity constraints. By introducing variables w_{ijk}, which express the amount transported from plant i over depot j to customer k, the following formulation results.

$$v(\text{TSCFLP}) = \min \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} \bar{c}_{ijk} w_{ijk} + \sum_{j \in J} f_j y_j \tag{12}$$

$$\sum_{j \in J} z_{kj} = 1, \quad \forall\, i \in I \tag{13}$$

$$\sum_{k \in K} d_k z_{kj} \leq s_j y_j, \quad \forall\, j \in J \tag{14}$$

$$z_{kj} - y_j \leq 0, \quad \forall\, k \in K, j \in J \tag{15}$$

$$\sum_{i \in I} w_{ijk} = d_k z_{kj}, \quad \forall\, j \in J, k \in K \tag{16}$$

$$\sum_{j \in J} \sum_{k \in K} w_{ijk} \leq p_i, \quad \forall\, i \in I \tag{17}$$

$$w_{ijk} \geq 0, \quad \forall\, i \in I, k \in K, j \in J \tag{18}$$

$$z_{kj} \in \{0,1\}, \quad \forall\, k \in K, j \in J \tag{19}$$

$$y_j \in \{0,1\}, \quad \forall\, j \in J\,, \tag{20}$$

where $\bar{c}_{ijk} = t_{ij} + c_{kj}/d_k$ is the cost per unit to supply customer k by plant i and depot j. By setting

$$z_{kj} = \frac{1}{d_k} \sum_{i \in I} w_{ijk} \text{ and } x_{ij} = \sum_{k \in K} w_{ijk}$$

it is easy to see, that both formulations are also equivalent with respect to the LP-relaxation. As Geoffrion & Graves (1974) mention, one advantage of the second formulation is a greater flexibility for some applications, since it allows to model transportation costs, which depend on the plant-customer relation. Geoffrion & Graves investigate a multicommodity version of this model and solve it using Benders' decomposition.

The formulations (1)–(10) and (12)–(20), resp., contain "single-source constraints", i. e. each customer's demand must be satisfied by a single depot. In general, this restriction makes sense from the viewpoint of administration,

marketing and accounting. The constraint can be relaxed by substituting (9) by $z_{kj} \geq 0$ for all $k \in K$ and $j \in J$. For fixed y_j, the problem reduces then to a minimum-cost network flow problem, while in the case of single-source constraints it still remains to solve a hard problem even if the y_j are fixed. This strategic character of the variables y_j makes the problem more suitable for branch and bound methods if single-source constraints are missing. An example of such a branch and bound method is the algorithm of Hindi & Basta (1994). They do not add the (trivial) clique constraints (4); this facilitates the computation of an optimal solution of the resulting (weak) LP-relaxation substantially, but at the expense of weaker lower bounds. More promising approaches to solve TSCFLP to optimality seem to be branch and cut algorithms based on polyhedral cuts. Polyhedral results for uncapacitated and capacitated facility location problems have been obtained by Aardal (1994), Aardal et al. (1995, 1994), Cho et al. (1983a, 1983b), Cornuejols et al. (1977), Cornuejols & Thizy (1982), Guignard (1980) and Leung & Magnanti (1989).

In this paper, the computation of sharp lower and upper bounds for the TSCFLP by Lagrangean relaxation is described. If the capacity constraints (3) and (6) are relaxed in a Lagrangean manner, which is equivalent to relaxing the constraints (14) and (17) in the second formulation, an UFLP with an additional knapsack constraint is obtained, which in general can be solved efficiently by branch and bound methods. The best lower bound resulting from this relaxation is computed by applying Dantzig-Wolfe decomposition to the Lagrangean dual. Unfortunately, Dantzig-Wolfe decomposition often shows a bad convergence behaviour. To overcome this difficulty, first different methods to initialize the algorithm have been tried, where some of these methods are based on cross decomposition. Second, a variant of Dantzig-Wolfe decomposition, introduced by Wentges (1994, 1997), which takes a convex combination of the best multipliers generated and the solution of the master problem, has been used. Feasible solutions of the TSCFLP are computed from the solution of the Lagrangean subproblem by applying reassignment heuristics. In the next section, the Lagrangean relaxation approach and the method to solve the Lagrangean dual is described. To this end, elements of cross decomposition, which combines Lagrangean relaxation and Benders' decomposition, are used. Therefore, the application of these decomposition methods to the TSCFLP is described briefly in Sect. 3 and Sect. 4, before the Lagrangean relaxation algorithm is described in detail in Sect. 5. Finally, in Sect. 6, some computational results are presented.

2 Lagrangean Relaxation of Capacity Constraints

Lower bounds for a mixed-integer programming problem (MIP)

$$v(\text{MIP}) = \min\ cx + fy \tag{21}$$

$$A_1 x + A_2 y \geq b \tag{22}$$

$$B_1 x + B_2 y \geq d \tag{23}$$

$$x \geq 0,\ y \in S, \tag{24}$$

where $S \neq \emptyset$ is a set of nonnegative integer vectors satisfying some side constraints, can be obtained by dropping some "complicating" constraints $A_1 x + A_2 y \geq b$ and incorporating them into the objective function with a penalty term $\eta(b - A_1 x - A_2 y)$, which yields the MIP

$$v(\mathrm{SD}_\eta) = \min \{cx + fy + \eta(b - A_1 x - A_2 y) : (x, y) \in F_{\mathrm{SD}}\}\ , \tag{25}$$

where

$$F_{\mathrm{SD}} = \{(x, y) : (23) \text{ and } (24)\}\ .$$

The program above is known as *Lagrangean relaxation* and gives a lower bound for $v(\mathrm{MIP})$ for every $\eta \geq 0$ (i.e. $v(\mathrm{SD}_\eta) \leq v(\mathrm{MIP})\ \forall\, \eta \geq 0$). For fixed η it is also called *Lagrangean subproblem.* The best lower bound available from the relaxation (25), is obtained from the optimal solution of the so-called *Lagrangean dual*

$$v(\mathrm{LD}) = \max_{\eta \geq 0} v(\mathrm{SD}_\eta)\ . \tag{26}$$

Since every MIP can be rewritten as a linear program by convexification, the linear program

$$v(\mathrm{MD}_{T_D}) = \max\ \eta_0 \tag{27}$$

$$\eta_0 \leq cx^t + fy^t + \eta(b - A_1 x^t - A_2 y^t), \quad \forall\, t \in T_D \tag{28}$$

$$\eta \geq 0\ , \tag{29}$$

where $\{(x^t, y^t) : t \in T_D\}$ is the set of all extreme points of the convex hull of F_{SD}, is equivalent to the Lagrangean dual (i.e. $v(\mathrm{LD}) = v(\mathrm{MD}_{T_D})$), if it is assumed for simplicity that F_{SD} is nonempty and bounded (see e.g. Nemhauser & Wolsey (1988)).

The linear program (27)–(29) is also known as *dual master problem.* It can be solved by *Dantzig-Wolfe decomposition.* That means, the constraints (28) – which are called *dual cuts* – are relaxed, and only a small subset $T_D^h \subseteq T_D$ is considered in every iteration h. Additional dual cuts are generated by solving the Lagrangean subproblem (25) with the Lagrangean multipliers η set to $\eta = \eta^h$, where η^h is the optimal solution of the relaxed master program with cut set T_D^h. This process is repeated until the upper bound $v(\mathrm{MD}_{T_D^h})$ for $v(\mathrm{LD})$ coincides with the best lower bound obtained from the solutions of the Lagrangean relaxation. The finiteness of the approach results from the following argument: If (η_0^h, η^h) is an optimal solution of the relaxed master problem with cut set $T_D^h \subset T_D$, then

$$v(\mathrm{MD}_{T_D^h}) = \eta_0^h \leq cx^t + fy^t + \eta^h(b - A_1 x^t - A_2 y^t) \quad \forall\, t \in T_D^h$$

holds. On the other hand, from an optimal solution (x^τ, y^τ), $\tau \in T_D$, of the Lagrangean subproblem (25) with $\eta = \eta^h$, one obtains

$$\begin{aligned} v(\mathrm{SD}_{\eta^h}) &= \min\{cx + fy + \eta^h(b - A_1x - A_2y) : (x, y) \in F_{\mathrm{SD}}\} \\ &= cx^\tau + fy^\tau + \eta^h(b - A_1x^\tau - A_2y^\tau) \\ &\leq cx^t + fy^t + \eta^h(b - A_1x^t - A_2y^t) \quad \forall\, t \in T_D\,. \end{aligned}$$

Therefore, either $v(\mathrm{SD}_{\eta^h}) = v(\mathrm{MD}_{T_D^h}) = v(\mathrm{LD})$, or $\tau \notin T_D^h$ and (x^τ, y^τ) yields a new dual cut. Since T_D is finite, the optimal solution of (27)–(29) is found in a finite number of steps by this procedure.

To compute lower bounds for the TSCFLP and to obtain heuristic solutions, Lagrangean relaxation of the capacity constraints is used. Relaxing these constraints with multipliers λ_i $(i \in I)$ for the supply constraints (6) and multipliers μ_j $(j \in J)$ for the depot capacity constraints (3), yields the Lagrangean subproblem

$$\begin{aligned} v(\mathrm{SD}_{\lambda,\mu}) = \min\Bigg\{ &\sum_{i\in I}\sum_{j\in J} \tilde{t}_{ij}x_{ij} + \sum_{k\in K}\sum_{j\in J} \tilde{c}_{kj}z_{kj} + \sum_{j\in J} \tilde{f}_j y_j \;:\; (x, y, z) \in F_{\mathrm{SD}} \Bigg\} \\ &- \sum_{i\in I} p_i\lambda_i \end{aligned} \tag{30}$$

with the feasible region

$$F_{\mathrm{SD}} = \{(x, y, z) : (2), (4), (5), (7), (8), (9), (10)\} \tag{31}$$

and

$$\tilde{t}_{ij} = t_{ij} + \lambda_i,\ \tilde{c}_{kj} = c_{kj} + d_k\mu_j,\ \tilde{f}_j = f_j - \mu_j s_j\,.$$

Now, the first and second stage of distribution are only connected by the constraints (7) of flow conservation. Since there are no restrictions, neither on the capacities of the depots, nor on the supply capacities of the plants, there always exists an optimal solution of (30), where a depot at site j is only supplied by its "cheapest source". Therefore, one can set

$$x_{ij} = \begin{cases} \sum\limits_{k\in K} d_k z_{kj}\,, \text{ for } i = \arg\min\{\tilde{t}_{lj} : l \in I\} \\ 0 \qquad\qquad\quad , \text{otherwise} \end{cases} \tag{32}$$

$$c^*_{kj} = \tilde{c}_{kj} + d_k \min_{i\in I} \tilde{t}_{ij}\,, \tag{33}$$

and the relaxation (30) reduces to the binary optimization problem

$$\begin{aligned} v(\mathrm{SD}_{\lambda,\mu}) = \min\Bigg\{ &\sum_{k\in K}\sum_{j\in J} c^*_{kj}z_{kj} + \sum_{j\in J} \tilde{f}_j y_j \;:\; (2), (4), (5), (9), (10) \Bigg\} \\ &- \sum_{i\in I} p_i\lambda_i\,, \end{aligned} \tag{34}$$

which is an UFLP with an additional knapsack constraint. The corresponding Lagrangean dual is given by

$$\max \ \lambda_0 \tag{35}$$

$$\lambda_0 \leq v(x^t, y^t, z^t) - \sum_{i \in I} N_i(x^t)\lambda_i - \sum_{j \in J} M_j(y^t, z^t)\mu_j, \quad \forall\, t \in T_D \tag{36}$$

$$\lambda_i \geq 0, \quad \forall\, i \in I \tag{37}$$

$$\mu_j \geq 0, \quad \forall\, j \in J \tag{38}$$

$$\lambda_0 \in \mathbb{R}, \tag{39}$$

where

$$\{(x^t, y^t, z^t) \mid t \in T_D\}$$

is the set of extreme points of the convex hull of F_{SD} in (31),

$$v(x, y, z) = \sum_{i \in I} \sum_{j \in J} t_{ij} x_{ij} + \sum_{k \in K} \sum_{j \in J} c_{kj} z_{kj} + \sum_{j \in J} f_j y_j$$

is the objective function value of a solution (x, y, z) to (30) and

$$M_j(y, z) = s_j y_j - \sum_{k \in K} d_k z_{kj} \text{ and } N_i(x) = p_i - \sum_{j \in J} x_{ij}$$

are the values of the "slack variables" corresponding to the constraints (3) and (6) in a solution (x, y, z) to the Lagrangean subproblem (30).

The computation of optimal multipliers for the relaxation (30) requires to solve an UFLP of the type (34) in each iteration. In many cases, this can be done efficiently by branch and bound methods, which use subgradient optimization for lower bounding (see Klose (1994)). But, since the UFLP is NP-hard (see Krarup (1983)), it is important to keep the number of iterations required small. Unfortunately, the approach of Dantzig-Wolfe decomposition often suffers from the problem of poor convergence. A problem closely related to this, is the identification of a good initial set of dual cuts, such that the relaxed dual master problem is bounded and gives meaningful multipliers. Some possibilities to derive initial cuts for the master problem (35)–(39) are discussed in Sect. 4. To further improve the convergence behaviour, an idea of Wentges (1994, 1997) can be used. He proposes to use the convex combination

$$\overline{\eta}^h = \alpha_h \eta^h + (1 - \alpha_h)\eta^* \qquad (0 < \alpha_h < 1) \tag{40}$$

as multipliers for the next Lagrangean subproblem (25), where η^h denotes the optimal solution of the relaxed master problem (27)–(29) in iteration h with cut set T_D^h and η^* the best multipliers obtained so far. The idea to weigh the generated multipliers is related to the approach of linear mean value cross decomposition (see Holmberg (1992b)). In contrast to conventional Dantzig-Wolfe decomposition, it is possible, that a cut already contained in T_D^h is

duplicated by the solution of the Lagrangean subproblem. But in this case, the lower bound must improve, as can be seen from the following argument (see Wentges (1994, 1997)): If an optimal solution (x^τ, y^τ) of (25) with $\eta = \overline{\eta}^h$ yields a cut already contained in T_D^h, then

$$\begin{aligned} v(\mathrm{SD}_{\overline{\eta}^h}) &= cx^\tau + fy^\tau + \overline{\eta}^h(b - A_1x^\tau - A_2y^\tau) \\ &= \alpha_h \left(cx^\tau + fy^\tau + \eta^h(b - A_1x^\tau - A_2y^\tau)\right) \\ &\quad + (1-\alpha_h)\left(cx^\tau + fy^\tau + \eta^*(b - A_1x^\tau - A_2y^\tau)\right) \\ &\geq \alpha_h \left(cx^\tau + fy^\tau + \eta^h(b - A_1x^\tau - A_2y^\tau)\right) + (1-\alpha_h)v(\mathrm{SD}_{\eta^*}) \end{aligned}$$

since $v(\mathrm{SD}_{\eta^*}) \leq cx^t + fy^t + \eta^*(b - A_1x^t - A_2y^t)$ for all $t \in T_D$. From

$$v(\mathrm{MD}_{T_D^h}) = \eta_0^h \leq cx^t + fy^t + \eta^h(b - A_1x^t - A_2y^t) \quad \forall\, t \in T_D^h$$

it follows

$$v(\mathrm{SD}_{\overline{\eta}^h}) \geq \alpha_h v(\mathrm{MD}_{T_D^h}) + (1-\alpha_h)v(\mathrm{SD}_{\eta^*})$$

and the lower bound improves at least by

$$\alpha_h \left(v(\mathrm{MD}_{T_D^h}) - v(\mathrm{SD}_{\eta^*})\right) .$$

Since T_D is finite and the set F_{SD} is bounded by assumption, the approach yields an ϵ-optimal solution of the Lagrangean dual in a finite number of steps. A more detailed description and proofs of further properties of the procedure can be found in the work of Wentges (1994, 1997).

Beside Dantzig-Wolfe decomposition, other techniques like subgradient optimization or Lagrangean ascent can be used to solve the Lagrangean dual. Subgradient optimization uses only the information of the last dual cut and in general requires many iterations to obtain good multipliers. Therefore, the method is better suited in the case of a Lagrangean relaxation, which is easy to solve. Lagrangean ascent methods (see Guignard & Rosenwein (1989) and Guignard & Opaswongkarn (1990)) try to increase the lower bound in each step by varying only one or few multipliers. To this end, some post-optimal analysis is necessary to determine by which amount the multipliers can be changed without altering the optimal solution of the Lagrangean relaxation. With respect to the TSCFLP and the Lagrangean relaxation (30) of the capacity constraints such an analysis is difficult and computationally burdensome.

3 Benders' Decomposition

Benders' decomposition was introduced by Benders (1962) to solve mathematical programming problems with mixed variables. The method has been applied with good success to a multicommodity version of the TSCFLP by Geoffrion & Graves (1974).

The technique is quite simple and allows the incorporation of various side constraints. The MIP (21)–(24) can be written as

$$v(\text{MIP}) = \min \{v(\text{SP}_y) \,:\, y \in S\}$$

where

$$v(\text{SP}_y) = fy + \min \{cx \,:\, A_1 x \geq b - A_2 y,\, B_1 x \geq d - B_2 y,\, x \geq 0\} \;. \quad (41)$$

For given $y \in S$, the program (41) is known as *primal subproblem.* Since it is a linear program, it can be dualized with dual variables η and ν, yielding

$$\begin{aligned} v(\text{SP}_y) = \max \; & (f - \eta A_2 - \nu B_2)\, y + \eta b + \nu d \\ & \eta A_1 + \nu B_1 \leq c \\ & \eta \geq 0,\, \nu \geq 0 \,. \end{aligned}$$

If it is assumed for simplicity that (41) has a feasible solution for every $y \in S$, then

$$v(\text{SP}_y) = \max \left\{ \left(f - \eta^t A_2 - \nu^t B_2\right) y + \eta^t b + \nu^t d \,:\, t \in T_P \right\} ,$$

where $\{(\eta^t,\, \nu^t) \,:\, t \in T_P\}$ denotes the set of all extreme points of the dual program above. Therefore, the MIP can be rewritten as

$$\min \; y_0 \quad (42)$$

$$y_0 \geq \left(f - \eta^t A_2 - \nu^t B_2\right) y + \eta^t b + \nu^t d, \quad \forall\, t \in T_p \quad (43)$$

$$y \in S,\; y_0 \in \mathbb{R} \,. \quad (44)$$

The reformulation (42)–(44) is known as *Benders' master problem* and solved by row generation, i. e. the constraints (43) – which are called *Benders' cuts* – are first relaxed and then successively added to the (relaxed) master problem by solving the primal subproblem. This process is repeated until the best upper bound obtained from the solutions of the primal subproblem coincides with the lower bound obtained from the solution of the relaxed master problem. The finiteness of the approach follows from a similar argument as the finiteness of Dantzig-Wolfe decomposition for the dual master problem (see e. g. Nemhauser & Wolsey (1988)).

If Benders' decomposition is applied to the TSCFLP, the primal subproblem is given by the transportation problem

$$v(\text{SP}_{(y,z)}) = \sum_{j \in J} \Big(f_j y_j + \sum_{k \in K} c_{kj} z_{kj} \Big) + \min \Big\{ \sum_{i \in I} \sum_{j \in J} t_{ij} x_{ij} \,:\, (6)\text{–}(8) \Big\} \quad (45)$$

with the corresponding dual program

$$v(\text{SP}_{(y,z)}) = \max \sum_{k \in K} \sum_{j \in J} (c_{kj} + \omega_j d_k) z_{kj} + \sum_{j \in J} f_j y_j - \sum_{i \in I} p_i \lambda_i$$

$$\begin{aligned} \omega_j - \lambda_i &\leq t_{ij}, \quad \forall\, i \in I,\, j \in J \\ \lambda_i &\geq 0, \quad \forall\, i \in I \\ \omega_j &\in \mathbb{R}, \quad \forall\, j \in J\,, \end{aligned}$$

where $(y,\, z) \in F_{\mathrm{SP}} = \{(y,\, z) : (2)\text{–}(4)\}$. Since the primal subproblem (45) has a feasible solution for every $(y,\, z) \in F_{\mathrm{SP}}$, the Benders' reformulation of the TSCFLP is given by the master problem

$$\min \quad y_0 \tag{46}$$

$$y_0 \geq \sum_{k \in K} \sum_{j \in J} (c_{kj} + \omega_j^t d_k) z_{kj} + \sum_{j \in J} f_j y_j - \sum_{i \in I} p_i \lambda_i^t \quad \forall\, t \in T_P \tag{47}$$

$$(y,\, z) \in F_{\mathrm{SP}},\ y_0 \in \mathbb{R}\,, \tag{48}$$

where $\{(\lambda^t,\, \omega^t) : t \in T_P\}$ is the set of extreme points of the dual program above. Benders' cuts can be derived from an optimal solution $(\lambda^*,\, \omega^*)$ of the transportation problem (45), with the set J of facilities restricted to the set $O = \{j \in J : y_j = 1\}$ of open facilities, by setting $\lambda_i = \lambda_i^*$ for all $i \in I$ and

$$\omega_j = \begin{cases} \omega_j^* & , \text{for } j \in O \\ \min\{t_{ij} + \lambda_i^* : i \in I\} & , \text{for } j \in J \setminus O\,. \end{cases}$$

The algorithmic problems associated with the method of Benders' decomposition are the efficient solution of the master problem, which is a linear mixed-integer programming problem, and the determination of "strong" Benders' cuts. The master problem (46)–(48) can be reformulated as the problem of finding an integer solution $(y,\, z) \in F_{\mathrm{SP}}$, which meets the constraints (47) with strict inequality for all generated Benders' cuts $t \in T_p^h \subseteq T_P$, if the variable y_0 is set to the best upper bound z_B found so far. This feasibility problem again, can be rewritten as a pure binary linear programming problem (see Geoffrion & Graves (1974) on this so-called ϵ-method) and solved by heuristic methods as long as one succeeds in finding an integer solution, which meets the requirements above. Another proposal is to first relax the integer requirements and then to generate Benders' cuts, which are also valid for the LP-relaxation, until a certain number of iterations has been reached or the LP has been solved (see McDaniel & Devine (1977)). Furthermore, because of the primal degeneration of transportation problems, there is some freedom in the choice of Benders' cuts. To derive "stronger" cuts, one can try to alter some of the dual variables λ_i and ω_j without changing the dual objective function value $v(\mathrm{SP}_{(y,z)})$ and without loosing dual feasibility. The flexibility in the choice of Benders' cuts can be enlarged if superfluous constraints as $x_{ij} \leq s_j y_j$ for all $i \in I$ and $j \in J$ are added. If the dual variables corresponding to these constraints are denoted by π_{ij}, a Benders' cut is given by

$$y_0 \geq \sum_{j \in J} \Bigg(\sum_{k \in K} (c_{kj} + \omega_j d_k) z_{kj} + \Big(f_j - s_j \sum_{i \in I} \pi_{ij} \Big) y_j \Bigg) - \sum_{i \in I} \lambda_i p_i\,,$$

where the dual variables λ, ω and π must be feasible (i. e. $\omega_j - \lambda_i - \pi_{ij} \leq t_{ij}$ for all $i \in I$ and $j \in J$) and choosen in such a way, that the right hand side of the constraint equals the objective function value $v(SP_{y,z})$ of the actual primal subproblem. A formal theory of strong Benders' cuts and methods to derive such cuts can be found in Magnanti & Wong (1981, 1989) and Wentges (1996).

4 Cross Decomposition

The method of *cross decomposition* to solve mixed-integer programming problems has been introduced and applied to the *Capacitated Facility Location Problem* (CFLP) by Van Roy (1980, 1983, 1986). The method has been further extended to linear, nonlinear and pure integer programming problems by Holmberg (1990, 1992a, 1994).

The technique combines the methods of Lagrangean relaxation and Benders' decomposition. The idea is to avoid the solution of master problems as long as possible. This is done by solving the Lagrangean subproblem (25) instead of the Benders' master problem (42)–(44) to generate a new integer solution, while new Lagrangean multipliers are obtained from an optimal dual solution of the primal subproblem (41). This *subproblem* phase is continued until it becomes apparent that this procedure does not converge against the optimal solution of the Lagrangean dual (26) or the MIP (21)–(24). This is the case, if an optimal dual solution $(\overline{\eta}, \overline{\nu})$ of the actual primal subproblem (41) violates one of the constraints

$$z_D < cx^t + fy^t + \overline{\eta}(b - A_1 x^t - A_2 y^t), \quad \forall\, t \in T_D^h\,, \tag{49}$$

where z_D is the best lower bound obtained so far and T_D^h the set of generated dual cuts. Accordingly, the subproblem phase cannot converge, if an optimal solution $\overline{y}$ of the actual Lagrangean subproblem (25) does not satisfy the constraints

$$z_B > \left(f - \eta^t A_2 - \nu^t B_2\right)\overline{y} + \eta^t b + \nu^t d, \quad \forall\, t \in T_P^h\,, \tag{50}$$

where z_B is the best upper bound obtained so far and T_P^h is the set of generated Benders' cuts. It is easy to see, that the dual solution $\overline{\eta}$ cannot lead to an improved lower bound, if (49) does not hold, and that the integer solution $\overline{y}$ cannot lead to an improved upper bound, if (50) is violated (for a proof see Van Roy (1983) or Holmberg (1990)). If this is the case, a Benders' master or dual master problem must be solved to restart the subproblem phase.

The method of cross decomposition requires that the problem under question has a primal and dual exploitable structure in such a way, that the Lagrangean subproblem is a relaxation of the Benders master problem and that the dual of the primal subproblem is a relaxation of the dual master problem. Therefore, a natural way to apply cross decomposition to the TSCFLP

is to define the Lagrangean subproblem by relaxing only the supply constraints (6). In this case, the Lagrangean relaxation is a CFLP with *single sourcing*, which is a very hard problem. Therefore, this way is not applicable. On the other hand, if the capacity constraints (3) are relaxed as well, dual variables correponding to these constraints cannot be generated from the transportation problem (45). At first sight, this difficulty seems to be eliminated, if one substitutes the constraints (3) by the constraints (11). If the constraints (6) and (11) are relaxed, the Lagrangean subproblem is still given by (30), and the resulting primal subproblem, which reads as

$$v(\mathrm{SP}_{(y,z)}) = \sum_{j\in J}\Big(f_j y_j + \sum_{k\in K} c_{kj} z_{kj}\Big) + \min\Big\{\sum_{i\in I}\sum_{j\in J} t_{ij} x_{ij} \;:\; (6)\text{–}(8),\ (11)\Big\}, \tag{51}$$

is now a relaxation of the resulting dual master problem. The linear program (51) is infeasible, if the solution (y, z) of the Lagrangean subproblem (30) violates one of the capacity constraints (3). Regarding the general MIP (21)–(24), this might not cause any problems, as Holmberg (1990) has shown: If the primal subproblem (41) is infeasible, an extreme ray

$$(\eta^r,\, \nu^r) \in \{(\eta,\, \nu) \;:\; \eta A_1 + \nu B_1 \leq 0\}$$

of the dual solution space is identified. This gives the "feasibility cut"

$$0 \geq \eta^r b + \nu^r d - (\eta^r A_2 + \nu^r B_2)\, y\ ,$$

which has to be added to the Benders' master problem. A new integer solution can then be generated by solving a "modified" Lagrangean subproblem, which is obtained if in (25) c and f are set to zero and η is set to η^r. In addition to the convergence tests (49) and (50), one has to check if a solution of the Lagrangean subproblem (25) satisfies all generated feasibility cuts. In the case of the TSCFLP, if the capacity of a depot $l \in J$ is violated by the solution $(y,\, z)$ of the Lagrangean subproblem, extreme rays of the dual of (51) are given by

$$\lambda_i = 0\ \forall\, i \in I,\ \omega_j = \mu_j = 0\ \forall\, j \in J \setminus \{l\} \text{ and } \omega_l = \mu_l = 1\ ,$$

where λ, ω and μ are dual variables corresponding to the constraints (6), (7) and (11). The feasibility cut is just the depot capacity constraint corresponding to depot l. A solution of the modified Lagrangean subproblem is to assign no customer to depot l, while all other customers can be arbitrarily assigned to the other depots; and the additional convergence test consists in checking, if a capacity constraint (3) has been violated a second time. Clearly, one could not expect from such an approach that ping-ponging between subproblems can be done for a sufficient number of steps. Rather, the method would degenerate to Benders' or Dantzig-Wolfe decomposition, resp., since it is often necessary to solve a master problem.

To summarize, the cross decomposition algorithm seems not to be applicable in a conventional way to solve the TSCFLP with single sourcing. On the other hand, the method can be used to get initial dual cuts for the dual master problem (35)–(39) in the context of Dantzig-Wolfe decomposition. To this end, the capacity constraints (3) are ignored and the subproblem phase of cross decomposition is applied to this relaxed problem – which is denoted by TSCFLP* – until one of the convergence criteria is not met. In this case, one has to switch to Dantzig-Wolfe Decomposition. The primal subproblem, which one has to solve in this initial phase, is the transportation problem (45). From an optimal dual solution $(\overline{\lambda}, \overline{\omega})$, a Benders' cut (47) is constructed. The dual variables $\overline{\lambda}_i$ $(i \in I)$ are then used as Lagrangean multipliers in the Lagrangean subproblem (30), while the multipliers μ_j $(j \in J)$ are set to zero. An optimal solution $(\overline{x}, \overline{y}, \overline{z})$ of (30) then defines a dual cut (36). Denote by z_D the best lower bound obtained so far in this way, by $\tilde{z}_B$ the objective function value of the best solution found to the relaxed problem TSCFLP*, and by T_D^h and T_P^h the index sets of the generated dual cuts and Benders' cuts, resp. The convergence tests (49) and (50) are then given by

$$z_D < v(x^t, y^t, z^t) - \sum_{i \in I} N_i(x^t)\overline{\lambda}_i \quad \forall\, t \in T_D^h \tag{52}$$

and

$$\tilde{z}_B > \sum_{k \in K} \sum_{j \in J} (c_{kj} + d_k \omega_j^t)\overline{z}_{kj} + \sum_{j \in J} f_j \overline{y}_j - \sum_{i \in I} p_i \lambda_i^t \quad \forall\, t \in T_P^h\,. \tag{53}$$

If the dual variables $\overline{\lambda}$, obtained from the optimal dual solution of the primal subproblem (45), or the binary variables $(\overline{y}, \overline{z})$, obtained from an optimal solution of the Lagrangean subproblem (30), do not meet the constraints (52) or (53), resp., the subproblem phase is finished and Dantzig-Wolfe decomposition is started.

Furthermore, the optimal solution of the primal subproblem can be used to define dual cuts. Let the actual optimal solution of the Lagrangean subproblem (30) be $(x, y, z) = (\overline{x}, \overline{y}, \overline{z})$ and denote the optimal primal solution of the transportation problem by $\tilde{x}$. Since in an optimal solution of the Lagrangean relaxation (30) every open depot is served exactly by one plant, the solution $\tilde{x}$ must be represented as convex combination of solutions x^t with

$$x_{ij}^t \in \Big\{0, \sum_{k \in K} d_k \overline{z}_{kj}\Big\} \text{ and } \sum_{i \in I} x_{ij}^t = \sum_{k \in K} d_k \overline{z}_{kj} \;\; \forall\, j \in J\,.$$

If such a convex combination is given by

$$\tilde{x} = \sum_{t \in T_E} \alpha_t x^t \text{ where } \sum_{t \in T_E} \alpha_t = 1 \text{ and } \alpha_t > 0 \;\; \forall\, t \in T_E\,,$$

dual cuts as defined in (36) can be derived from the solutions $(x^t, \overline{y}, \overline{z})$ for every $t \in T_E$. As Van Roy (1986) shows, the set T_E of cuts defines an *efficient*

set of cuts for the "Dantzig-Wolfe reformulation" of the primal subproblem (45). (A set $T_E \subseteq T$ of cuts is called *efficient*, if it is a minimal subset of the set T of all cuts, such that $v(M_{T_E}) = v(M_T)$, where $v(M_{T'})$ is the value of the problem including only the cuts $t \in T'$.) Adding these cuts to the (relaxed) dual master program, with the additional restriction that the multipliers μ_j have to be 0, ensures that the program is bounded and leads to a value less than or equal to $v(SP_{(\overline{y},\overline{z})})$. To find the above convex combination and thereby to derive dual cuts from the solution of the primal subproblem, Van Roy (1986) proposes the following algorithm.

Algorithm 1: Dual cuts from the primal subproblem (Van Roy (1986))

Notation:
$(\overline{y}, \overline{z})$ actual solution of Lagrangean subproblem (34),
$\tilde{x}$ solution of transportation problem,
T_D^h set of generated dual cuts.

Initialization:
For all $j \in J$ **Do** $D_j := \sum_{k \in K} d_k \overline{z}_{kj}$;
$O := \{j \in J : D_j > 0\}$;
$Stop := false$;
For all $j \in O$ **Do**
  $I_j := \{i \in I : \tilde{x}_{ij} > 0\}$;
  For all $i \in I_j$ **Do** $\tilde{x}_{ij} := \dfrac{\tilde{x}_{ij}}{D_j}$;
Endfor;

Main loop:
$t := |T_D^h|$;
While Not *Stop* **Do**
  $t := t + 1$;
  For all $j \in O$ **Do** $i(j) := \arg\min \{\tilde{x}_{ij} : i \in I_j\}$;
  $j^* := \arg\min \{\tilde{x}_{i(j),j} : j \in O\}$;
  $\alpha_t := \tilde{x}_{i(j^*),j^*}$;
  $y^t := \overline{y}$;
  $z^t := \overline{z}$;
  For all $j \in O$ **Do**
    For all $i \in I_j$ **Do**
      If $i = i(j)$ **Then**
        $x_{ij}^t := D_j$
        $\tilde{x}_{ij} := \tilde{x}_{ij} - \alpha_t$
      Else
        $x_{ij}^t := 0$;
      Endif;
    Endfor;
  Endfor;

From the solution (x^t, y^t, z^t) generate a dual cut as in (36);
$T_D^h := T_D^h \cup \{t\}$;
$I_{j^*} := I_{j^*} \setminus \{i(j^*)\}$;
$Stop := I_{j^*} = \emptyset$;
Endwhile;

5 The Lagrangean Heuristic

In this section, the Lagrangean relaxation approach to solve the TSCFLP is described in detail. The entire approach consists in the following parts:

- the heuristic to compute a feasible solution to the TSCFLP from a solution of the Lagrangean subproblem (34),
- the procedure to get an initial set of dual cuts, and
- the Dantzig-Wolfe Decomposition approach to solve the Lagrangean dual (35)–(39).

5.1 Computation of Feasible Solutions

Denote by $(\overline{y}, \overline{z})$ an optimal solution of the Lagrangean subproblem (34) for given multipliers λ and μ and by $O = \{j : \overline{y}_j = 1\}$ the set of depot sites selected in this solution. To create a feasible solution to the TSCFLP from this solution, one can proceed in two steps.

Step 1: Find a good feasible solution to the *Generalized Assignment Problem* (GAP)

$$v(\text{GAP}) = \min \sum_{k \in K} \sum_{j \in O} c^*_{kj} z_{kj} \tag{54}$$

$$\sum_{j \in O} z_{kj} = 1, \quad \forall\, k \in K \tag{55}$$

$$\sum_{k \in K} d_k z_{kj} \leq s_j, \quad \forall\, j \in O \tag{56}$$

$$z_{kj} \in \{0,1\}, \quad \forall\, k \in K, j \in O\,, \tag{57}$$

where c^*_{kj} is defined by (33).

Step 2: If a feasible solution $\tilde{z}$ to the GAP has been found, solve the transportation problem

$$v(\text{TP}) = \min \sum_{i \in I} \sum_{j \in O} t_{ij} x_{ij} \tag{58}$$

$$\sum_{i \in I} x_{ij} = D_j, \quad \forall\, j \in O \tag{59}$$

$$\sum_{j \in O} x_{ij} \leq p_i, \quad \forall\, i \in I \tag{60}$$

$$x_{ij} \geq 0, \quad \forall\, i \in I, j \in O\,, \tag{61}$$

where $D_j = \sum_{k \in K} d_k \tilde{z}_{kj}$. The objective function value z_H of the resulting solution with respect to the TSCFLP is then given by

$$z_H = v(TP) + \sum_{k \in K} \sum_{j \in O} c_{kj} \tilde{z}_{kj} + \sum_{j \in O} f_j\,.$$

Since solving the transportation problem is easily done, the main problem is to find a good feasible solution to the GAP, which is an NP-hard problem and difficult to solve to optimality (for an overview on the GAP, and exact and heuristic solution methods see e. g. Martello & Toth (1990) and Osman (1995)). Francis et al. (1992) mention the following approach to find a solution to the GAP. Relax the integer requirements and solve the resulting transportation problem. Then try to find a degenerate solution of the transportation problem, which meets the requirements of single sourcing, by performing a number of pivoting steps. Finally, if a solution is found in this way, try to improve it by reassigning customers. This procedure is an adaption of the heuristic of Balas & Martin (1980) for linear binary optimization problems to this special problem. But, since the GAP is only a subproblem, which has to be solved repeatedly in the overall algorithm, this approach is computational expensive.

A simpler method is to use a reassignment procedure. Such approaches have been proposed by Barcelo & Casanovas (1984), Klincewicz & Luss (1986) and Barcelo et al. (1991) in the context of Lagrangean heuristics for the CFLP with single sourcing. Barcelo & Casanovas propose the following heuristic. Denote by

$$K_j = \{k \in K : \overline{z}_{kj} = 1\}$$

the set of customers assigned to facility j in an optimal solution of (34) and by

$$V = \left\{ j \in O : r_j = s_j - \sum_{k \in K_j} d_k < 0 \right\}$$

the set of open facilities whose capacity is violated. Customers $k \in K_j$, with $j \in V$, are then reassigned to an open facility $h(k)$ according to the optimal solution of the knapsack problem

$$\min\left\{ \sum_{k \in K_j} \rho_{h(k)} z_k \;:\; \sum_{k \in K_j} d_k z_k \geq -r_j,\; z \in \{0,1\}^{|K_j|} \right\}. \tag{62}$$

In (62), $\rho_{h(k)}$ estimates the costs of reassigning customer k. Barcelo & Casanovas use the following estimate

$$\rho_{h(k)} = \min_{l \in O \setminus V} \left\{ c^*_{kl} - c^*_{kj} \;:\; r_l \geq d_k \right\} \quad (k \in K_j)\,.$$

The variable z_k in (62) takes on the value 1 if customer $k \in K_j$ has to be reassigned from facility j to facility $h(k)$. If some customers have been reassigned, the set V and the costs $\rho_{h(k)}$ are redefined, and the procedure is repeated until $V = \emptyset$ or no reassignment has been made. The heuristics of Barcelo et al. and Klincewicz & Luss are similar, except that the knapsack problem (62) is solved heuristically. Algorithm 2 gives a pseudo–pascal code of the reassignment heuristic of Barcelo et al.

Algorithm 2: Reassignment heuristic (Barcelo et al. (1991))

Notation:
$(\overline{y}, \overline{z})$ optimal solution of (34),
O set of open facilities, $O = \{j \in J : \overline{y}_j = 1\}$,
K_j set of customers assigned to depot j, $K_j = \{k \in K : \overline{z}_{kj} = 1\}$,
c^*_{kj} defined by (33).

Initialization:
For all $j \in O$ **Do**
$D_j = \sum_{k \in K_j} d_k$;
Endfor;
$V = \{j \in O \mid s_j < D_j\}$;
$Change = True$;

Main loop:
While *Change* **And** $V \neq \emptyset$ **Do**
$Change := False$;
For all $j \in V$ **And** $k \in K_j$ **Do**
$\rho_{h(k)} := \min_{l \in O \setminus V} \left\{ \frac{c^*_{kl} - c^*_{kj}}{d_k} : s_l - D_l \geq d_k \right\}$;
Endfor;
Order the $\rho_{h(k)}$ in nondecreasing order;
While (list of $\rho_{h(k)} \neq \emptyset$) **And** $(V \neq \emptyset)$ **Do**
Pick first $\rho_{h(k)}$ in the list;
If $s_{h(k)} \geq D_{h(k)} + d_k$ **Then**
$Change := True$;
$K_j := K_j \setminus \{k\}$;
$K_{h(k)} := K_{h(k)} \cup \{k\}$;
$D_j := D_j - d_k$;
$D_{h(k)} := D_{h(k)} + d_k$;
If $s_j \leq D_j$ **Then** $V := V \setminus \{j\}$;
Endif;
Drop $\rho_{h(k)}$ from the list;
Endwhile;
Endwhile;

Even if the GAP (54)–(57) has a feasible solution, there is no guarantee that the reassignment procedure will find one. Therefore, one can try to

improve the solution, i. e. to decrease its infeasibility or its objective function value, by a simple local search heuristic, which reassigns customers from one depot to another or exchanges customers between depots. To measure the infeasibility of a solution, the sum of the violations of the depot capacities is used. The improvement of a solution is then measured as the increase in cost per unit decrease of infeasibility. Such measures of infeasibility and improvement have been used by Hillier (1969) in his heuristic for linear pure integer programming problems. Algorithm 3 describes this simple local search procedure.

Algorithm 3: Local search heuristic for the GAP

Notation:
Ω actual partition of customer set K, $\Omega = \{K_1, \ldots, K_n\}$,
$C(\Omega)$ cost of a solution Ω, $C(\Omega) = \sum_{j \in J} \sum_{k \in K_j} c^*_{kj}$,
$U(\Omega)$ infeasibility of a solution Ω, $U(\Omega) = \sum_{j \in J} \max\Big\{0, \sum_{k \in K_j} d_k - s_j\Big\}$,
$N(\Omega)$ neighborhood of a solution Ω, where
$N(\Omega) = \{\Omega' : \Omega' = \{K'_1, \ldots, K'_n\}\}$,
$K'_j = K_j \quad \forall\, j \in J \setminus \{l, q\}$,
$K'_l = (K_l \setminus \overline{K}_l) \cup \overline{K}_q$,
$K'_q = (K_q \setminus \overline{K}_q) \cup \overline{K}_l$,
$(|\overline{K}_l|, |\overline{K}_q|) \in \{(0,1), (1,0), (1,1)\}$.

Search for a feasible solution:
Stop := *False*;
While $(U(\Omega) > 0)$ **And Not** *Stop* **Do**
  If $\{\Omega' \in N(\Omega) : U(\Omega') = 0\} \neq \emptyset$ **Then**
    $\Omega^* := \arg\min \{C(\Omega') : \Omega' \in N(\Omega),\ U(\Omega') = 0\}$;
    $\Omega := \Omega^*$;
  Else If $\{\Omega' \in N(\Omega) : U(\Omega') < U(\Omega)\} \neq \emptyset$ **Then**
    $\Omega^* := \arg\min \left\{ \dfrac{C(\Omega') - C(\Omega)}{U(\Omega) - U(\Omega')} : \Omega' \in N(\Omega),\ U(\Omega') < U(\Omega) \right\}$;
    $\Omega := \Omega^*$;
  Else
    Stop := *True*;
  Endif;
Endwhile;

Search for improved solutions:
Improve := $U(\Omega) = 0$;
While *Improve* **Do**
  If $\{\Omega' \in N(\Omega) : U(\Omega') = 0,\ C(\Omega') < C(\Omega)\} \neq \emptyset$ **Then**
    $\Omega := \Omega^* \in \{\Omega' \in N(\Omega) : U(\Omega') = 0,\ C(\Omega') < C(\Omega)\}$;
    Improve := *True*;
  Endif

Else
Improve := *False*;
Endwhile;

The local search above can be easily extended by allowing shifts and interchanges of up to two customers between two depots. Osman (1995) uses such a neighborhood in simulated annealing and tabu search approaches for the GAP. Since this extended search is time-consuming, it is only used at the end of the overall algorithm, to further improve the assignment of customers in the best solution found to the TSCFLP.

5.2 Computation of an Initial Set of Dual Cuts

Consider a feasible solution $(\tilde{x}, \tilde{y}, \tilde{z})$ to the TSCFLP. If Dantzig-Wolfe decomposition is applied to the primal subproblem (45), with $y = \tilde{y}$ and $z = \tilde{z}$, the resulting master problem is equivalent to (45) and a relaxation of the complete dual master problem (35)–(39) (see Van Roy (1986)). Furthermore, an efficient set T_D^h of dual cuts for this relaxation can be generated by applying algorithm 1 to the solution $(\tilde{x}, \tilde{y}, \tilde{z})$. By this way, an initial set T_D^h of cuts has been found, which guarantees that the corresponding (relaxed) master problem (35)–(39) is bounded (see Van Roy (1986)). In general, a feasible solution to the TSCFLP is easily obtained by solving the Lagrangean subproblem (30) with $\lambda_i = 0$ for all $i \in I$ and $\mu_j = f_j/s_j$ for all $j \in J$, and applying the reassignment heuristics described in the preceding section.

On the other hand, as mentioned in Sect. 4, an initial set of dual cuts can be computed by applying the subproblem phase of cross decomposition to the TSCFLP without the capacity constraints (3). This procedure can be summarized as follows:

Algorithm 4: Procedure to obtain an initial set of dual cuts

Notation:

t	iteration counter,
z_B^*	best upper bound for problem without constraints (3),
z_D, z_B	best lower and upper bound,
T_P^t, T_D^t	index set of generated Benders' cuts and dual cuts, resp.,
(x^t, y^t, z^t)	actual solution of Lagrangean subproblem (30),
μ^t, λ^t	multipliers of capacity constraints (3) and (6),
ω^t	dual variables of constraints (7),
μ^B, λ^B	best multipliers found so far.

Initialization:
Stop := *False*;
$t := 0$;
$T_P^t = T_D^t := \emptyset$;
$z_B = z_B^* := \infty$;
$z_D := -\infty$;
$\mu^1 = \lambda^1 := 0$;

Main loop:
While Not *Stop* **Do**
  $t := t + 1$;
  Call *DualSubproblem*(λ^t, μ^t);
  If Not *Stop* **Then**
    Generate a dual cut (36) and set $T_D^t := T_D^{t-1} \cup \{t\}$;
    If (x^t, y^t, z^t) does not meet condition (53) **Then**
      $Stop := True$
    Else
      Call *PrimalSubproblem*;
    Endif
  Endif
Endwhile

Procedure *DualSubproblem*
Solve the Lagrangean subproblem (30) with multipliers (λ^t, μ^t) and obtain an optimal solution (x^t, y^t, z^t);
If $v(SD_{(\lambda^t,\mu^t)}) > z_D$ **Then**
  $z_D := v(SD_{(\lambda^t,\mu^t)})$;
  $(\lambda^B, \mu^B) := (\lambda^t, \mu^t)$;
Endif;
Obtain a feasible solution with total costs z_H^t by applying algorithm 2 and 3;
If $z_H^t < z_B$ **Then** $z_B := z_H^t$;
If $z_D = z_B$ **Then** $Stop := True$;

Procedure *PrimalSubproblem*
Solve the primal subproblem (45);
Obtain an optimal dual solution $(\lambda^{t+1}, \omega^{t+1})$ and set $\mu^{t+1} = 0$;
If $v(SP_{(y^t,z^t)}) < z_B^*$ **Then** $z_B^* := v(SP_{(y^t,z^t)})$;
Generate a Benders' cut (47) and set $T_P^t := T_P^{t-1} \cup \{t\}$;
If λ^{t+1} does not meet condition (52) **Then**
  $Stop := True$
Else
  Derive additional dual cuts by applying algorithm 1;

In total, three different methods to get an initial set of dual cuts have been tried:

Method 1: Generate a feasible solution to the TSCFLP by solving (30) with $\lambda_i = 0$ $(\forall\, i \in I)$ and $\mu_j = f_j/s_j$ $(\forall\, j \in J)$, and applying algorithm 2 and 3. Generate dual cuts from this solution by applying algorithm 1.
Method 2: Generate a feasible solution as in method 1. Generate a dual cut from this solution. Then apply algorithm 4.
Method 3: Apply method 1 and algorithm 4.

5.3 Solving the Lagrangean Dual

After an initial set of dual cuts has been obtained, the Lagrangean dual is solved using Dantzig-Wolfe decomposition or "weighed" Dantzig-Wolfe decomposition, resp. In each iteration of the algorithm, a feasible solution to the TSCFLP is constructed from the solution of the Lagrangean relaxation (30) with the help of the reassignment heuristics described in Sect. 5.1. The procedure can be decribed as follows:

Algorithm 5: Lagrangean heuristic for the TSCFLP

Notation:

t	iteration counter,
$\epsilon > 0$	tolerance parameter,
z_D, z_B	best lower and upper bound,
T_P^t, T_D^t	index set of generated Benders' cuts and dual cuts, resp.,
(x^t, y^t, z^t)	actual solution of Lagrangean subproblem (30),
$v(\mathrm{MD}_{T_D^t})$	objective function value of dual master problem (35)–(39) with cut set T_D^t,
$\tilde{\mu}^t$, $\tilde{\lambda}^t$	optimal solution of dual master problem (35)–(39) with cut set T_D^t,
μ^B, λ^B	best multipliers found so far,
α_t	weight for multipliers used in (40), where $\alpha \geq \alpha_{\min}$, ($\alpha_{\min} = 1$, if conventional Dantzig-Wolfe decomposition is used, and $\alpha_{\min} = 0.1$ otherwise),
μ^t, λ^t	multipliers of capacity constraints (3) and (6),
Flag	indicates if dual master problem has to be solved.

Initialization:
Obtain an initial set of cuts by one of the methods described in Sect. 5.2; denote this set by T_D^0; let z_B and z_D be the best upper and lower bound found in this initialization phase, and μ^B, λ^B the multipliers corresponding to z_D; set $t := 0$;

Main Loop:
$Stop := z_D = z_B$;
$Flag := True$;
While Not *Stop* **Do**
  $t := t + 1$;
  If $Flag = True$ **Then**
    Call *DualMaster*;
  Else
    $(\tilde{\lambda}^t, \tilde{\mu}^t) := (\tilde{\lambda}^{t-1}, \tilde{\mu}^{t-1})$;
    $v(\mathrm{MD}_{T_D^{t-1}}) := v(\mathrm{MD}_{T_D^{t-2}})$;
  Endif;
  If $z_D \cdot (1 + \epsilon) \geq v(\mathrm{MD}_{T_D^{t-1}})$ **Then**
    $Stop := True$

Else
$\alpha_t := \max\{\alpha_{\min}, 1/t\}$;
$(\lambda^t, \mu^t) := \alpha_t(\tilde{\lambda}^t, \tilde{\mu}^t) + (1 - \alpha_t)(\lambda^B, \mu^B)$;
Call *DualSubproblem*;
If $(z_D = z_B)$ **Or** $(z_D(1 + \epsilon) = v(\mathrm{MD}_{T_D^{t-1}}))$ **Then**
Stop := *True*
Else If $(x^t, y^t, z^t) \notin \{(x^h, y^h, z^h) : h \in T_D^{t-1}\}$ **Then**
Generate a dual cut as in (36) and set $T_D^t := T_D^{t-1} \cup \{t\}$
Flag := *True*;
Else
Flag := *False*;
Endif;
Endif;
Endwhile;

Procedure *DualMaster*
Solve (35)–(39) with cut set T_D^{t-1} and the additional constraint $\lambda_0 \leq z_B$. Obtain optimal solution $(\tilde{\lambda}^t, \tilde{\mu}^t)$ with objective function value $v(\mathrm{MD}_{T_D^{t-1}})$.

Procedure *DualSubproblem* (See Sect. 5.2)

6 Computational Results

The algorithm has been coded in Sun Pascal and tested on 48 randomly generated test problems of different sizes (see Table 1) and different characteristics, concerning the relative size of fixed depot costs, throughput costs and depot capacities (see Table 2). The procedure was terminated, if a maximum number of 150 dual cuts was reached, or if the gap between the (relaxed) dual master problem and the best lower bound was less than or equal to 1 %.

Table 1. Problem classes

Class	A	B	C	D	E	F
$\|I\|$	10	10	10	20	20	20
$\|J\|$	25	25	50	25	25	50
$\|K\|$	50	100	100	50	100	100

Table 2. Problem characteristics

Problem	1	2	3	4	5	6	7	8
Capacity	S	S	S	S	L	L	L	L
Throughput costs	S	S	L	L	S	S	L	L
Fixed costs	S	L	S	L	S	L	S	L

S = "small", L = "large"

Table 3 shows the results, averaged over all test problems in a class, for the different "cut initialization methods" (see Sect. 5.2), if conventional Dantzig-Wolfe decomposition was applied to solve the Lagrangean dual. The entries in this table and the following tables have the following meaning:

It denotes the number of iterations (number of times, the UFLP (34) has been solved),

t is the computation time in CPU-seconds on a Sun Ultra (166 Mhz),

Δ_1 denotes the percentage deviation between the lower and upper bound computed by the Lagrangean heuristic,

Δ_2 is the percentage gap between the solution value of the last dual master problem and the lower bound computed by the Lagrangean heuristic,

Δ_3 is the percentage deviation of the upper bound computed by the Lagrangean heuristic and a lower bound obtained by the branch and bound method of CPLEX within a time limit of 3 hours, (within this time limit CPLEX was able to find the optimal solution of the 9 test problems 1, 5 and 8 in class A, 7 and 8 in class B, 6 and 8 in class D, 5 and 7 in class E),

Δ_4 is the percentage deviation of the lower bound found by the heuristic and the lower bound found using CPLEX.

As can be seen from Table 3, the best results using "unweighed" Dantzig-Wolfe decomposition have been obtained, if method 3 was used to get an initial set of dual cuts (an average gap of 1.38 % between the upper bound and CPLEX's lower bound and an average gap of 3.21 % between the lower bound and CPLEX's lower bound). The gap between the dual master problem and the lower bound is quite large (3.75 %), which means that for most of the test problems, the Lagrangean dual could not be solved to the desired ϵ-optimality of 1 % within the limit of 150 dual cuts.

Table 4 shows the corresponding results, which have been obtained by using "weighed" Dantzig-Wolfe decomposition to solve the Lagrangean dual. As can be seen from the value of Δ_2 in that table, now nearly for all test problems, except for some of the larger ones, the desired 1 %-optimality of the solution to the Lagrangean dual has been reached. Furthermore, there is a substantial improve in the lower and the upper bounds found, as well. The difference in performance between the cut initialization methods vanishes (although there is a slight advantage for method 3). On the other hand, the computation times increase. That means, the Lagrangean subproblems have been much more difficult to solve than in the case of "unweighed" decomposition. But, this increase in computation time is justified by the better bounds.

To further improve the lower bounds, it has been tried to incorporate some valid inequalities. Valid inequalities for the TSCFLP can be easily derived from *lifted cover inequalities* for the knapsack polytope

$$\operatorname{conv}\left(F_{\mathrm{KP}}^{K}\right) = \operatorname{conv}\left(\left\{z \in \{0,1\}^{|K|} : \sum_{k \in K} d_k z_k \leq s_j\right\}\right).$$

A *minimal cover* $C \subseteq K$ is a minimal subset of K, such that

$$\sum_{k \in C} d_k > s_j \text{ and } \sum_{k \in C \setminus \{l\}} d_k \leq s_j \text{ for any } l \in C\,.$$

Table 3. Results "unweighed" Dantzig-Wolfe decomposition

Class	It	t	Δ_1	Δ_2	Δ_3	Δ_4
Cut initialization method 1						
A	133.25	65.10	3.88	3.46	0.90	2.99
B	124.75	158.99	3.79	3.10	1.01	2.79
C	143.75	123.31	9.01	9.01	5.10	4.10
D	132.25	37.81	7.10	6.72	1.67	5.53
E	130.13	108.70	5.32	4.19	1.26	4.11
F	132.38	111.08	11.16	11.16	3.79	7.65
Average			6.71	6.27	2.29	4.53
Cut initialization method 2						
A	113.38	67.96	3.37	2.82	1.12	2.28
B	82.38	197.98	2.61	1.53	1.03	1.58
C	124.88	166.48	5.02	4.91	2.05	3.02
D	90.25	31.75	6.74	6.32	1.13	5.68
E	86.13	100.62	5.02	3.93	1.23	3.84
F	87.88	99.31	9.15	9.04	3.35	6.02
Average			5.32	4.76	1.65	3.74
Cut initialization method 3						
A	136.00	710.75	2.45	1.66	0.90	1.56
B	91.38	456.65	2.23	1.13	0.92	1.30
C	132.13	354.89	5.05	4.91	2.11	3.01
D	99.75	118.47	4.51	3.59	0.51	4.01
E	104.38	270.44	4.40	3.03	1.10	3.33
F	106.50	219.05	8.60	8.14	2.76	6.02
Average			4.54	3.75	1.38	3.21

From a minimal cover C, one obtains the valid inequality

$$\sum_{k \in C} z_k \leq |C| - 1 \,,$$

which is a *facet* of $\text{conv}(F_{\text{KP}}^C)$, i. e. the knapsack polytope with K substituted by C. By sequential lifting, which here implies the solving of a sequence of knapsack problems, the inequality can be lifted to a facet

$$\sum_{k \in C} z_k + \sum_{k \in K \setminus C} \alpha_k z_k \leq |C| - 1$$

of $\text{conv}(F_{\text{KP}}^K)$ (see e. g. Nemhauser & Wolsey (1988)). More general lifted cover inequalities can be derived by partitioning the cover C into two subsets and sequential up- and down-lifting (see Gu, Nemhauser and Savelsbergh (1995) on this topic and many other details regarding the implementation of cover inequalities). Here, only the simple lifted cover inequality above has been used. Furthermore, as Barcelo et al. (1990) have done in their approach

Table 4. Results "weighed" Dantzig-Wolfe decomposition

Class	It	t	Δ_1	Δ_2	Δ_3	Δ_4
Cut initialization method 1						
A	60.88	192.83	1.89	0.98	0.95	0.94
B	53.50	370.77	2.20	0.93	0.94	1.25
C	101.75	819.03	1.97	0.96	1.27	0.70
D	81.00	247.92	2.34	0.96	0.31	2.03
E	84.63	444.58	2.73	0.96	1.00	1.73
F	119.63	838.95	3.70	2.20	1.20	2.50
Average			2.47	1.17	0.94	1.53
Cut initialization method 2						
A	48.75	134.46	1.95	0.94	1.00	0.95
B	41.13	290.41	2.12	0.94	0.91	1.21
C	83.75	792.56	2.00	0.94	1.31	0.70
D	63.75	207.00	2.34	0.95	0.35	2.00
E	64.00	424.58	2.76	1.00	1.03	1.73
F	86.88	628.55	4.13	2.72	1.33	2.83
Average			2.55	1.25	0.99	1.57
Cut initialization method 3						
A	49.00	183.10	1.90	0.93	0.98	0.93
B	46.13	366.26	2.11	0.97	0.91	1.19
C	83.25	916.50	2.09	0.96	1.40	0.70
D	59.75	218.73	2.35	0.96	0.34	2.02
E	65.38	540.97	2.66	0.96	1.00	1.66
F	97.88	1000.04	3.54	1.84	1.20	2.37
Average			2.44	1.10	0.97	1.48

for the CFLP with single sourcing, these inequalities have only been used to separate integer solutions of the Lagrangean subproblem, and not to cut off a noninteger optimal solution of the primal of the Lagrangean dual problem. The generation of cover inequalities has been started, if the gap between the lower bound and the dual master falls below a value of 5 %; each time this gap improves by 5 %, the process was repeated. Table 5 shows the results obtained by using "weighed" Dantzig-Wolfe decomposition and cut initialization method 1 (in this table #cuts is the average number of generated inequalities, and #active is the average number of these inequalities, which are active in the optimal solution of the last dual master program). Regarding the lower bounds, there is only a very small improve on the average. On the other hand, the incorporation of these inequalities contributed to an improve in the convergence of the decomposition. This improve in convergence is even more significant if these inequalities are used in combination with "unweighed decomposition, as can be seen from Table 6, (the results in Table 6 have been obtained by generating cover inequalities in every iteration).

Table 5. Results "weighed" decomposition with cover inequalities

Cut initialization method 1								
Class	*It*	*t*	Δ_1	Δ_2	Δ_3	Δ_4	#cuts	#active
A	50.88	115.11	1.83	0.93	0.95	0.89	8.00	1.75
B	43.13	378.73	2.20	0.93	0.94	1.26	7.00	0.75
C	74.13	582.59	1.92	0.97	1.34	0.59	15.63	1.50
D	73.38	178.50	2.30	0.96	0.32	1.98	8.00	0.63
E	77.63	422.64	2.70	0.94	1.04	1.67	8.88	1.38
F	119.13	827.97	3.32	1.71	1.07	2.27	18.75	2.50
Average			2.38	1.07	0.94	1.44		

Table 6. Results "unweighed" decomposition with cover inequalities

Cut initialization method 1								
Class	*It*	*t*	Δ_1	Δ_2	Δ_3	Δ_4	#cuts	#active
A	98.88	140.38	1.78	0.93	0.86	0.92	14.13	0.63
B	66.38	471.51	2.12	0.87	0.94	1.17	10.63	0.13
C	141.50	745.73	3.65	3.23	1.74	1.95	34.25	0.50
D	120.88	168.75	3.45	2.36	0.38	3.08	16.00	0.00
E	120.38	358.01	4.64	3.29	1.08	3.60	15.00	0.13
F	132.38	399.78	8.59	8.59	1.88	6.84	33.50	0.00
Average			4.04	3.21	1.15	2.93		

Finally, the Lagrangean heuristic has been compared to a Lagrangean relaxation approach, which relaxes constraints (2) and (7). A similar approach has been used by Sridharan (1993) for the CFLP with single sourcing. The subproblem, which results from this relaxation, can be solved by computing a sequence of knapsack problems. The Lagrangean dual is approximately solved using subgradient optimization; upper bounds can be generated from a solution of the subproblem in a similar manner, as shown for the relaxation of the capacity constraints, by applying reassignment heuristics. Table 7 shows the results obtained with this approach. Of course, the computation times are much lower, compared to the relaxation of the capacity constraints, but the lower bounds and even the upper bounds are much worse.

To summarize, as the computational results show, the Lagrangean relaxation approach to the TSCFLP described here, produces sharp lower and upper bounds. The upper bounds obtained where within a 1% deviation from optimality, the lower bounds were less than 1.5% from the value of an optimal solution. A key to this performance has been the use of the "weighed" Dantzig-Wolfe decomposition scheme proposed by Wentges (1994, 1997). Further improvements should be possible if classes of valid inequalities are used to cut off a noninteger optimal or near optimal solution of the primal of the Lagrangean dual problem. If combined with such cuts, this bounding scheme

Table 7. Results for Lagrangean relaxation of constraints (2) and (7)

Class	t	Δ_1	Δ_3	Δ_4
A	14.20	12.84	4.49	8.81
B	124.65	9.15	2.82	6.54
C	164.09	7.20	2.87	4.44
D	15.05	13.34	2.81	10.86
E	125.11	14.27	2.92	11.68
F	176.87	10.71	2.30	8.57
Average		11.25	3.04	8.48

can be used in a branch and cut algorithm. Furthermore, the approach can be extended to a multicommodity version of the TSCFLP.

Acknowledgement: This research has been supported by the "Swiss Federal Commission for Technology and Innovation (KTI)".

References

Aardal, K. (1994): Capacitated Facility Location: Separation Algorithm and Computational Experience. CentER Discussion Paper 9480, Tilburg. (available via ftp on ftp://ftp.cs.ruu.nl/pub/ruu/cs/techreps).

Aardal, K. / Labbé, M. / Leung, J. / Queyranne, M. (1994): On the Two-Level Uncapacitated Facility Location Problem. CentER Discussion Paper 9486, Tilburg. (available via ftp on ftp://ftp.cs.ruu.nl/pub/ruu/cs/techreps).

Aardal, K. / Pochet, Y. / Wolsey, L. A. (1993): Capacitated Facility Location: Valid Inequalities and Facets. *Mathematics of Operations Research*, 20:552–582.

Balas, E. / Martin, R. (1980): Pivot and Complement: A Heuristic for 0-1 Programming. *Management Science*, 26:86–96.

Barcelo, J. / Casanovas, J. (1984): A Heuristic Lagrangean Algorithm for the Capacitated Plant Location Problem. *European Journal of Operational Research*, 15:212–226.

Barcelo, J. / Fernandez, E. / Jörnsten, K. (1991): Computational Results from a New Lagrangean Relaxation Algorithm for the Capacitated Plant Location Problem. *European Journal of Operational Research*, 52:38–45.

Barcelo, J. / Hallefjord, Å. / Fernandez, E. / Jörnsten, K. (1990): Lagrangean Relaxation and Constraint Generation Procedures for Capacitated Plant Location Problems with Single Sourcing. *Operations Research-Spektrum*, 12:79–88.

Benders, J. F. (1962): Partitioning Procedures for Solving Mixed-Variables Programming Problems. *Numerische Mathematik*, 4:238–252.

Cho, D. C. / Johnson, E. L. / Padberg, M. W. / Rao, M. R. (1983a): On the Uncapacitated Plant Location Problem I: Valid Inequalities and Facets. *Mathematics of Operations Research*, 8:579–589.

Cho, D. C. / Johnson, E. L. / Padberg, M. W. / Rao, M. R. (1983b): On the Uncapacitated Plant Location Problem II: Facets and Lifting Theorems. *Mathematics of Operations Research*, 8:590–612.

Cornuejols, G. / Fisher, M. L. / Nemhauser, G. L. (1977): On the Uncapacitated Location Problem. *Annals of Discrete Mathematics*, 1:163–177.

Cornuejols, G. / Thizy, J.-M. (1982): Some Facets of the Simple Plant Location Polytope. *Mathematical Programming*, 23:50–74.

Francis, R. L. / McGinnis, L. F. / White, J. A. (1992): *Facility Layout and Location: An Analytical Approach.* Prentice-Hall, Englewood Cliffs NJ.

Geoffrion, A. M. / Graves, G. W. (1974): Multicommodity Distribution System Design by Benders Decomposition. *Management Science*, 20:822–844.

Gu, Z. / Nemhauser, G. L. / Savelsbergh, M. W. P. (1995): Lifted Cover Inequalities for 0-1 Linear Programs: Computation. Report 94-09, Logistic Optimization Center, Georgia Institute of Technology. (available on http:// akula.isye.gatech.edu:80/~mwps/).

Guignard, M. (1980): Fractional Vertices, Cuts and Facets of the Simple Plant Location Problem. *Mathematical Programming Study*, 12:150–162.

Guignard, M. / Opaswongkarn, K. (1990): Lagrangean Dual Ascent Algorithms for Computing Bounds in Capacitated Plant Location Problems. *European Journal of Operational Research*, 46:73–83.

Guignard, M. / Rosenwein, M. B. (1989): An Application-Oriented Guide for Designing Lagrangean Dual Ascent Algorithms. *European Journal of Operational Research*, 43:197–205.

Hillier, F. S. (1969): Efficient Heuristic Procedures for Integer Linear Programming. *Operations Research*, 17:600–637.

Hindi, K. S. / Basta, T. (1994): Computationally Efficient Solution of a Multiproduct, Two-Stage Distribution-Location Problem. *The Journal of the Operational Research Society*, 45:1316–1323.

Holmberg, K. (1990): On the Convergence of Cross Decomposition. *Mathematical Programming*, 47:269–296.

Holmberg, K. (1992a): Generalized Cross Decomposition Applied to Nonlinear Integer Programming Problems: Duality Gaps and Convexification in Parts. *Optimization*, 23:341–356.

Holmberg, K. (1992b): Linear Mean Value Cross Decomposition: A Generalization of the Kornai-Liptak Method. *European Journal of Operational Research*, 62:55–73.

Holmberg, K. (1994): Cross Decomposition Applied to Integer programming Problems: Duality Gaps and Convexification in Parts. *Operations Research*, 42:657–668.

Klincewicz, J. G. / Luss, H. (1986): A Lagrangian Relaxation Heuristic for Capacitated Facility Location with Single Source Constraints. *The Journal of the Operational Research Society*, 37:495–500.

Klose, A. (1994): A Branch and Bound Algorithm for an Uncapacitated Facility Location Problem with a Side Constraint. Working paper, Institut für Unternehmensforschung (Operations Research), Hochschule St. Gallen.

Krarup, J. / Pruzan, P. M. (1983): The Simple Plant Location Problem: Survey and Synthesis. *European Journal of Operational Research*, 12:36–81.

Leung, J. M. Y. / Magnanti, T. L. (1989): Valid Inequalities and Facets of the Capacitated Plant Location Problem. *Mathematical Programming*, 44:271–291.

Magnanti, T. L. / Wong, R. T. (1981): Accelerating Benders Decomposition: Algorithmic Enhancement and Model Selection Criteria. *Operations Research*, 29:464–484.

Magnanti, T. L. / Wong, R. T. (1989): Decomposition Methods for Facility Location Problems. In: P. B. Mirchandani and R. L. Francis, editors, *Discrete Location Theory*, Wiley-Interscience Series in Discrete Mathematics and Optimization, pages 209–262. John Wiley & Sons, Chichester New York.

Martello, S. / Toth, P. (1990): *Knapsack Problems – Algorithms and Computer Implementations.* Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Chichester New York.

McDaniel, D. / Devine, M. (1977): A Modified Benders' Partitioning Algorithm for Mixed Integer Programming. *Management Science*, 24:312–319.

Nemhauser, G. L. / Wolsey, L. A. (1988): *Integer and Combinatorial Optimization.* Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Chichester New York.

Osman, I. H. (1995): Heuristics for the Generalized Assignment Problem: Simulated Annealing and Tabu Search Approaches. *Operations Research-Spektrum*, 17:211–225.

Sridharan, R. (1993): A Lagrangian Heuristic for the Capacitated Plant Location Problem with Single Source Constraints. *European Journal of Operational Research*, 66:305–312.

Van Roy, T. J. (1980): *Cross Decomposition for Large-Scale, Mixed Integer Linear Programming with Application to Facility Location on Distribution Networks.* PhD thesis, Katholieke Universiteit Leuven, Leuven.

Van Roy, T. J. (1983): Cross Decomposition for Mixed Integer Programming. *Mathematical Programming*, 25:46–63.

Van Roy, T. J. (1986): A Cross Decomposition Algorithm for Capacitated Facility Location. *Operations Research*, 34:145–163.

Wentges, P. (1994): *Standortprobleme mit Berücksichtigung von Kapazitätsrestriktionen: Modellierung und Lösungsverfahren.* Dissertation Nr. 1620, Hochschule St. Gallen.

Wentges, P. (1996): Accelerating Benders' Decomposition for the Capacitated Facility Location Problem. *Mathematical Methods of Operations Research*, 44:267–290.

Wentges, P. (1997): Weighted Dantzig-Wolfe Decomposition for Linear Mixed-Integer Programming. *International Transactions in Operational Research*, 4:151–162.

Parametric Analysis of Fixed Costs in Uncapacitated Facility Location

Andreas Klose and Paul Stähly

Universität St. Gallen, 9000 St. Gallen, Switzerland

Abstract. Solving facility location problems does not only require to compute optimal or nearby optimal solutions, but also to perform a sensitivity and parametric analysis. It is necessary to provide insight into the behaviour of the total cost in dependence on the number of facilities to locate and into the possible variations of the data, which do not affect the optimality of a solution. Such an information is needed because locational decisions have a long-term planning horizon and the cost and demand data are subject to unforeseeable changes. Furthermore, only the information of the cost curve in the neighborhood of the optimum allows the decision maker to assess the consequences of a deviation from the optimal solution, which may be desirable for certain reasons.

Regarding the *Uncapacitated Facility Location Problem* (UFLP), such a sensitivity and parametric analysis of the fixed costs can be done by means of Lagrangean relaxation. In this paper, we will describe this approach theoretically and demonstrate its use on two real depot location problems.

1 Introduction

Sensitivity analysis in mathematical programming investigates the robustness of an optimal solution against variations in the data. It defines a range of parameter values for which the computed solution remains optimal. A natural extension of sensitivity analysis is parametric analysis. In a parametric analysis, the data of a mathematical programming problem are represented as functions of one or more parameters. Parametric programming then tries to identify optimal solutions for a given range of these parameters.

The need for sensitivity and parametric analysis in mathematical programming arises mainly from the fact of uncertainty in the data, estimation errors, incomplete information, the approximate nature of certain data or the mathematical model used. Furthermore, for many planning problems it is often more important to learn about the effects of data variations, than to know a single mathematical optimal solution.

An important instrument, which facilitates sensitivity and parametric analysis, is duality theory. Regarding linear programming, duality theory is well established and sensitivity results can be easily derived. For a linear program

$$v(\mathrm{LP}) = \min\{cx : Ax \geq b,\ x \geq 0\} \tag{1}$$

it is well known that the value functions

$$\Phi_{\mathrm{LP}}(d) = \min\{cx : Ax \geq d,\ x \geq 0\} \tag{2}$$

and

$$\Psi_{\text{LP}}(g) = \min\{gx \,:\, Ax \geq b,\ x \geq 0\} \tag{3}$$

are piecewise linear, and convex and concave, resp. (see e. g. Minoux (1986)). The break-points of these functions can be determined with the help of sensitivity analysis. Futhermore, an optimal primal solution of the LP (1) yields a subgradient of the function Ψ_{LP} at the point $g = c$ while an optimal dual solution is a subgradient of the function Φ_{LP} at the point $d = b$.[1] In this sense, dual variables act as sensitivity measures in postoptimal analysis.

Regarding an integer program (IP)

$$v(\text{IP}) = \min\left\{cx \,:\, Ax \geq b,\ x \in \mathbb{Z}_+^n\right\}\ , \tag{4}$$

where $\mathbb{Z}_+$ denotes the set of nonnegative integers, the situation is quite more complex. In integer programming no unique dual problem exists and strong dual problems, which can be the basis for sensitivity and parametric analysis, are very difficult to obtain. Some aspects and results of integer programming duality are discussed in Walukiewicz (1981) and further literature on this topic can be found in the paper of Jenkins (1990).

Since the IP (4) can be – at least theoretically – reformulated as the linear program

$$v(\text{IP}) = \min\left\{cx \,:\, x \in \text{conv}\{x \in \mathbb{Z}_+^n \,:\, Ax \geq b\}\right\}\ ,$$

where $\text{conv}(S)$ denotes the convex hull of a set S, the value function

$$\Psi_{\text{IP}}(g) = \min\left\{gx \,:\, Ax \geq b,\ x \in \mathbb{Z}_+^n\right\} \tag{5}$$

of the IP (4), with respect to the coefficients of the objective function, is still a piecewise linear and concave function. Therefore, to evaluate a parametric function like

$$\psi_{\text{IP}}(\theta) = \Psi_{\text{IP}}\big(c(\theta)\big)\ ,\ \text{where } c(\theta) = c + \theta\tilde{c} \text{ and } 0 \leq \theta \leq 1\ , \tag{6}$$

for a given "change vector" $\tilde{c}$, one can use its concavity by iteratively refining piecewise linear and concave lower and upper functions $LB(\theta)$ and $UB(\theta)$ until these functions coincide for every linear segment (see e. g. Jenkins (1990)). Like the value function Φ_{LP} of a linear program, the value function

$$\Phi_{\text{IP}}(d) = \min\left\{cx \,:\, Ax \geq d,\ x \in \mathbb{Z}_+^n\right\} \tag{7}$$

of the IP (4) with respect to the right-hand side is also piecewise linear, nondecreasing and subadditive (i. e. $\Phi_{\text{IP}}(d^1) + \Phi_{\text{IP}}(d^2)$ is an upper bound for

[1] A vector $s \in \mathbb{R}^n$ is a *subgradient* of the convex or concave function $f : \mathbb{R}^n \to \mathbb{R}$ at the point $x = x^*$ if $s(x - x^*) = 0$ is a supporting hyperplane of f at $x = x^*$, i. e. $f(x) \geq f(x^*) + s(x - x^*)$ for all x if f is convex, and $f(x) \leq f(x^*) + s(x - x^*)$ for all x if f is concave, resp. The subdifferential of the function f at $x = x^*$ is the set of all subgradients at this point.

$\Phi_{\text{IP}}(d^1 + d^2)$), but unfortunately not convex. One possibility to evaluate a parametric function like

$$\phi_{\text{IP}}(\alpha) = \Phi_{\text{IP}}\big(b(\alpha)\big) \text{ , where } b(\alpha) = b + \alpha\tilde{b} \text{ and } 0 \leq \alpha \leq 1 \text{ ,} \tag{8}$$

for a pure IP, where the given change vector $\tilde{b}$ is restricted to be either positive or negative, is therefore to determine those points α at which $\phi_{\text{IP}}(\alpha)$ has a discontinuity and to solve the corresponding IP at these points (see Jenkins (1990)). On the other hand, one major concern of integer programming duality is to approximate the value function Φ_{IP} from below (see e. g. Nemhauser & Wolsey (1988) on this topic). If only a sub-vector of d in (7) is allowed to vary, one way do to this, is to compute the convex hull of the resulting value function. This approach is connected to Lagrangean relaxation and Lagrangean duality. Furthermore, the parametric function (6) is also related to Lagrangean relaxation, since it results from adding the constraint $\tilde{c}x = \tilde{c}_0$ to (4) and relaxing this constraint with a multiplier θ.

In the sequel, we describe the use of Lagrangean relaxation for parametric analysis of the *Uncapacitated Facility Location Problem* (UFLP), which is a special structured IP. In Sect. 2 we briefly describe the UFLP and the objectives of the parametric analysis. Section 3 summarizes some theoretical results which form the basis of the algorithm to perform the analysis. The algorithmic details are described in Sect. 4. Since only variations of the fixed costs in the UFLP are considered here, we remark on the parametric analysis of the variable costs in Sect. 5. Finally, we report on two case-studies in Sect. 6.

2 The Uncapacitated Facility Location Problem

The UFLP consists in finding the locations of uncapacitated depots from a set of potential depot sites and the assignment of the customers to the selected depots in such a way as to minimize the total cost. The total cost consists of the fixed costs for operating the depots and the costs for distributing commodities between the depots and the customers. Mathematically, the problem can be formulated as follows

$$v(\text{UFLP}) = \min \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} z_{ij} + \sum_{j=1}^{n} f_j y_j \tag{9}$$

subject to

$$\sum_{j=1}^{n} z_{ij} = 1, \quad \forall\, i \tag{10}$$

$$z_{ij} - y_j \leq 0, \quad \forall\, i, j \tag{11}$$

$$z_{ij},\, y_j \in \{0, 1\}, \quad \forall\, i, j \text{ ,} \tag{12}$$

where i indexes the customers and j the potential depot sites; m is the number of customers and n the number of potential depots; c_{ij} is the cost of meeting all the demand of customer i from depot j and f_j the fixed cost of operating depot j. The variable y_j equals 1 if depot j is open, and 0 otherwise. The variable z_{ij} equals 1 if customer i is assigned to depot j, and 0 otherwise.

The UFLP has a variety of applications which have nothing to do with distribution logistics and depot location (see e. g. Boffey (1989), Cornuejols et al. (1977), Current & Weber (1994), Krarup & Pruzan (1983) and Nauss & Markland (1981)). But, if applied to depot location, it models in general a planning problem with a long-term planning horizon. Decisions on the locations of the depots have a long-standing influence on the profit situation of the firm and are only reversible at high cost. Furthermore, there is uncertainty with respect to the demand development, the fixed depot costs and the variable costs for satisfying the demands. Therefore, sensitivity and parametric analysis is important, if the UFLP is used to support locational decision making. Even though NP-hard (see Krarup & Pruzan (1983)), the UFLP is usually relatively easy to solve, since in most cases the LP-relaxation gives a sharp lower bound (see e. g. Morris (1978)). This behaviour of the UFLP facilitates parametric analysis substantially (on the other hand of course, as Ahn et al. (1988) show, it is easy to construct problem instances with a huge linear programming duality gap).

Since variations of the demand data do only affect the variable costs c_{ij}, parametric analysis in the UFLP refers to the cost coefficients c_{ij} and f_j, i. e. has to evaluate a function like

$$\psi_{\text{ufl}}(\theta) = \min \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{ij} + \theta \tilde{c}_{ij}) z_{ij} + \sum_{j=1}^{n} (f_j + \theta \tilde{f}_j) y_j \ : \ (10)\text{–}(12) \right\} , \quad (13)$$

where $0 \leq \theta \leq 1$ and $\tilde{c}$ and $\tilde{f}$ are given change vectors. As it has been mentioned in Sect. 1, the function (13) is piecewise-linear and concave.

In this paper, we are primarily concerned with variations of the fixed costs, where the changes $\tilde{f}_j$ of the fixed costs are equal. However, the methods described here, can be applied in a similar way in the case of a general change vector $\tilde{f}$. Furthermore, in the case of equal fixed costs, a parametric analysis of the variable costs can be easily derived from the analysis of the fixed costs.

Consider the UFLP with the additional cardinality constraint

$$\sum_{j=1}^{n} y_j = p \, . \quad (14)$$

If this constraint is relaxed in a Lagrangean fashion, one obtains the Lagrangean relaxation

$$v(\text{LR}_\mu) = \min \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} z_{ij} + \sum_{j=1}^{n} (f_j + \mu) y_j \ : \ (10)\text{–}(12) \right\} - p\mu \, . \quad (15)$$

Assuming identical changes, parametric analysis of the fixed costs in the UFLP is equivalent to the evaluation of the function $v(\mathrm{LR}_\mu)$. This can be done with the help of the knowledge of the value function

$$\Phi_{\mathrm{ufl}}(p) = \min\left\{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}z_{ij} + \sum_{j=1}^{n} f_j y_j \;:\; (10)\text{–}(12),\ (14)\right\},\quad p = 1, \ldots, n\,. \tag{16}$$

Strictly speaking, the knowledge of the convex hull $\overline{\Phi}_{\mathrm{ufl}}$ of the cost curve Φ_{ufl} suffices to compute $v(\mathrm{LR}_\mu)$ for all μ. Of course, the cost curve Φ_{ufl} is itself of interest in the context of a depot location problem. Fortunately, as computational experience shows, the cost curve Φ_{ufl} usually coincides with its convex hull for most values of p.

Regarding the relaxation (15) and its Lagrangean dual, some important results (Geoffrion (1974) and Greenberg (1977)) are known, which directly lead to the so-called "tangential approximation" method to compute the convex hull of the function $\overline{\Phi}_{\mathrm{ufl}}$ at a given point p. This method can be easily extended to compute the complete convex hull. In the following section, we describe these results with respect to the general linear integer programming problem and the relaxation of a single constraint.

3 The One-Dimensional Multiplier Problem

Assume that the set $\{x \in \mathbb{Z}_+^n \;:\; Ax \geq b\}$ of feasible solutions of the IP in (4) is given by $F \cap \{x \in \mathbb{R}^n \;:\; hx = d_0\}$ where $F = \{x \in \mathbb{Z}_+^n \;:\; A^*x \geq b^*\}$, i. e.

$$v(\mathrm{IP}) = \min\{cx \;:\; hx = d_0,\ x \in F\}\,. \tag{17}$$

Relaxing the single constraint $hx = d_0$ in a Lagrangean manner yields the Lagrangean relaxation

$$v(\mathrm{LR}_\lambda) = \min\{cx + \lambda(hx - d_0) \;:\; x \in F\}\,. \tag{18}$$

The "one-dimensional multiplier problem" is then to find an optimal multiplier λ^*, i. e. to solve the Lagrangean dual

$$v(\mathrm{LD}) = \max_{\lambda}\, v(\mathrm{LR}_\lambda)\,. \tag{19}$$

There is a close connection between the optimization problem (19) and the convex hull of the value function or "perturbation function"

$$\Phi_{\mathrm{IP}}(d) = \min\{cx \;:\; hx = d,\ x \in F\}\,,\quad d \in D\,, \tag{20}$$

where D is the set of all d for which the integer program in (20) has an optimal solution. As mentioned in Sect. 1, the function (20) is in general not

convex. Now, from the theory of Lagrangean relaxation, it is well known that the Lagrangean dual (19) is dual to the linear program

$$\min \{cx : hx = d_0,\ x \in \operatorname{conv}(F)\} , \tag{21}$$

i. e. $v(\mathrm{LD})$ can be determined by solving the linear program above (see e. g. Nemhauser & Wolsey (1988)). Furthermore, as Geoffrion (1974) has shown, the value function

$$\overline{\Phi}_{\mathrm{IP}}(d) = \min \{cx : hx = d,\ x \in \operatorname{conv}(F)\} , \tag{22}$$

of the linear program (21) is the convex hull of the value function (20). Therefore, the connection between the Lagrangean relaxation (18), the Lagrangean dual (19) and the value functions (20) and (22) can be depicted as in Fig. 1 and Fig. 2. (In Fig. 1, the functions Φ_{IP} and $\overline{\Phi}_{\mathrm{IP}}$ are shown as continuous functions only for illustrative purposes). In Fig. 1, optimal multipliers for the Lagrangean relaxation are indicated as the negative of a subgradient of the function $\overline{\Phi}_{\mathrm{IP}}$. For example, in this figure $-\lambda_3$ is a subgradient of $\overline{\Phi}_{\mathrm{IP}}$ at the point $d = d_0 = d_6$, and $v(\mathrm{LD}) = v(\mathrm{LR}_{\lambda_3}) = \overline{\Phi}_{\mathrm{IP}}(d_0)$ follows. Since the plane $-\lambda_3(d - d_0) = 0$ does not support the value function Φ_{IP} at the point $d = d_0$, a duality gap exists, i. e. $v(\mathrm{IP}) = \Phi_{\mathrm{IP}}(d_0) > v(\mathrm{LD}) = \overline{\Phi}_{\mathrm{IP}}(d_0)$.

The subdifferential of the function $v(\mathrm{LR}_\lambda)$ at the point $\lambda = \overline{\lambda}$ is given by the set

$$\partial v(\mathrm{LR}_{\overline{\lambda}}) = \left\{d - d_0 : v(\mathrm{LR}_{\overline{\lambda}}) = \Phi_{\mathrm{IP}}(d) + \overline{\lambda}(d - d_0)\right\} .$$

It can be shown that $\partial v(\mathrm{LR}_{\overline{\lambda}})$ is a convex set whose vertices are given by $\{d_k - d_0 : d_k = hx^k,\ k \in K'\}$ where $\{x^k : k \in K\}$ is the set of all extreme points of $\operatorname{conv}(F)$ and $v(\mathrm{LR}_{\overline{\lambda}}) = \Phi_{\mathrm{IP}}(d_k) + \overline{\lambda}(d_k - d_0)$ for all $k \in K' \subset K$ (see e. g. Nemhauser & Wolsey (1988)). Figure 2 shows the subgradients of the Lagrangean function $v(\mathrm{LR}_\lambda)$, which corresponds to the value function Φ_{IP} in Fig. 1. For example, subgradients of $v(\mathrm{LR}_\lambda)$ at the point $\lambda = \lambda_3$ are given by $\alpha d_5 + (1 - \alpha)d_7 - d_0$ with $0 \leq \alpha \leq 1$. Therefore, optimal solutions x^* of the Lagrangean subproblem (18) with $\lambda = \lambda_3$ give $hx^* \in \{d_5, d_7\}$, while $-\lambda_3$ is a subgradient of $\overline{\Phi}_{\mathrm{IP}}$ at all points $d = \alpha d_5 + (1 - \alpha)d_7 - d_0$.

In the sequel, the connection between optimal Lagrangean multipliers and global subgradients of the value function Φ_{IP} and subgradients of its convex hull $\overline{\Phi}_{\mathrm{IP}}$, respectively, is considered in more detail. To this end, it is useful to rewrite the Lagrangean relaxation (18) and the Lagrangean dual (19) in terms of the perturbation function. Furthermore, we assume for simplicity that $\operatorname{conv}(F)$ is nonempty and bounded. The Lagrangean relaxation (18) and the Lagrangean dual (19) can then be rewritten in terms of the function Φ_{IP} as follows

$$\begin{aligned} v(\mathrm{LR}_\lambda) &= \min \{cx + \lambda(hx - d_0) : x \in \operatorname{conv}(F)\} \\ &= \min \left\{cx^k + \lambda(hx^k - d_0) : k \in K\right\} \\ &= \min \{\Phi_{\mathrm{IP}}(d_k) + \lambda(d_k - d_0) : k \in K\} \\ &= \min \{\Phi_{\mathrm{IP}}(d) + \lambda(d - d_0) : d \in D\} , \end{aligned} \tag{23}$$

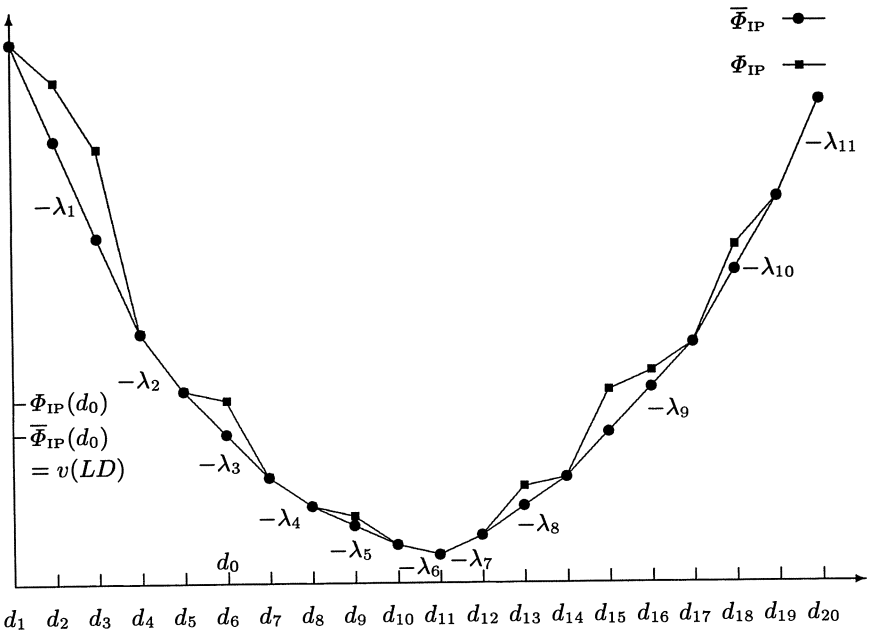


Fig. 1. Perturbation function, its convex hull and Lagrangean multipliers

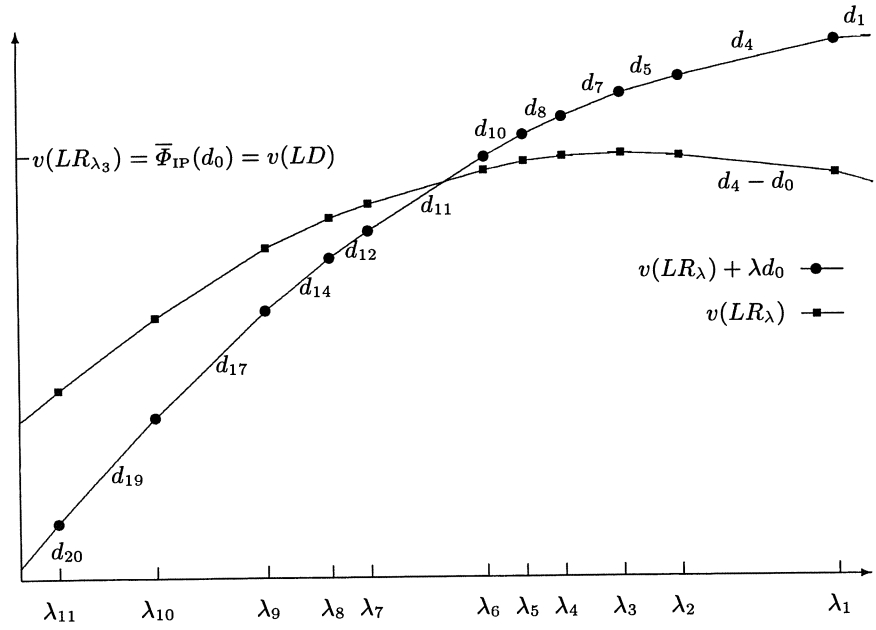


Fig. 2. Lagrangean relaxation and the convex hull of the perturbation function

and

$$\begin{aligned} v(\mathrm{LD}) &= \max_{\lambda} \min \{\Phi_{\mathrm{IP}}(d) + \lambda(d - d_0) \ : \ d \in D\} \\ &= \max_{\lambda} \min \{\Phi_{\mathrm{IP}}(d_k) + \lambda(d_k - d_0) \ : \ k \in K\} \\ &= \max_{\lambda_0,\, \lambda} \{\lambda_0 \ : \ \lambda_0 \leq \Phi_{\mathrm{IP}}(d_k) + \lambda(d_k - d_0) \ \ \forall\, k \in K\} \ . \end{aligned} \tag{24}$$

The dual of the linear program (24) is given by

$$v(\mathrm{LD}) = \min \sum_{k \in K} \mu_k \Phi_{\mathrm{IP}}(d_k) \tag{25}$$

$$\sum_{k \in K} \mu_k = 1 \tag{26}$$

$$\sum_{k \in K} \mu_k d_k = d_0 \tag{27}$$

$$\mu_k \geq 0 \quad \forall\, k \in K \, . \tag{28}$$

Optimality conditions for a feasible solution (λ_0, λ) of (24) and μ of (25)–(28) are then given by the complementary slackness conditions

$$\mu_k \left(\Phi_{\mathrm{IP}}(d_k) + \lambda(d_0 - d_k) - \lambda_0\right) = 0 \quad \forall\, k \in K \, . \tag{29}$$

Given this formulation of the Lagrangean dual and the complementary slackness conditions (29), it is straightforward to derive some conditions for the optimality of a multiplier λ^*.

Proposition 1. (see Geoffrion 1974, Theorem 3) *There exists no duality gap, i. e.* $v(\mathrm{LD}) = \Phi_{\mathrm{IP}}(d_0)$, *if and only if there exists a global subgradient* s *of the function* Φ_{IP} *at the point* $d = d_0$, *i. e. if and only if*

$$\Phi_{\mathrm{IP}}(d) \geq \Phi_{\mathrm{IP}}(d_0) + s(d - d_0) \quad \forall\, d \in D \, .$$

Proof. If $v(\mathrm{IP}) \equiv \Phi_{\mathrm{IP}}(d_0) = v(\mathrm{LD})$ then there exists an optimal multiplier λ^* such that

$$\begin{aligned} \Phi_{\mathrm{IP}}(d_0) = v(\mathrm{LR}_{\lambda^*}) &\equiv \min_{k \in K} \{\Phi_{\mathrm{IP}}(d_k) + \lambda^*(d_k - d_0)\} \\ &= \min_{d \in D} \{\Phi_{\mathrm{IP}}(d) + \lambda^*(d - d_0)\} \\ &\leq \Phi_{\mathrm{IP}}(d) + \lambda^*(d - d_0) \quad \forall\, d \in D \, , \end{aligned}$$

i. e. $-\lambda^*$ is a global subgradient of Φ_{IP} at $d = d_0$. On the other hand, if s^* is a global subgradient of Φ_{IP} at $d = d_0$, we have

$$\begin{aligned} &\Phi_{\mathrm{IP}}(d) \geq \Phi_{\mathrm{IP}}(d_0) + s^*(d - d_0) \quad \forall\, d \in D \\ \Leftrightarrow\ &\Phi_{\mathrm{IP}}(d_0) \leq \min_{d \in D} \{\Phi_{\mathrm{IP}}(d) - s^*(d - d_0)\} = v(\mathrm{LR}_{-s^*}) \, . \end{aligned}$$

Since $v(\mathrm{LR}_\lambda) \leq \Phi_{\mathrm{IP}}(d_0) \ \forall\ \lambda$ this implies $\Phi_{\mathrm{IP}}(d_0) = v(\mathrm{LR}_{-s^*}) = v(\mathrm{LD})$ and $-s^*$ is an optimal multiplier. □

From Proposition 1, one directly obtains the following corollary, which gives a range of optimal multipliers in the case of a nonexisting duality gap.

Corollary 2. *If there is no duality gap, then λ^* is an optimal multiplier if and only if*

$$\max_{d>d_0} \frac{\Phi_{\mathrm{IP}}(d_0) - \Phi(d)}{d - d_0} \leq \lambda^* \leq \min_{d<d_0} \frac{\Phi_{\mathrm{IP}}(d) - \Phi(d_0)}{d_0 - d} .$$

In the case of a duality gap, the following proposition gives a necessary and sufficient condition for the optimality of an multiplier λ.

Proposition 3. *If there is a duality gap, i.e. $v(\mathrm{LD}) < v(\mathrm{IP})$, then λ^* is an optimal multiplier if and only if there exists $d_l < d_0$ and $d_r > d_0$ such that*

$$\Phi_{\mathrm{IP}}(d) \geq \Phi_{\mathrm{IP}}(d_l) - \lambda^*(d - d_l) = \Phi_{\mathrm{IP}}(d_r) - \lambda^*(d - d_r) \quad \forall\, d \in D ,$$

i. e. if and only if $-\lambda^*$ defines a global subgradient of the perturbation function Φ_{IP} at the two points $d_l < d_0$ and $d_r > d_0$. Furthermore, this condition is sufficient if $\Phi_{\mathrm{IP}}(d_0) = v(\mathrm{LD})$.

Proof. Assume λ^* is an optimal multiplier, i.e. an optimal solution of the linear program (24). Then we have

$$\begin{aligned} v(\mathrm{LD}) &= \min_{d \in D} \{\Phi_{\mathrm{IP}}(d) + \lambda^*(d - d_0)\} \\ &= \Phi_{\mathrm{IP}}(d_k) + \lambda^*(d_k - d_0) \quad \forall\, k \in K' \subset K \end{aligned}$$

for some nonempty subset K' of K. From the slackness condition (29) it follows for the corresponding solution of the dual (25)–(28)

$$\mu_k = 0 \ \ \forall\, k \in K \setminus K' \quad \text{and} \quad \sum_{k \in K'} \mu_k = 1 . \tag{30}$$

Since $v(\mathrm{LD}) < \Phi_{\mathrm{IP}}(d_0)$ by assumption, we have

$$\lambda^* \neq 0 \text{ and } d_0 \notin \{d_k : k \in K'\} .$$

Therefore,

$$\sum_{k \in K} \mu_k(d_k - d_0) = \sum_{k \in K'} \mu_k(d_k - d_0) = 0$$

implies together with (28) that there exists $d_l < d_0$ and $d_r > d_0$ $(l, r \in K')$ such that

$$\begin{aligned} \Phi_{\mathrm{IP}}(d_l) + \lambda^*(d_l - d_0) &= \Phi_{\mathrm{IP}}(d_r) + \lambda^*(d_r - d_0) \\ &\leq \Phi_{\mathrm{IP}}(d) + \lambda^*(d - d_0) \quad \forall\, d \in D \\ \Leftrightarrow \qquad \Phi_{\mathrm{IP}}(d_l) + \lambda^* d_l &= \Phi_{\mathrm{IP}}(d_r) + \lambda^* d_r \\ &\leq \Phi_{\mathrm{IP}}(d) + \lambda^* d \quad \forall\, d \in D \\ \Leftrightarrow \quad \Phi_{\mathrm{IP}}(d_l) - \lambda^*(d - d_l) &= \Phi_{\mathrm{IP}}(d_r) - \lambda^*(d - d_r) \\ &\leq \Phi_{\mathrm{IP}}(d) \quad \forall\, d \in D . \end{aligned}$$

To prove sufficiency, assume $-\lambda^*$ is a global subgradient of Φ_{IP} at the two points $d = d_l < d_0$ and $d = d_r > d_0$. Then we have

$$\begin{aligned}
&\Phi_{\text{IP}}(d_l) - \lambda^*(d - d_l) = \Phi_{\text{IP}}(d_r) - \lambda^*(d - d_r) \leq \Phi_{\text{IP}}(d) \quad \forall\, d \in D \\
\Rightarrow\ &\Phi_{\text{IP}}(d_l) + \lambda^*(d_l - d_0) = \Phi_{\text{IP}}(d_r) + \lambda^*(d_r - d_0) \\
&\qquad\qquad \leq \Phi_{\text{IP}}(d) + \lambda^*(d - d_0) \quad \forall\, d \in D,
\end{aligned}$$

and the solution (λ_0^*, λ^*) where

$$\lambda_0^* = \Phi_{\text{IP}}(d_l) + \lambda^*(d_l - d_0) = \Phi_{\text{IP}}(d_r) + \lambda^*(d_r - d_0)$$

is feasible for (24). The following dual solution

$$\mu_l^* = \frac{d_r - d_0}{d_r - d_l}\,,\ \mu_r^* = \frac{d_0 - d_l}{d_r - d_l} \text{ and } \mu_k^* = 0\ \forall\, k \in K \setminus \{l, r\}$$

is feasible for (25)–(28). The optimality of the multiplier λ^* follows then from

$$\begin{aligned}
\sum_{k \in K} \mu_k^* \Phi_{\text{IP}}(d_k) &= \frac{1}{d_r - d_l}\Big((d_r - d_0)\Phi_{\text{IP}}(d_l) - (d_l - d_0)\Phi_{\text{IP}}(d_r)\Big) \\
&= \frac{(d_r - d_0)\Big(\Phi_{\text{IP}}(d_r) + \lambda^*(d_r - d_l)\Big) - (d_l - d_0)\Phi_{\text{IP}}(d_r)}{d_r - d_l} \\
&= \Phi_{\text{IP}}(d_r) + \lambda^*(d_r - d_0) = \lambda_0^*\,,
\end{aligned}$$

which completes the proof. □

Fig. 3 illustrates the determination of the optimal multiplier λ^* in Proposition 3. Furthermore, it is straightforward to show, that the contents of Proposition 3 can also be stated as follows (for a complete proof see Klose (1993)).

Corollary 4. *Assume* $\{k \in K : d_k < d_0\} \neq \emptyset$. *Define*

$$\lambda(d_l, d_r) = \frac{\Phi_{\text{IP}}(d_l) - \Phi_{\text{IP}}(d_r)}{d_r - d_l}\,.$$

The multiplier $\lambda(d_l, d_r)$ *is an optimal multiplier if*

$$\begin{aligned}
&\Phi_{\text{IP}}(d_l) + \lambda(d_l, d_r)(d_l - d_0) \\
&= \min\left\{\Phi_{\text{IP}}(d_i) + \lambda(d_i, d_j)(d_i - d_0) \,:\, d_i < d_0 < d_j\right\}.
\end{aligned}$$

If $v(\text{LD}) < \Phi_{\text{IP}}(d_0)$ *the condition above is also necessary. Furthermore, the multiplier* $\lambda(d_l, d_r)$ *is unique.*

With respect to the UFLP, the Lagrangean relaxation (15) and the cost curve Φ_{ufl} in (16), the results concerning the solution of the one-dimensional multiplier problem can be summarized as follows.

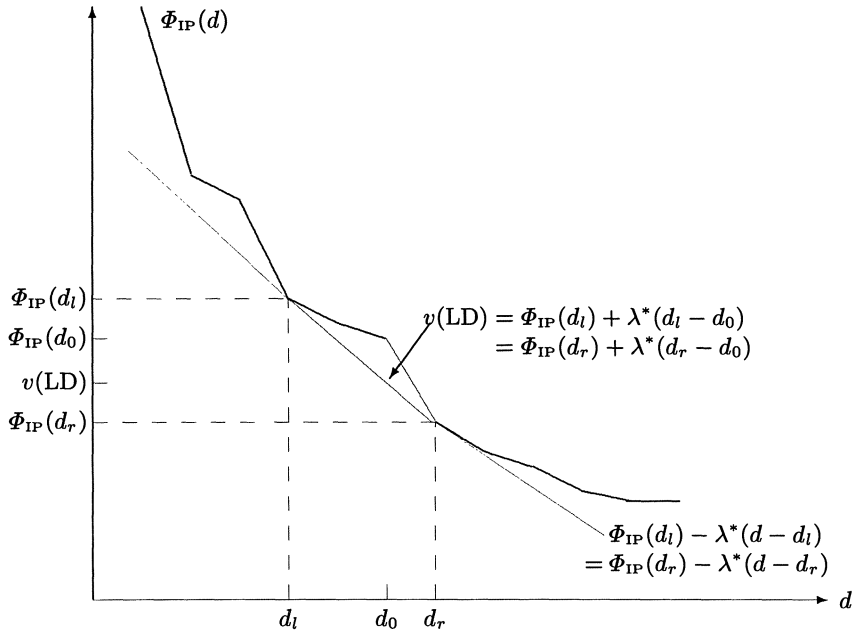


Fig. 3. Solution of the one-dimensional multiplier problem

1. There exists no duality gap, i.e. $\max_\mu v(\text{LR}_\mu) = \Phi_{\text{ufl}}(p)$, if and only if there exists a global subgradient s of the cost curve Φ_{ufl} in (16) at the point $k = p$ (see Proposition 1).
2. If there is no duality gap at the point p, optimal multipliers μ^* lie in the range $\mu_{\min} \leq \mu^* \leq \mu_{\max}$, where

$$\mu_{\min} = \max_{k>p} \left\{ \frac{\Phi_{\text{ufl}}(p) - \Phi_{\text{ufl}}(k)}{k - p} \right\} \text{ and } \mu_{\max} = \min_{k<p} \left\{ \frac{\Phi_{\text{ufl}}(k) - \Phi_{\text{ufl}}(p)}{p - k} \right\}$$

   (see Corollary 2).
3. A multiplier μ^* is optimal if $-\mu^*$ is a global subgradient of the cost curve Φ_{ufl} in (16) at the two points $r < p$ and $R > p$. If there is a duality gap at the point p, this condition is also necessary (Proposition 3).
4. If

$$\Phi_{\text{ufl}}(r^*) + \mu(r^*, R^*)(r^* - p) = \min \left\{ \Phi_{\text{ufl}}(r) + \mu(r, R)(r - p) \,:\, r < p < R \right\},$$

   where

$$\mu(r, R) = \frac{\Phi_{\text{ufl}}(r) - \Phi_{\text{ufl}}(R)}{R - r} \text{ for } r < p < R,$$

   then $\mu(r^*, R^*)$ is an optimal multiplier (Corollary 4).

4 The Tangential Approximation Algorithm

The results of the preceding section lead to the so-called "tangential approximation" method to compute the convex hull of the perturbation function at a given point. We describe this method with respect to the UFLP and the Lagrangean relaxation (15) of the cardinality constraint (14). The algorithm can be applied in a similar way to any other one-dimensional multiplier problem. The method has been introduced by Greenberg (1977). He shows that tangential approximation is equivalent to the application of Dantzig-Wolfe-Decomposition to the linear program (24) and thereby proves the sufficiency of the condition of Proposition 3. The algorithm has also been used by Mirchandani et al. (1985) to solve the Lagrangean dual of the relaxation (15) in the context of a p-median problem.

By Proposition 3, the search for optimal multipliers of the Lagrangean relaxation (15) can be restricted to multipliers

$$\mu(r, R) = \frac{\Phi_{\text{ufl}}(r) - \Phi_{\text{ufl}}(R)}{R - r}$$

where $r = 1, \ldots, p-1$ and $R = p+1, \ldots, n$. If k^* depots are open in an optimal solution of the Lagrangean relaxation (15) with the multiplier μ set to $\mu = \mu(r, R)$, we have

$$\Phi_{\text{ufl}}(k^*) + \mu(r, R)(k^* - p) \leq \Phi_{\text{ufl}}(k) + \mu(r, R)(k - p) \quad \text{for } k = 1, \ldots, n\,,$$

and $-\mu(r, R)$ is a global subgradient of Φ_{ufl} at the point k^*. Therefore, $\mu(r, R)$ is an optimal multiplier if $k^* = p$ by Proposition 1 or if $k^* \in \{r, R\}$ by Proposition 3 and the definition of $\mu(r, R)$. This observation can be exploited as follows to solve the one-dimensional multiplier problem, i. e. to compute the convex hull $\overline{\Phi}_{\text{ufl}}$ of the cost curve Φ_{ufl} at a given point p.

Tangential Approximation Algorithm

Step 1: Set $t = 0$ and $\mu_0 = 0$. Solve the Lagrangean relaxation $[\text{LR}_{\mu_0}]$ in (15). Denote by k_0 the number of depots open in this solution. If $k_0 = p$, then stop. Otherwise, set $(r_0, R_0) = (1, k_0)$ if $k_0 > p$, and $(r_0, R_0) = (k_0, n)$ if $k_0 < p$.

Step 2: Set $\mu_{t+1} = (\Phi_{\text{ufl}}(r_t) - \Phi_{\text{ufl}}(R_t))\,/(R_t - r_t)$, $t := t + 1$ and go to the next step.

Step 3: Solve the Lagrangean relaxation $[\text{LR}_{\mu_t}]$ and let k_t be the number of depots open in an optimal solution. If $k_t \in \{r_{t-1}, p, R_{t-1}\}$ then μ_t is an optimal multiplier. Otherwise set $(r_t, R_t) = (k_t, R_{t-1})$ if $r_{t-1} < k_t < p$, and $(r_t, R_t) = (r_{t-1}, k_t)$ if $p < k_t < R_{t-1}$. Go back to step 2.

Figure 4 illustrates the mechanics of the method. In every iteration of the procedure, one has to solve an UFLP. This can be done by the dual ascent

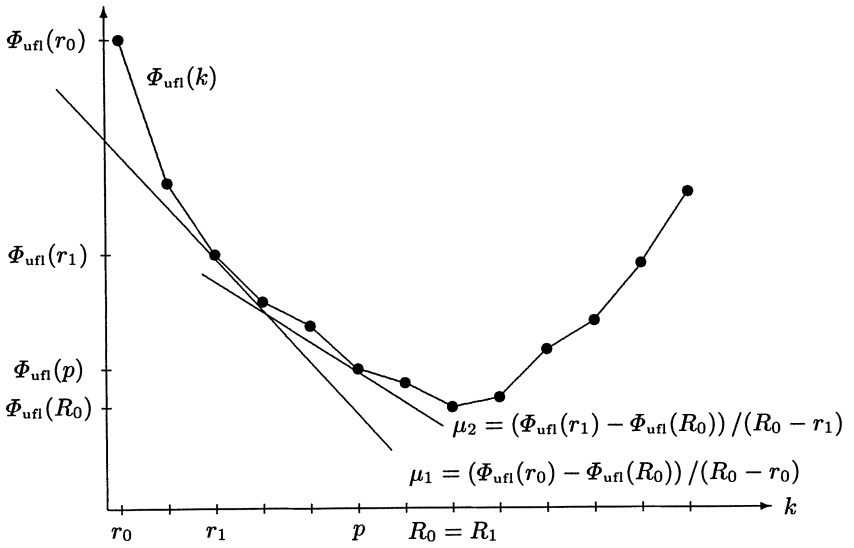


Fig. 4. Tangential approximation algorithm

approach of Erlenkotter (1978) or by improved versions of this algorithm for larger problem instances, like the algorithm of Körkel (1989). Instead of a dual ascent method, we use a branch and bound algorithm which is based on subgradient optimization and the relaxation of the assignment constraints (10) (see Klose (1993, 1994)). The procedure is a modified version of the branch and bound algorithms of Christofides & Beasley (1982) and Hanjoul & Peeters (1985) for the p-median problem. Computational experience has shown that this approach is competitive to dual ascent methods for larger problem instances (see Klose (1995)).

The tangential approximation method can be extended to compute the whole convex hull $\overline{\Phi}_{\text{ufl}}$ of the cost curve Φ_{ufl} and to determine ranges of optimal multipliers for every p. To this end, one determines functions LB and UB which approximate $\overline{\Phi}_{\text{ufl}}$ from below and above. By solving the Lagrangean relaxation (15), these functions are iteratively refined until they coincide for every point p. The following algorithm describes this method.

Computation of the convex hull of Φ_{ufl}

Step 1: Set $t = 0$ and $\mu_t = 0$. Solve the Lagrangean relaxation $[\text{LR}_{\mu_t}]$ in (15) and denote by k_0 the number of depots open in an optimal solution. Set

$$\Phi_{\text{ufl}}(k_0) := v(\text{LR}_0)\,,$$

$$\Phi_{\text{ufl}}(1) := LB_t(1) := \min_{j=1,\ldots,n} \left\{ \sum_{i=1}^{m} c_{ij} + f_j \right\},$$

$$\Phi_{\text{ufl}}(n) := LB_t(n) := \sum_{i=1}^{m} \min_{j=1,\ldots,n} c_{ij} + \sum_{j=1}^{n} f_j \,,$$

$$UB_t(k) := \Phi_{\text{ufl}}(1) - \frac{\Phi_{\text{ufl}}(1) - \Phi_{\text{ufl}}(k_0)}{k_0 - 1}(k-1) \text{ for } k = 1, \ldots, k_0 \,,$$

$$UB_t(k) := \Phi_{\text{ufl}}(k_0) - \frac{\Phi_{\text{ufl}}(k_0) - \Phi_{\text{ufl}}(n)}{n - k_0}(k - k_0) \text{ for } k = k_0, \ldots, n$$

and $LB_t(k) = \Phi_{\text{ufl}}(k_0)$ for $k = 2, \ldots, n-1$. Go to step 2.

Step 2: If $LB_t(k) = UB_t(k)$ for $k = 1, \ldots, n$, set

$$\begin{aligned}
\overline{\Phi}_{\text{ufl}}(j) &:= LB_t(j) \quad \text{for } j = 1, \ldots, n \,, \\
\overline{\mu}_1 &:= \infty \,, \\
\overline{\mu}_j &:= \overline{\Phi}(j-1) - \overline{\Phi}(j) \quad \text{for } j = 2, \ldots, n \,, \\
\underline{\mu}_j &:= \overline{\Phi}(j) - \overline{\Phi}(j+1) \quad \text{for } j = 1, \ldots, n-1 \,, \\
\underline{\mu}_n &:= -\infty
\end{aligned}$$

and terminate. Otherwise go to step 3.

Step 3: Select $p \in \{k : LB_t(k) < UB_t(k)\}$. Set $t := t+1$ and

$$\begin{aligned}
r_t &:= \max_{k<p}\{k : LB_{t-1}(k) = UB_{t-1}(k)\} \\
R_t &:= \min_{k>p}\{k : LB_{t-1}(k) = UB_{t-1}(k)\} \\
\mu_t &:= \frac{\Phi(r_t) - \Phi(R_t)}{R_t - r_t} \,.
\end{aligned}$$

Solve the Lagrangean relaxation $[\text{LR}_{\mu_t}]$ and denote by k_t the number of depots open in an optimal solution. Set

$$\Phi_{\text{ufl}}(k_t) := v(\text{LR}_{\mu_t}) + p\mu_t \,,$$

$$UB_t(k) := \Phi_{\text{ufl}}(k_t) - \frac{\Phi_{\text{ufl}}(r_t) - \Phi_{\text{ufl}}(k_t)}{k_t - r_t}(k - k_t) \text{ for } k = r_t + 1, \ldots, k_t \,,$$

$$UB_t(k) := \Phi_{\text{ufl}}(k_t) - \frac{\Phi_{\text{ufl}}(k_t) - \Phi_{\text{ufl}}(R_t)}{R_t - k_t}(k - k_t) \text{ for } k = k_t, \ldots, R_t - 1$$

and

$$LB_t(k) := \max\{LB_{t-1}(k),\ \Phi_{\text{ufl}}(k_t) - \mu_t(k - k_t)\} \,.$$

for $k = r_t + 1, \ldots, R_t - 1$. Go back to step 2.

Figure 5 illustrates the first three iterations of the algorithm above. After its execution, the convex hull $\overline{\Phi}_{\text{ufl}}$ of the cost curve Φ_{ufl} in (16) is determined together with ranges $\underline{\mu}_p \le \mu_p^* \le \overline{\mu}_p$ for the optimal multiplier μ_p^* of the

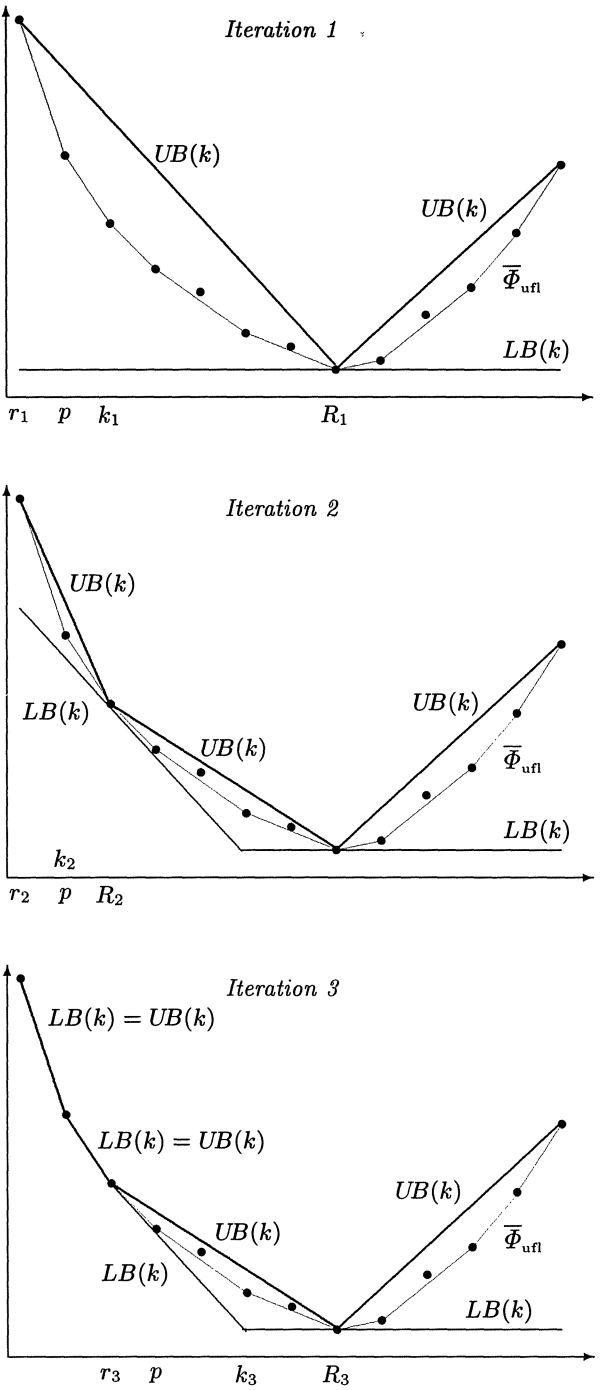


Fig. 5. Computation of the convex hull of the cost curve

Lagrangean relaxation (15) for $p = 1, .., n$. The determination of the function $\overline{\Phi}_{\text{ufl}}$ requires the computation of $2l - 1$ UFLPs where l is the number of linear segments of $\overline{\Phi}_{\text{ufl}}$. Furthermore, all points k of the cost curve Φ_{ufl} where

$$\Phi_{\text{ufl}}(k) = \overline{\Phi}_{\text{ufl}}(k) > \overline{\Phi}_{\text{ufl}}(r) - \frac{\overline{\Phi}_{\text{ufl}}(r) - \overline{\Phi}_{\text{ufl}}(R)}{R - r}(k - r) \text{ for } r < k < R$$

have been determined by the method above.

The knowledge of the function $\overline{\Phi}_{\text{ufl}}$ now allows the following analysis.

1. If k_0 depots are open in an optimal solution of the UFLP (9)–(12) this solution remains optimal as long as all fixed costs f_j do not decrease by more than $-\underline{\mu}_{k_0}$ or increase by more than $\overline{\mu}_{k_0}$.
2. If the fixed cost f_j change to $f_j + \mu$ where $\underline{\mu}_p \leq \mu \leq \overline{\mu}_p$, optimal solutions of the corresponding UFLP are all solutions corresponding to those points k of the cost curve Φ_{ufl} for which

$$\Phi_{\text{ufl}}(k) = \overline{\Phi}_{\text{ufl}}(p) - \mu(k - p)$$

holds.

3. In the case of no changes in the variable costs ($\tilde{c}_{ij} = 0$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$) and equal changes in the fixed costs ($\tilde{f}_j = \tilde{f}$ for $j = 1, \ldots, n$), the parametric function ψ_{ufl} in (13) is given by

$$\psi_{\text{ufl}}(\theta) = v(\text{LR}_\mu) + \mu p = \overline{\Phi}_{\text{ufl}}(p) + \mu p$$

for $\mu = \theta \cdot \tilde{f} \in [\underline{\mu}_p, \overline{\mu}_p]$ and $p = 1, \ldots, n$.

5 Parametric Analysis of Variable Costs

A parametric analysis of the variable costs c_{ij} in (9) is relatively easy if the fixed costs f_j are equal. In this case, a change of the fixed costs from $f_j = f > 0$ for all j to $f_j = f + \mu$ for all j with $\mu > -f$ is equivalent to a change of all variable costs from c_{ij} to $\delta \cdot c_{ij}$ where $\delta = f/(f + \mu)$. If μ^* is an optimal multiplier for the Lagrangean relaxation (15), i. e. $\underline{\mu}_p \leq \mu^* \leq \overline{\mu}_p$, then

$$\begin{aligned}\overline{\Phi}_{\text{ufl}}(p) + \mu^* p &= \min\left\{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} z_{ij} + \sum_{j=1}^{n}(f + \mu^*) y_j \;:\; (10)\text{–}(12)\right\} \\ &= \frac{1}{\delta}\min\left\{\sum_{i=1}^{m}\sum_{j=1}^{n} \delta c_{ij} z_{ij} + \sum_{j=1}^{n} f y_j \;:\; (10)\text{–}(12)\right\}\end{aligned}$$

holds for $\delta = f/(f + \mu*)$ by the results of the previous section. Therefore, the parametric function

$$\psi^*_{\text{ufl}}(\delta) = \min\left\{\sum_{i=1}^{m}\sum_{j=1}^{n} \delta c_{ij} z_{ij} + \sum_{j=1}^{n} f y_j \;:\; (10)\text{–}(12)\right\} \quad (\delta > 0)$$

is given by

$$\psi^*_{\text{ufl}}(\delta) = \delta\left(\overline{\Phi}(p) + p(1/\delta - 1)f\right) \text{ for } \underline{\delta}_p \leq \delta \leq \overline{\delta}_p \text{ and } p = 1, \ldots, n\ ,$$

where $\underline{\delta}_p = f/(f + \overline{\mu}_p)$ and $\overline{\delta}_p = f/(f + \underline{\mu}_p)$.

Theoretically, a general parametric analysis of the variable and fixed costs, defined by the parametric program (13), can be carried out by similar means used for the analysis of the fixed costs in the preceding section. To this end, one can consider the family of integer programs

$$\min\left\{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}z_{ij} + \sum_{j=1}^{n} f_j y_j \ :\ (10)\text{–}(12) \text{ and } \sum_{i=1}^{m}\sum_{j=1}^{n} \tilde{c}_{ij}z_{ij} + \sum_{j=1}^{n} \tilde{f}_j y_j = \gamma\right\}$$

and the Lagrangean relaxation of the additional constraint with multiplier θ, i. e.

$$\min\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}(c_{ij} + \theta\tilde{c}_{ij})z_{ij} + \sum_{j=1}^{n}(f_j + \theta\tilde{f}_j)y_j \ :\ (10)\text{–}(12)\right\} - \theta\cdot\gamma\ .$$

However, the computation of the convex hull of the corresponding value or perturbation function is quite more difficult than the computation of the convex hull $\overline{\Phi}_{\text{ufl}}$ of the cost curve Φ_{ufl} in (16), since it can be expected to consist of far more linear segments. Another approach to parametric analysis of the variable costs in the UFLP has been proposed by Tcha et al. (1995). They consider the parametric cost problem (13) where the change vector $\tilde{f}$ of the fixed costs is set to zero and the change matrix $\tilde{c}$ of the variable costs is set to $\tilde{c}_{ij} = c_{ij}\alpha_i$ with $1 + \alpha_i > 0$ for all i. Such cost changes are caused by changes in the demands if transportation costs are linear functions of the amount transported. To evaluate the resulting function $\psi_{\text{ufl}}(\theta)$ in (13), Tcha et al. develop a parametric version of the Erlenkotter algorithm based on lower and upper bounding functions $LB(\theta)$ and $UB(\theta)$, respectively, which are computed using a dual ascent method.

6 Applications

The fixed depot costs f_j and the variable supply costs c_{ij} in the UFLP are highly aggregated cost components. To estimate these costs, they must be decomposed into the costs of the driving factors and processes, which can vary from application to application. Since detailed cost analysis are often too expensive, or not possible because of missing information, or lead to cost functions which are too complicated to be handled efficiently in a mathematical model, only approximations of these costs are available in general. Beside the long-term planning horizon, the approximate nature of the cost data stresses the importance of sensitivity and parametric analysis in the

UFLP. Furthermore, the restructuring of a distribution system often requires large organizational changes, and it is not known in advance if the solution computed by a mathematical model can be really realized. Therefore, with respect to depot location problems, it is more important to provide alternative solutions, to evaluate deviations from the theoretical optimal solution and to provide insight into the effects of data variations than to compute a single "optimal" solution. A good means to this end is the analysis described in the preceding sections. In the sequel, we describe its application to two similar depot location problems.

6.1 Distribution of Dairy Products in Switzerland

The depot location problem to be discussed here, was proposed by a large producer and distributor of dairy products in Switzerland. The firm resulted from the merger of smaller companies producing dairy products. Each company had its own distribution system whose structure was more the result of the firm's history than of economic reasoning.

The whole firm operated 28 depots which serve approximately 18,000 single customers (supermarkets, retailers, ...) in Switzerland. To reduce the distribution costs, the management decided to close some of the existing depots (except 7 depots) and to change the allocation of customers to depots.

Since there were no restrictive capacity constraints, we proposed to use the UFLP as a model to support the decision on the new number and locations of depots and the customer assignment. For the purpose of the analysis, the 18,000 single customers were aggregated by the firm on the basis of postal codes to 1,400 customer nodes whose geographical distribution is shown in Fig. 6.

The management was – for several reasons – not able to provide data on the unit throughput costs and the fixed costs at the depots. Thus, we assumed equal throughput costs of the depots and decided – because of the unknown fixed costs – to compute the transportation cost function, i.e. the function Φ_{ufl} in (16) where $f_j = 0$ for $j = 1, \ldots, n$.

To estimate the variable transportation costs c_{ij}, we used two different approaches. In a first approach, these costs were computed as

$$c_{ij} = b_i \cdot 2 \cdot d_{ij} \cdot \omega_D \,,$$

where b_i is the annual demand of customer i, d_{ij} the length of the shortest path from depot j to customer node i in a road network of Switzerland and ω_D the transportation cost per unit weight and distance (provided by the firm). In a second approach, these costs were computed as

$$c_{ij} = 250 \cdot (2 \cdot d_{ij} \cdot k_i) \cdot T_c \,,$$

where T_c is the transportation cost per unit distance and k_i is an estimate of the number of vehicles, which visit customer node i daily (250 days per

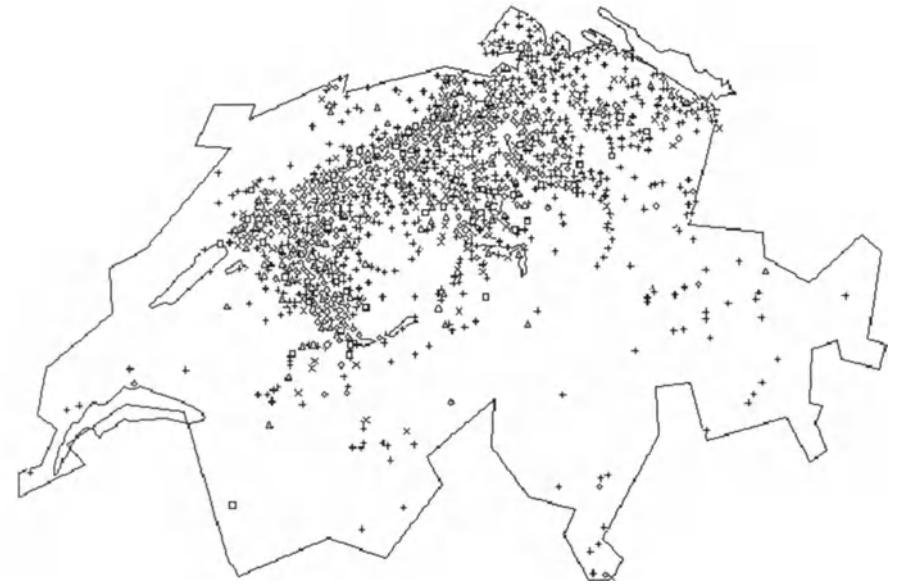


Fig. 6. Distribution of customer nodes (+: lowest sales, ◇: highest sales)

year; $k_i = \lceil (b_i/250)/Q \rceil$, Q = average vehicle capacity, $\lceil x \rceil$ smallest integer greater or equal than x).

The last approach expresses the daily demand of a customer in "units of vehicles" and assumes that every customer node is visited daily. Since the 1,400 customer nodes contained nodes with very low sales, a further aggregation of the 1,400 customer nodes were necessary. Using an aggregation scheme proposed by Daskin et al. (1989), we obtained 570 customer regions as follows: The first aggregate node is given by the node with the largest demand. In every step, all remaining nodes ("mini-nodes") are then assigned to the nearest aggregate node. The next aggregate node is the mini-node with greatest weighted distance (distance × demand) to its current aggregate node. This process is repeated until the desired number of aggregate nodes has been selected.

Instead of computing the cost curve Φ_{ufl} directly, we computed its convex hull $\overline{\Phi}_{\text{ufl}}$ with the help of the algorithm described in Sect. 4. Figure 7 shows the transportation cost curve (in 1,000 sFr. per year) which in this case coincides with its convex hull. The computation of the curve took only a few minutes on a 486 PC. Figure 8 shows the corresponding ranges $[\underline{\mu}_p, \overline{\mu}_p]$ of the optimal multiplier for $p = 1, \ldots, n$. Since the cost curve Φ_{ufl} is convex in this example, an optimal solution of the UFLP with fixed depot cost f where $\underline{\mu}_p \leq f \leq \overline{\mu}_p$ is given by the solution corresponding to $\Phi_{\text{ufl}}(p)$.

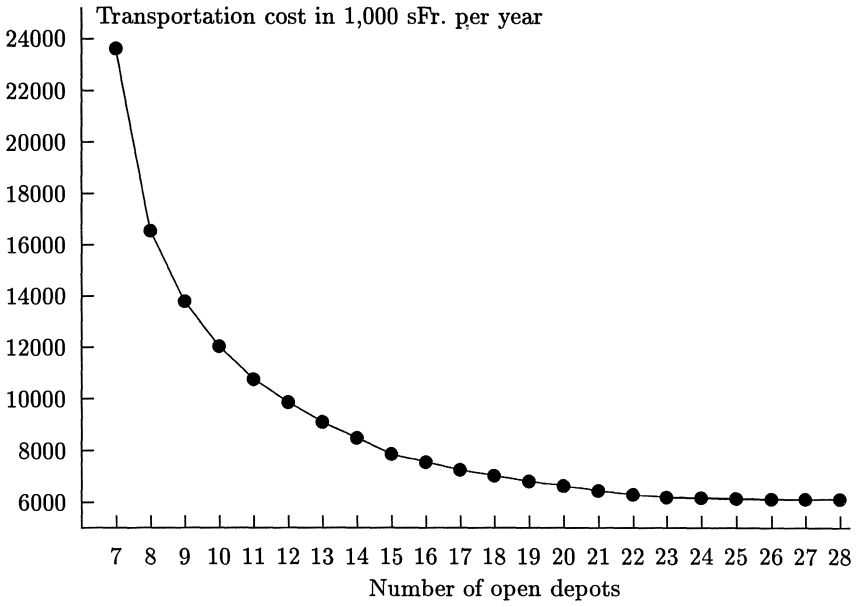


Fig. 7. Transportation cost curve

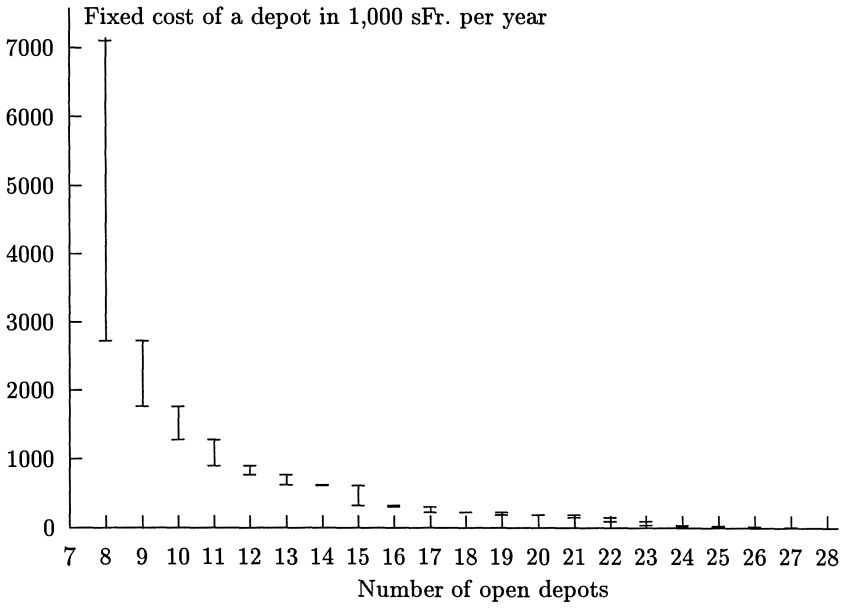


Fig. 8. Analysis of fixed costs

The transportation cost function and the associated "sensitivity analysis" of the fixed costs, based on the method of tangential approximation, were very helpful information for the management to decide on the new depot locations. Other helpful information in this respect were also tables of the different solutions corresponding to $\Phi_{\text{ufl}}(p)$ for $p = 1, \ldots, n$. These indicate the robustness of certain (sub-)solutions by showing which potential depot j belongs to a certain solution. This example shows that a mathematical location model can still give valuable results, despite the absence of detailed data.

6.2 A Depot Location Problem of a Food Producer

The second application is a location-allocation problem of a large food producer. For detailed information on this case-study, we refer to the paper of Tüshaus & Wittmann (1997) in this volume. The general structure of the distribution system is a two-stage system. The products, preserved food, are shipped to regional distribution centers from a central plant by external carriers. From the regional distribution centers, where the goods are stored and invoiced, the products are delivered to 15,000 customers on routes by the company's own vehicle fleet.

Since the structure of the actual distribution system is the result of a historical process, the management expected considerable savings from a restructuring of the distribution system. Therefore, the main objective of the study was to determine the optimal number and locations of the depots and to assess the cost saving potential of such a new distribution structure. To this end, the problem was modelled as an UFLP, since the capacity of the (existing) depots is sufficiently flexible.

In order to obtain a good coverage of the whole market area, the management selected 72 potential depot locations, which include the existing depots. The fixed depot costs have been assumed to be independent of the location and the throughput of the depots. A good estimate of these costs (rents, wages of the administrative staff, cost of electricity and maintenance, depreciation etc.) was available from the company's cost accounting. The cost c_{ij} to supply a customer i from a depot j consists of three components, the variable depot costs – which are nearly identical for all depots and can therefore be omitted –, the costs to transport the demand quantity from the plant to the depot – for which a freight tariff per unit distance and quantity is given –, and the cost $\hat{c}_{ij}$ to deliver the demand from the depot to the customer.

To estimate the costs $\hat{c}_{ij}$ is difficult, because the customers are served on vehicle routes. To obtain approximations, we had to estimate a customer's share t_{ij}^A of the travel time T_R of a route R which starts and ends at depot j. The estimations performed were based on empirical data of the vehicle tours for all existing depots. These data were provided by the firm's route planning system, where the single customers are aggregated to over 3,200

customer locations, which have been further aggregated to 2,600 customer nodes according to a 4-digit coordinate system, after the cost matrix $(\hat{c}_{ij})$ has been estimated.

In a first approach ("cost apportionment") a customer's share on travel time was obtained as

$$t_{ij}^{A} = \frac{2t_{ij} - \bar{t}_i}{\overline{n}_{ij}} + \bar{t}_i \ ,$$

where t_{ij} is the time to travel from j to i in a road network, $\bar{t}_i$ is the mean travel time to the next customer on a route, which serves customer i, and $\overline{n}_{ij}$ is the mean number of stops of a route, which serves customer i and starts in depot j. Since the travel time of a route may not exceed a certain maximum $T_{\max}$ the number of stops were estimated from

$$\overline{n}_{ij} = \min \left\{ \frac{T_{\max} - 2t_{ij} + \bar{t}_i}{\bar{t}_i + \bar{t}_i^{u}}, \ \overline{n}_i \right\} \ ,$$

where $\bar{t}_i^{u}$ is the mean load time and $\overline{n}_i$ the average number of stops on routes serving customer i. A second approach was to use the regression equation

$$T_R = \alpha \sum_{i \in R} t_{ij} + \beta \sum_{i \in R} \bar{t}_i + \epsilon \ ,$$

which explains the travel time T_R of a route R in dependence on the depot-customer travel times t_{ij} and the mean travel time $\bar{t}_i$ from the customers $i \in R$ on that route to the next two customers. On this basis, the share t_{ij}^{A} has been computed as

$$t_{ij}^{A} = (\alpha t_{ij} + \beta \bar{t}_i) \frac{\overline{n}}{\overline{n}_{ij}} \ ,$$

where $\overline{n}$ is the mean number of stops of all routes. Finally after multiplying t_{ij}^{A} with the cost per unit travel time and with the delivery frequency, the delivery costs $\hat{c}_{ij}$ have been obtained.

The algorithm to compute the convex hull of the perturbation function Φ_{ufl} (see Sect. 4) has been applied to the depot location problem of the food producer for the two different methods to estimate the service cost c_{ij}. The computations take approximately 14 hours (wallclock time) on a Sun Ultra (166 mhz) for each run. Figure 9 shows the convex hull $\overline{\Phi}_{\text{ufl}}$ of the cost curve Φ_{ufl} for a range of 2 to 20 open depots (at least 2 open depots are necessary to cover all customers). The underlying cost matrix has been obtained by the regression approach. The curve for the apportionment approach looks very similar. This, together with an analysis of the actual system costs, validates the cost estimation approach. The convex hull shown in Fig. 9 deviates from the cost curve Φ_{ufl} only at the point $p = 16$, where a small duality gap exists. The optimal solution is reached if 4 depots are open. But, the cost curve is flat around the optimum, which is typically the case for uncapacitated facility location problems, and solutions with 3 or 5 open depots are not much more

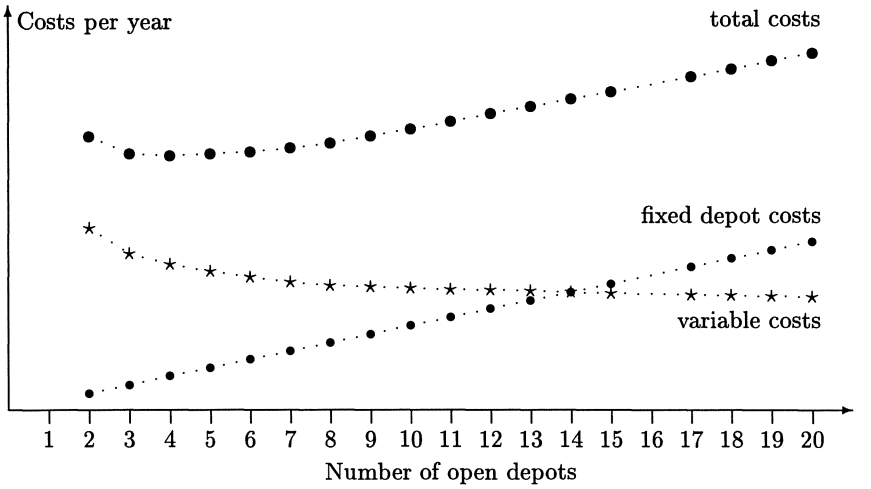


Fig. 9. Cost curve (regression)

expensive than the optimal solution. Since a solution with 5 open depots is easier to implement from an organizational point of view, the management gave priority to this solution. A comparison with the actual situation gives a potential saving of approximately 8.2 % of the actual system costs, which can be realized by implementing the optimal solution.

The results of the parametric analysis of the fixed and variable costs are shown in Fig. 10 and Fig. 11. The figures show the possible range of the fixed and variable costs (in percent of the actual fixed and variable costs), which do not affect the optimality of the solution corresponding to the given number of open depots. For example, the solution with 4 open depots would remain optimal if the fixed costs increased by 23 % or decreased by 20 %. Since the depots have equal fixed costs, this is equivalent to a decrease of the variable costs by 18 % and an increase of the variable costs of 26 %. Therefore, one can conclude, that the solution found is very stable and that estimation errors or data changes of reasonable ranges have no influence on its optimality.

7 Conclusions

In the context of depot location, the UFLP is a mathematical model for a strategic planning problem. Due to the long-term planning horizon, there is an uncertainty regarding the relevant cost components and the demand development. Furthermore, the cost coefficients in the UFLP, which measure the costs of assigning customers and the costs to operate the depots, are highly aggregated entities which are composed of a variety of different cost factors and can often only be estimated.

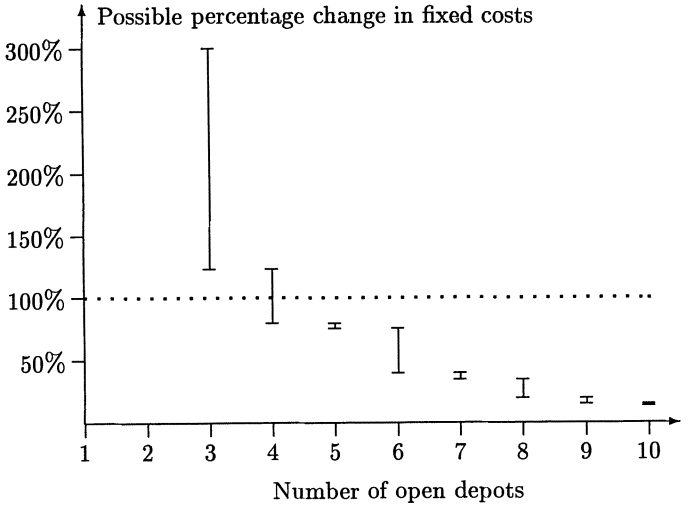


Fig. 10. Parametric analysis of fixed costs

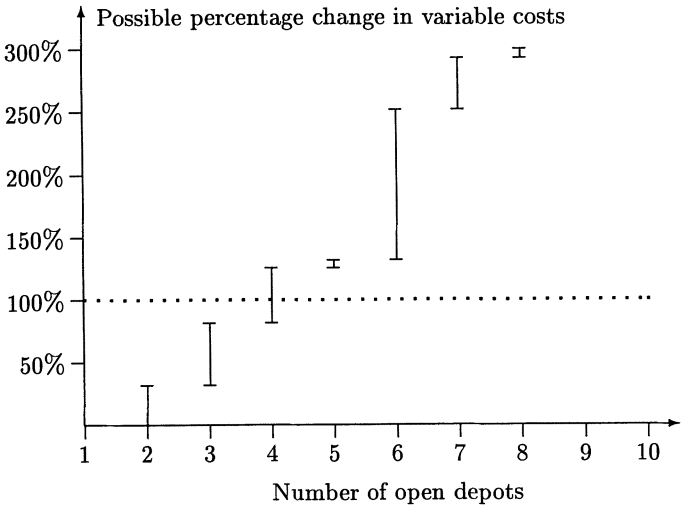


Fig. 11. Parametric analysis of variable costs

solutions and the effects of changes in fixed and variable costs. Regarding general integer programming problems, the analysis of the sensitivity of solutions and parameter changes is computationally a very burdensome task. Fortunately, the UFLP has a simple structure which facilitates parametric analysis substantially. Assuming identical fixed cost changes for all potential depots, a parametric analysis of the fixed depot costs can be performed by Lagrangean relaxation of a cardinality restriction and computing the convex hull of the resulting value function with the help of the tangential approximation algorithm. Futhermore, in the case of equal fixed costs, the results of this analysis can also be used to assess the effects of changes in the variable costs.

The approach of using Lagrangean relaxation and the concept of the perturbation function to perform a parametric analysis is attractive from a computational point of view if the convex hull of the pertubation function does not consist of too much linear segments. Regarding the UFLP, this holds for a parametric analysis of the fixed costs with equal fixed cost changes and may also hold if general fixed cost change vectors are considered, but – due to the high dimensionality of the parameter space – this is surely not true if a general analysis of changes in the assignment costs has to be performed. In this case direct parametric algorithmic approaches are required.

Acknowledgement: This research has been supported by the "Swiss Federal Commission for Technology and Innovation (KTI)".

References

Ahn, S. / Cooper, C. / Cornuejols, G. / Frieze, A. M. (1988): Probabilistic Analysis of a Relaxation for the k-Median Problem. *Mathematics of Operations Research*, 13:1–31.

Boffey, T. B. (1989): Location Problems Arising in Computer Networks. *The Journal of the Operational Research Society*, 40:347–354.

Christofides, N. / Beasley, J. E. (1982): A Tree Search Algorithm for the p-Median Problem. *European Journal of Operational Research*, 10:196–204.

Cornuejols, G. / Fisher, M. L. / Nemhauser, G. L. (1977): Location of Bank Accounts to Optimize Float: An Analytic Study of Exact and Approximate Algorithms. *Management Science*, 23:163–177.

Current, J. / Weber, C. (1994): Application of Facility Location Modeling Constructs to Vendor Selection Problems. *European Journal of Operational Research*, 76:387–392.

Daskin, M. S. / Haghani, A. E. / Khanal, M. / Malandraki, C. (1989): Aggregation Effects in Maximum Covering Models. *Annals of Operations Research*, 18:115–140.

Erlenkotter, D. (1978): A Dual-Based Procedure for Uncapacitated Facility Location. *Operations Research*, 26:992–1009.

Geoffrion, A. M. (1974): Lagrangean Relaxation for Integer Programming. *Mathematical Programming Study*, 2:82–114.

Greenberg, H. J. (1977): The One-Dimensional Generalized Lagrange Multiplier Problem. *Operations Research*, 25:338–345.

Hanjoul, P. / Peeters, D. (1985): A Comparison of two Dual-Dased Procedures for Solving the p-Median Problem. *European Journal of Operational Research*, 20:387–396.

Jenkins, L. (1990): Parametric Methods in Integer Linear Programming. *Annals of Operations Research*, 27:77–96.

Klose, A. (1993): *Das kombinatorische p-Median-Modell und Erweiterungen zur Bestimmung optimaler Standorte.* Dissertation Nr. 1464, Hochschule St. Gallen.

Klose, A (1994): A Branch and Bound Algorithm for an Uncapacitated Facility Location Problem with a Side Constraint. Working paper, Institut für Unternehmensforschung (Operations Research), Hochschule St. Gallen.

Klose, A. (1995): A Comparison Between the Erlenkotter Algorithm and a Branch and Bound Algorithm Based on Subgradient Optimization to Solve the Uncapacitated Facility Location Problem. In: Ulrich Derigs, Achim Bachem, and Andreas Drexl, editors, *Operations Research Proceedings 1994*, pages 335–339, Berlin Heidelberg New York, Springer.

Körkel, M. (1989): On the Exact Solution of Large-Scale Simple Plant Location Problems. *European Journal of Operational Research*, 39:157–173.

Krarup, J. / Pruzan, P. M. (1983): The Simple Plant Location Problem: Survey and Synthesis. *European Journal of Operational Research*, 12:36–81.

Minoux, M. (1986): *Mathematical Programming.* Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Chichester, New York.

Mirchandani, P. B. / Oudjit, A. / Wong, R. T. (1985): Multidimensional Extensions and a Nested Dual Approach for the m-Median Problem. *European Journal of Operational Research*, 21:121–137.

Morris, J. G. (1978): On the Extent to which Certain Fixed-Charge Depot Location Problems Can Be Solved by LP. *The Journal of the Operational Research Society*, 29:71–76.

Nauss, R. M. / Markland, R. E. (1981): Theory and Application of an Pptimizing Procedure for Lock Box Location Analysis. *Management Science*, 27:855–865.

Nemhauser, G. L. / Wolsey, L. A (1988): *Integer and Combinatorial Optimization.* Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Chichester, New York.

Tcha, D.-W. / Myung, Y.-S. / Chung, K.-H. (1995): Parametric Uncapacitated Facility Location. *European Journal of Operational Research*, 86:469–479.

Tüshaus, U. / Wittmann, S. (1997): Strategic Logistic Planning by Means of Simple Plant Location: A Case Study. (this volume).

Walukiewicz. S. (1981): Some Aspects of Integer Programming Duality. *European Journal of Operational Research*, 7:196–202.

Strategic Logistic Planning by Means of Simple Plant Location: A Case Study

Ulrich Tüshaus[1] and Stefan Wittmann[2]

[1] Universität der Bundeswehr Hamburg, 22039 Hamburg, Germany
[2] Universität St. Gallen, 9000 St. Gallen, Switzerland

Abstract. The solution of a distribution facility location problem in Switzerland is discussed. Possible structures of the system were either a warehouse system with two stages (plant(s) → warehouses → clients) or a transshipment-point system with three stages (plant(s) → warehouses → transshipment-points → clients). Both alternatives were formulated as Simple Plant Location Problems. For this model, the determination of the necessary data basis is described, taking especially into account the less-than-truckload-mode on the last distribution stage. Results of the optimizations as well as the conducted sensitivity analyses are given.

1 Introduction

Locating facilities for a distribution system means to determine the number, the size, and the locations of facilities—together with the allocation of customers to the facilities—such that the costs of distribution are minimized. An optimal balance has to be found between the fixed costs of operating facilities and the variable costs of serving the customers from those facilities.

This task had to be carried out for a producer of non-perishable food for the distribution in Switzerland (see also Tüshaus and Wittmann 1997 as well as Klose and Stähly 1997b). The overall structure of the distribution system was predetermined by the management, to be either a two-stage warehouse or a three-stage transshipment-point system. In both alternatives the last distribution stage, the transport of the goods from the warehouses or transshipment-points, resp., to the clients is carried out by the company's own vehicle fleet in routes. All other transport—"Plant(s) → Warehouses" and "Warehouses → Transshipment-points"—is done by contractors for a given cost unit rate per ton-kilometer.

To solve these types of problems a variety of mathematical facility location models are in principle applicable. The general rule for the selection of an appropriate model is to be as simple as possible and to use more complex models only if the additional insight won justifies the increased complexity. Following this rule, the decision was made to use the most basic locational model, the well known Simple Plant Location Problem (SPLP).

When using the SPLP model, it is necessary to assume that the cost to deliver the demand of a customer i from a potential facility site j can be expressed as a function of some variable(s) depending on the customer and/or

the facility but not on any other customer. Implicitly, this means that the cost to supply a customer is independent of the supply of other customers. But, if several customers are served together on one route, this implicit assumption is not valid. In this case, the calculation of the necessary cost coefficients is not a straightforward task.

In the following, a description of the problem is given and its modeling as Simple Plant Location Problems is discussed. The determination of the problem data, of the fixed costs f_j for potential facility sites and especially of the cost coefficients c_{ij} for serving the customers is shown, as well as the results of the calculations and sensitivity analyses conducted.

2 Problem Description

For a producer of non-perishable food with a single production plant the distribution system for Switzerland with more than 15'000 clients, supplied at least once a week, had to be revised. As general structure of the system, the management of the company had chosen to implement either a two-stage structure with only regional warehouses (W) or a three-stage structure with regional warehouses and transshipment-points (T) (see Fig. 1).

In the case of the two-stage structure "Warehouses" the distribution of the goods is carried out as follows: From the single production plant the goods are transported in a first distribution stage to the regional warehouses by contractors for a given rate per ton-kilometer. After warehousing and repacking they are delivered on routes to the individual clients in a second step by the company's own vehicle fleet. This flow of goods can be summarized as:

$$\text{Plant(s)} \rightarrow \text{Warehouses} \rightarrow \text{Clients}.$$

The structure "Transshipment-Points" is characterized by the fact that the distribution generally takes place in three steps: First, the goods are transported from the production plant to regional warehouses. In a second step, they are—after warehousing and repacking—transported to the transshipment-points. Having arrived at the transshipment-points, the goods are immediately (without further warehousing) repacked and finally shipped to the individual clients. The resulting flow of goods is given by:

$$\text{Plant(s)} \rightarrow \text{Warehouses} \rightarrow \text{Transshipment-Points} \rightarrow \text{Clients}.$$

While the first two stages are realized by contractors for a given rate per ton-kilometer, the last stage, the fine distribution, is carried out in routes by the company's own vehicle fleet.

The clients of the firm range from small mountain restaurants to large supermarkets; private clients are not supplied. Spread all over Switzerland, these commercial clients sum up to more than 15'000 individual drop places,

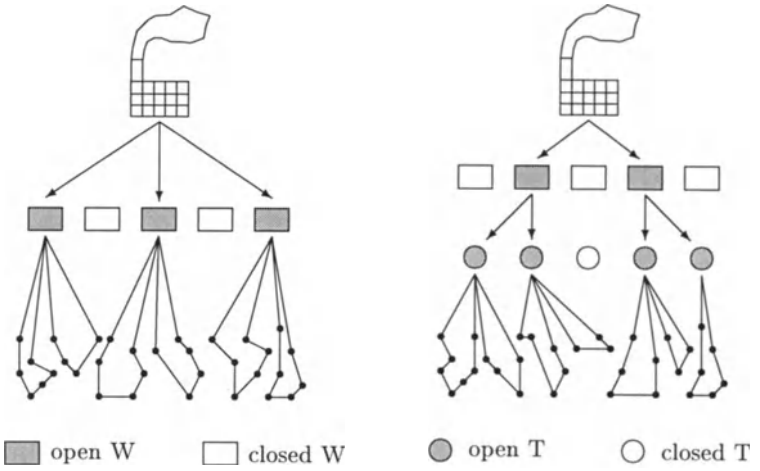


Fig. 1. "Warehouses"- vs. "Transshipment-Points"-structure

located in more than 3'200 villages and postal delivery areas. Their delivery frequency ranges from every day delivery to once or twice a week.

Potential sites for warehouses and/or transshipment-points were determined by the management. Besides legal and organizational aspects of building a facility at a specific site, the selection was guided by the following three considerations: First, every current warehouse site had to be a potential site for a warehouse or a transshipment-point in the new system; second, the whole distribution area had to be covered more or less evenly, taking into account the special topographic structure of Switzerland (valleys, mountain ranges etc.); and third, the number of potential sites should not exceed 100, forcing an actual selection. Within this framework the management came up with 72 potential sites, for both warehouses and transshipment-points.

3 Modeling the Problem

Modeling a specific problem means first, to select an adequate mathematical model, and second, to determine appropriately the model parameters. This is done in the following: The selected mathematical model, the Simple Plant Location Problem (SPLP), is introduced, the general idea how to model the two structures as SPLPs is given and the determination of the parameters is discussed in detail.

3.1 Mathematical Model: SPLP

Due to the special structure of the problem—a single production plant and sufficiently flexible capacities of the facilities—both, the two-stage "Ware-

houses" structure as well as the three-stage "Transshipment-Points" structure could be modeled and solved as uncapacitated one- and two-stage systems, resp. (see below), using the well know Simple Plant Location Problem (SPLP) formulation (see e.g. Krarup and Pruzan 1983):

$$v(\text{SPLP}) = \min \left(\sum_{j=1}^{m} f_j \cdot y_j + \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \cdot x_{ij} \right) \tag{1}$$

$$\text{s.t.} \quad \sum_{j=1}^{m} x_{ij} = 1 \qquad \forall i \tag{2}$$

$$x_{ij} \leq y_j \qquad \forall i, j \tag{3}$$

$$x_{ij},\, y_j \in \{0,1\} \qquad \forall i, j\,, \tag{4}$$

where:

c_{ij} cost coefficient (cost to supply customer i from facility site j

f_j fixed cost of a facility at site j

i index of customers; $i = 1, \ldots, n$

j index of potential facility sites; $j = 1, \ldots, m$

m number of potential facility sites

n number of customers

x_{ij} assignment of customer i to the facility at site j:

$$x_{ij} = \begin{cases} 1 & \text{if assigned} \\ 0 & \text{otherwise} \end{cases}$$

y_j opening of the facility at site j:

$$y_j = \begin{cases} 1 & \text{if open} \\ 0 & \text{otherwise} \end{cases}$$

The objective is to minimize the sum of the fixed costs caused by operating the facilities ($\sum_{j=1}^{m} f_j \cdot y_j$) and of the variable costs to serve the customers ($\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \cdot x_{ij}$) under the restriction that every customer is assigned to one and only one open facility.

Decision variables of the problem are the y_j and the x_{ij}: Opening a facility at a specific site j is expressed by the value of the y_j (1 if site j is used in a solution; 0 otherwise); which customers are assigned to a facility can be seen from the values of the x_{ij} (1 if i is assigned to j; 0 otherwise).

The problem parameters are the fixed cost f_j of every potential facility site j and the cost c_{ij} to serve the demand of customer i from the potential facility site j.

3.2 Modeling the Structures as SPLPs

To model the two structures "Warehouses" and "Transshipment-Points" as Simple Plant Location Problems, the problem parameters f_j and c_{ij} had to be determined adequately. In the following, the general idea is given, how to this end the costs of the different facility types and those of the single transportation stages are used. Thereby, as facility type related costs, the fixed costs f_j^{W} and f_j^{T} of operating a warehouse and a transshipment-point, resp., had to be

considered. As costs of the single transportation stages the costs c_{ij}^{PW}, c_{ij}^{WC}, c_{ij}^{WT}, and c_{ij}^{TC} are used for the stages "Plant $\rightarrow$ Warehouse", "Warehouse $\rightarrow$ Client", "Warehouse $\rightarrow$ Transshipment-Point", and "Transshipment-Point $\rightarrow$ Client", resp.

Structure "Warehouses". Although the structure "Warehouses" implies two stages, it could be modeled and solved directly as a Simple Plant Location Problem (a single-stage model): Since only one production plant supplies all m potential warehouse sites, the allocation of warehouses to plants was no question. By consequence, the cost c_{ij} to supply client i from warehouse j, caused by the transport over the two stages "Plant $\rightarrow$ Warehouse" and "Warehouse $\rightarrow$ Client", could be determined by adding the two transportation cost components. The fixed costs f_j of the problem were given by the fixed costs f_j^{W} of operating the warehouses. Thus, the parameters of the model to solve were given by:

$$f_j = f_j^{\mathrm{W}} \quad \forall j \quad \text{and} \quad c_{ij} = c_{ij}^{\mathrm{PW}} + c_{ij}^{\mathrm{WC}} \quad \forall i, j. \tag{5}$$

Structure "Transshipment-Points". To be able to model the structure "Transshipment-Points" again with the help of the simple SPLP model, it had to be decomposed into two parts: on the one hand the determination of the warehouses and on the other the one of the transshipment points. Note, thereby only optimal solutions for the two subproblems—and not a general optimum—are obtained.

The first step was the determination of transshipment-points. Concerning the fixed costs of operating transshipment-points and the costs to transport the client demands from the transshipment-points to the clients, formally:

$$f_j = f_j^{\mathrm{T}} \quad \forall j \qquad \text{and} \qquad c_{ij} = c_{ij}^{\mathrm{TC}} \quad \forall i, j\,, \tag{6}$$

the optimal number $p^\star$ and locations of transshipment-points j ($j = 1, \ldots, m$) as well as the optimal allocation of clients i ($i = 1, \ldots, n$) to those transshipment-points were determined with a SPLP-algorithm.

Using the so obtained transshipment-points, the number and location of warehouses were calculated in a second step with a SPLP-algorithm, too. This time the customers i were the transshipment-points ($i = 1, \ldots, p^\star$) obtained in the first step. Their demand was given by the aggregated demand of all clients allocated to a specific transshipment-point in the first step. In analogy to the structure "Warehouses" the cost coefficients of this step modeled not only one distribution stage, but two ("Plant $\rightarrow$ Warehouses" and "Warehouses $\rightarrow$ Clients"). As fixed costs the costs of operating a warehouse were used. This way the optimal number $w^\star$ and locations of warehouses j ($j = 1, \ldots, m$) with regard to the fixed costs of operating the warehouses and the costs to supply the transshipment-points established in the first step

from the plant over the warehouses, namely:

$$f_j = f_j^{\mathrm{W}} \quad \forall j \qquad \text{and} \qquad c_{ij} = c_{ij}^{\mathrm{PW}} + c_{ij}^{\mathrm{WT}} \quad \forall i, j\,, \tag{7}$$

could be determined.

To guarantee that orders placed until early in the morning will be delivered the very same day, it was necessary to ensure—by an additional restriction—that each transshipment-point is allocated to a warehouse not more than an hour drive away.

3.3 Determining the Model Parameters

The model parameters f_j and c_{ij} were determined on the basis of (historical) client and cost-accounting information and a road-network of Switzerland. Due to the fact that every client is delivered at least once a week, as underlying time period a week was chosen.

Fixed Costs f_j. Fixed costs are caused by the operation of facilities. Since the main portion of the cost of a facility was the labor cost, the problem parameters "fixed costs of the facilities" were more or less independent of the geographic position. Thus, they were assumed to be equal for all potential sites:

$$f_j = f \quad \forall j\,. \tag{8}$$

For a warehouse, the fixed cost was estimated mainly on the basis of cost accounting data, concerning existing warehouses. Since up to then transshipment-points have not been operated, the fixed cost of a transshipment-point had to be estimated by the sum of the expected labor costs, rents, energy, and telecommunication costs as well as expected depreciations on investments.

Comparing the resulting values, it turned out that operating a transshipment-point is more expensive then expected: The fixed cost f^{T} of a transshipment-point is about 10% of the fixed cost f^{W} of a warehouse.

Cost Coefficients c_{ij}. Much more difficult than the determination of the fixed costs was the one of the cost coefficients c_{ij}: First, more than one distribution stage had to be considered, and second, the fine distribution (serving the clients from the warehouses or transshipment-points, resp.), which causes the main portion of the transportation cost, is carried out in the less-than-truck-load mode.

Plant → Warehouses: c_{ij}^{PW}. Since the transport of the goods from the single production plant to the warehouses is done by contractors for a given rate r (cost unit rate per ton-kilometer), the transportation costs c_{ij}^{PW} are:

$$c_{ij}^{\mathrm{PW}} = r \cdot d_{pj} \cdot b_i \quad \forall i, j\,. \tag{9}$$

Here, d_{pj} is the road network distance (in km) between the production plant p and the potential warehouse site j and b_i is the demand of customer i (in tons per period). In case of the structure "Warehouses", the customers i were the clients of the firm and for the structure "Transshipment-Points", the i represented the transshipment-points for the second optimization step.

Warehouses → Transshipment-Points: c_{ij}^{WT}. Analogous to the transportation stage "Plant → Warehouses", the transport of the goods from the warehouses to the transshipment-points is carried out by contractors. The costs c_{ij}^{WT} are in principle:

$$c_{ij}^{\mathrm{WT}} = r \cdot d_{ij} \cdot b_i \quad \forall i, j\,, \tag{10}$$

where d_{ij} is the road network distance (in km) between the warehouse j and the transshipment-point i, and the term b_i denotes the aggregated demand of all clients assigned to transshipment-point i (in tons per period).

To incorporate the additional restriction that a transshipment-point has to be reached within one hour from the respective warehouse it is allocated to, Eqn.(10) was modified to:

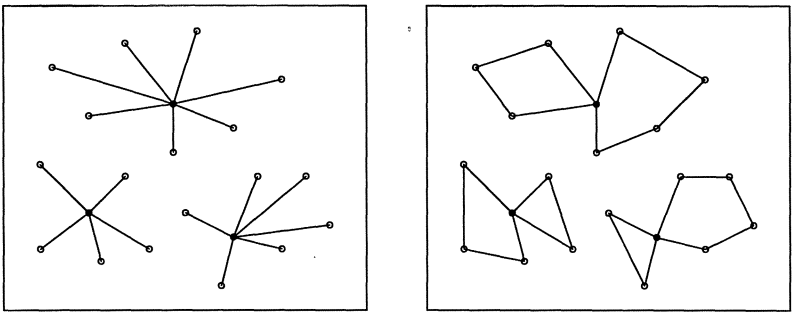
$$c_{ij}^{\mathrm{WT}} = \begin{cases} r \cdot d_{ij} \cdot b_i & \text{if } t_{ij} \leq 60\,\mathrm{min} \\ \infty & \text{otherwise} \end{cases} \quad \forall i, j\,, \tag{11}$$

with t_{ij} as the time given by the road network to drive from warehouse j to the transshipment-point i.

Warehouses or Transshipment-Points → Clients: c_{ij}^{WC} *or* c_{ij}^{TC}. When using the SPLP model, it is implicitly assumed that the cost to supply the clients are independent of each other. But, if several customers are served together on one route, this implicit assumption is not valid. Hence, the determination of the coefficients c_{ij}^{WC} or c_{ij}^{TC} is not a straightforward task.

"Model" vs. "Reality": The fine distribution is realized with the company's own vehicle fleet in the less-than-truckload mode. On routes the goods are transported from the warehouses or transshipment-points to the clients, resp. Using a simple function of the radial distance to calculate the costs c_{ij}^{WC} or c_{ij}^{TC}, it would have been implicitly assumed that the costs for a specific client are independent of the delivery of other clients. Since the clients are served on routes, this implicit assumption is not valid (see Fig. 2). Therefore, a different approach had to be taken: Client i's share c_{ij}^{WC} or c_{ij}^{TC} of the total cost of a tour starting at warehouse or transshipment-point j, resp., and serving this client (among others) had to be estimated.

Databasis: The databasis for this estimation was detailed tour information collected over one year, for the more than 15'000 individual clients on the aggregated client level of 3'200 villages and postal delivery areas (as distinguished by the firm's vehicle routing software). Since the locational road



(a) Model (b) Reality

Fig. 2. Relationship "Model" vs. "Reality"

network used was not the same as the one used by the firm's routing software, a second aggregation step had to be done. Every aggregated client i' (village or postal delivery area), out of the total of 3'200, was assigned to its closest node (euclidean distance) in the locational road network; the corresponding demand data were aggregated adequately. This second step resulted in about 2'600 client nodes i, out of a total of 6'677 network nodes, which were the finally used "clients". (With a Swiss surface of $41'293\,\mathrm{km}^2$ and 6'677 nodes on an average, every node covers a square with an edge length of 2.5 km.)

Parameter Components: From cost accounting it was known that the transportation costs of the current distribution system were mainly time dependent labor cost. Due to that, a time-based calculation of the cost components c_{ij}^{WC} and c_{ij}^{TC} was used. For this, the working day of a tour driver was divided into three components: First, the time spent in the morning and the evening at the facility site he is stationed at (loading etc); second, the actual driving time, and third, the time needed to unload the merchandise at the clients.

On the basis of the collected tour data and following this decomposition, the cost c_{ij}^{WC} and c_{ij}^{TC} caused by supplying client node i from the potential facility j during one period could be expressed as follows: It is given by the sum of the corresponding service costs $c_{i'j}^{\mathrm{WC}}$ (or $c_{i'j}^{\mathrm{TC}}$ resp.) of all villages and postal delivery areas i' assigned to client node i ($i' \in I'(i)$). The cost $c_{i'j}^{\mathrm{WC}}$ (or $c_{i'j}^{\mathrm{TC}}$ resp.) for a single village i' again is the sum of its portion $c_{i'}^{\mathrm{l}}$ of the cost at the facility site, its portion $c_{i'j}^{\mathrm{d}}$ of the actual driving cost, and the cost $c_{i'}^{\mathrm{u}}$ to unload (all per delivery), weighted with the delivery frequency $h_{i'}$ per period; formally (for $c_{i'j}^{\mathrm{TC}}$ holds the same):

$$c_{ij}^{\mathrm{WC}} = \sum_{i' \in I'(i)} c_{i'j}^{\mathrm{WC}} = \sum_{i' \in I'(i)} h_{i'} \cdot \left(c_{i'}^{\mathrm{l}} + c_{i'j}^{\mathrm{d}} + c_{i'}^{\mathrm{u}}\right) \quad \forall\, i, j\,. \tag{12}$$

While the coefficients $c_{i'}^{\mathrm{l}}$ and $c_{i'}^{\mathrm{u}}$ were easy to determine, the $c_{i'j}^{\mathrm{d}}$ were not.

Loading and Unloading Costs: Both, the portion of the cost caused by the time spent at the facility site and the cost caused by the time needed to unload at the individual clients of a village, were calculated by multiplying the actual time needed ($t^{\mathrm{l}}_{i'}$ and $t^{\mathrm{u}}_{i'}$ resp.) by a common cost unit rate $p^{\mathrm{l,u}}$:

$$c^{\mathrm{l}}_{i'} = t^{\mathrm{l}}_{i'} \cdot p^{\mathrm{l,u}} \qquad \text{and} \qquad c^{\mathrm{u}}_{i'} = t^{\mathrm{u}}_{i'} \cdot p^{\mathrm{l,u}} \quad \forall\, i' \,. \tag{13}$$

The village specific times $t^{\mathrm{u}}_{i'}$ for unloading could be extracted directly from the collected tour data. For every village an equal share $t^{\mathrm{l}}_{i'} = t^{\mathrm{l}}$ $(\forall\, i')$ of the time spent at the facility site could be determined by subtracting the total unloading time and the total driving time (given in the tour data) from the total working time of the drivers during the data collection period and dividing this value by the total number of villages visited during the period. In this sense, the terms $t^{\mathrm{l}}_{i'}$ are residual components.

The necessary cost unit rates $p^{\mathrm{l,u}}$ could be gathered on the basis of cost accounting information. The main cost component were the driver wages (about 75%). In addition to that, overhead charges had to be considered.

These two cost components, $c^{\mathrm{l}}_{i'}$ and $c^{\mathrm{u}}_{i'}$, are independent of the location of the facilities and the allocation of clients; in consequence, they do not influence the structure of the solution. Nevertheless, they were considered for the following reasons: First, to obtain total system cost in the model comparable to total system cost in cost accounting, and second, to partition cost components correctly, often subsumized incorrectly in a cost unit rate per distance unit. In logistical practice, often the rate per distance unit is determined by dividing some sum of vehicle fleet costs by the total annual distance traveled by the fleet, resulting in a heavily overestimated rate (see Sec. 4.1).

Driving Costs: Since the fine distribution is carried out in a less-than-truckload-mode, the driving cost for a client is not independent of the one for the other clients. In consequence, for each village i' its share $c^{\mathrm{d}}_{i'j}$ of the driving costs has to be estimated. To this end, a much more difficult approach than discussed so far had to be taken.

Again, the basic principle used to calculate the coefficients $c^{\mathrm{d}}_{i'j}$ was to weight the driving time share $t^{\mathrm{d}}_{i'j}$ of village i' with a cost unit rate p^{d} for the driving time:

$$c^{\mathrm{d}}_{i'j} = t^{\mathrm{d}}_{i'j} \cdot p^{\mathrm{d}} \quad \forall\, i' \,. \tag{14}$$

Similar to the loading and unloading costs, also for the costs caused by the actual driving, a purely time based cost element had to be used for the cost unit rate p^{d}: The share, falling to village i', of the cost caused by the time needed to drive the whole tour. In addition to that time dependent cost share, a share of the distance dependent costs of the trucks also falls to a village i'. To obtain a common measuring unit "time" for this second cost element, the distance related cost factor p^{d}_{d} for the distance dependent costs—this could be easily established from cost accounting—was transformed into a time related

cost factor p_t^{d}. (The distance dependent factor was multiplied with the total distance driven during the data collection period, and this value was divided by the total driving time during the period.) The cost factor p^{d} for the driving time was then given by the sum of those two factors:

$$p^{\mathrm{d}} = p^{\mathrm{l,u}} + p_t^{\mathrm{d}}\,, \tag{15}$$

where the purely time driven factor $p^{\mathrm{l,u}}$, mainly due to the drivers' wages (see above), makes up more than 72% of the total factor p^{d}.

With respect to the time share $t_{i'j}^{\mathrm{d}}$ of village i', both estimation schemes discussed in the literature were used: apportionment and regression (see e.g. Fleischmann 1979 as well as Klose and Tüshaus 1995).

The idea of the apportionment scheme is to estimate the time share of a village i' by apportioning the total cost of a tour to the corresponding villages. As main cost factors of a tour could be identified: the stem time, the stop time, and the variable running time. The first is the time needed to travel from the facility to the first client of a tour and from the last client back to the facility, the second is the time needed for unloading, and the third the time needed to travel to the next client (see Fig. 3).

Based on these elements, the total time $T(\cdot)$ needed for a tour including village i' could be estimated as:

$$T(\bar{t}_{i'}, \bar{t}_{i'}^{\mathrm{u}}, t_{i'j}, n_{i'j}) = 2 \cdot t_{i'j} + (n_{i'j} - 1) \cdot \bar{t}_{i'} + n_{i'j} \cdot \bar{t}_{i'}^{\mathrm{u}}\,. \tag{16}$$

Here, twice the driving time $t_{i'j}$ between the facility j and the village i' was used as an estimate for the stem time; the variable running time was estimated by the product of the number of stops $n_{i'j}$ (minus one); and the mean driving time $\bar{t}_{i'}$ to the next village on routes, which serve village i', and the stop time was estimated by the product of the number of possible stops and the mean unloading time $\bar{t}_{i'}^{\mathrm{u}}$ at village i'.

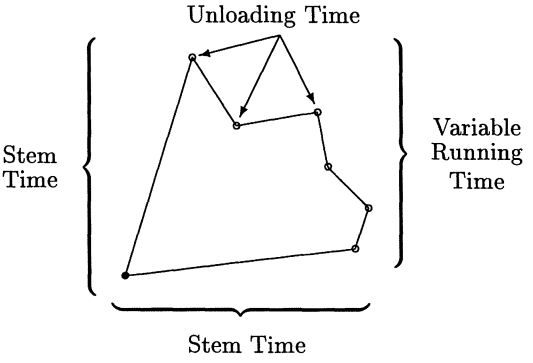


Fig. 3. Components of a tour

The driving time $t_{i'j}$ between the potential depot j and the village i' was evaluated using the road network based on the information given in the tour data. The same way the mean driving times $\bar{t}_{i'}$ to the respective two closest villages were determined. Thereby, every village was represented by the closest node (euclidean distance) in the road network. Problems, arising from assigning more than one village to a network node, were taken care of by dividing the values in question by the number of villages assigned to the node. The mean unloading time $\bar{t}^{\mathrm{u}}_{i'}$ at village i' was taken from the tour data directly. The same holds in principle for the number of stops $n_{i'j}$. Here, the average number $\bar{n}_i$ of stops on routes serving village i' could be used. However, a correction had to be made to take into account the maximum legal duration $T_{\max}$ of a tour, given by state regulations for drivers. A maximum for $n_{i'j}$ was given by:

$$2 \cdot t_{i'j} + n_{i'j} \cdot (\bar{t}_{i'} + \bar{t}^{\mathrm{u}}_{i'}) - \bar{t}_{i'} \leq T_{\max}\,, \tag{17}$$

the relationship of this maximum legal route duration and the total route duration. In consequence, the number $n_{i'j}$ of stops was estimated as the average number $\bar{n}_i$ of stops (taken from the data) adjusted to match the maximum driving time:

$$n_{i'j} = \min\left\{\frac{T_{\max} - 2 \cdot t_{i'j} + \bar{t}_{i'}}{\bar{t}_{i'} + \bar{t}^{\mathrm{u}}_{i'}}, \bar{n}_i\right\}\,. \tag{18}$$

Dividing the total time $T(\cdot)$ needed for a tour including village i' by the number of visited villages $n_{i'j}$ then gave the time share of a tour (including stop time) which falls to village i', formally:

$$\frac{2 \cdot t_{i'j} - \bar{t}_{i'}}{n_{i'j}} + \bar{t}_{i'} + \bar{t}^{\mathrm{u}}_{i'}\,. \tag{19}$$

Thus, according to the apportionment approach the shares of the driving time were estimated as:

$$t^{\mathrm{d}}_{i'j} = \frac{2 \cdot t_{i'j} - \bar{t}_{i'}}{n_{i'j}} + \bar{t}_{i'} \quad \forall\, i', j\,. \tag{20}$$

In a similar way, the regression approach uses characteristic factors to determine the share of the tour driving time which falls to a single village. Using multiple regression, the relationship between the driving time of a tour and some independent variables is found on the basis of tour data. This relationship can then be used to calculate the time share for all allocations of villages to potential depots.

To be able to calculate the time share for all allocations, it is necessary that the used variables are independent of the assignment of the villages. Examples for such variables are the times $t_{i'j}$ needed to travel from the depots to the villages and, as a measure for the spatial closeness of the villages, the mean travel times $\bar{t}_{i'}$ to the respective two nearest villages.

Using these variables, the driving time T_r^{d} (without stop time) needed for a tour r could be expressed as regression equation with the parameters α and β and the residual ϵ as:

$$T_r^{\mathrm{d}}(t_{i'j}, \bar{t}_{i'}) = \alpha \cdot \sum_{i' \in I'(r)} t_{i'j} + \beta \cdot \sum_{i' \in I'(r)} \bar{t}_{i'} + \epsilon \quad \forall\, r\,, \tag{21}$$

where $I'(r)$ was the set of villages visited on tour r.

After determining the values of the regression parameters on the basis of the tour data, in principle, it was possible to calculate the driving time share $t_{i'j}^{\mathrm{d}}$ for any allocation of a village i' to a potential depot j according to:

$$t_{i'j}^{\mathrm{d}} = \alpha \cdot t_{i'j} + \beta \cdot \bar{t}_{i'}\,. \tag{22}$$

But, for the regression approach, too, held that the number of villages visited on a tour was restricted by the maximum tour duration. Parameters determined by regression analysis evenly level the actual values implicitly to some "mean" value. Thus, the actual values for a specific village i' had to be corrected by the quotient of the mean value $\bar{n}$ and the assignment specific value $n_{i'j}$ for the estimated number of visited villages.

According to the regression approach, the shares of the tour driving time were then estimated as:

$$t_{i'j}^{\mathrm{d}} = \left(\alpha \cdot t_{i'j} + \beta \cdot \bar{t}_{i'}\right) \cdot \frac{\bar{n}}{n_{i'j}} \quad \forall\, i', j\,, \tag{23}$$

where the required values $t_{i'j}$, $\bar{t}_{i'}$, and $n_{i'j}$ were evaluated in the same way as for the apportionment approach. The value $\bar{n}$ was taken from the tour data.

Summary: Figure 4 gives a graphical summary of the components and their sources needed to calculate the cost coefficients.

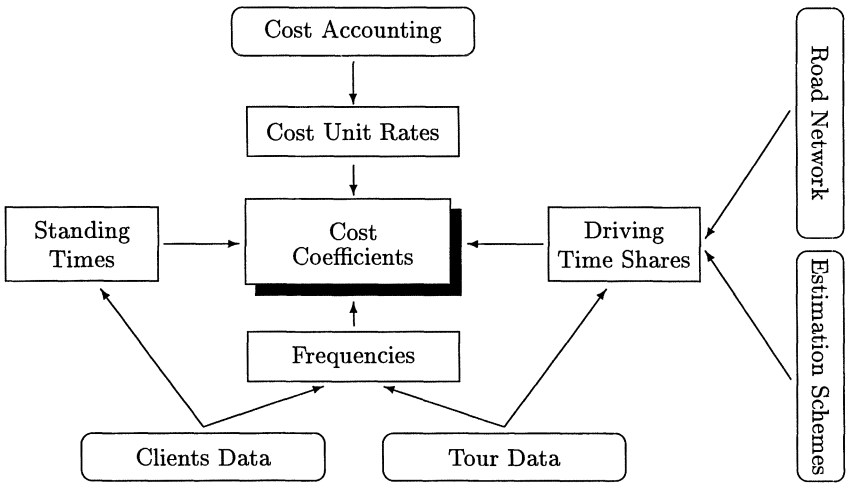


Fig. 4. Components of the cost coefficients and their sources

4 Analysis of the Problem

Based on the modeling approach discussed in the previous section, the problem was analyzed: The actual system was depicted by the model, and optimization calculations, sensitivity analyses as well as some further analyses were carried out for the two structures.

4.1 Analysis of the Actual System

In order to validate the modeling approach taken, the actual two-stage structure with only regional warehouses was analyzed on the basis of the model "Warehouses" in two ways: time- and cost-oriented.

Time Comparison. A purely time-oriented comparison of the reality and its representation by the model was done to verify the estimation schemes used to determine the time share of a single client. For this, the model parameters of the structure "Warehouses" were determined as discussed in Sec. 3.3, but with all cost unit rates set to zero, except p^{d} which was set to one. Using these parameters, a solution was generated with the actual number and sites of warehouses and allocation of clients to those warehouses. The resulting values are given in Tab. 1 as time component "Driving" in percent of the actual times according to the tour data.

Table 1. Times in percent of the actual tour data values

Time Component	Estimation Approach		Share (of Total)
	Apportionment	Regression	
Driving	76.3	93.4	42.9
Standing	100.0	100.0	41.1
Rest	100.0	100.0	16.0
Total	89.5	97.2	100.0

As it can be seen by the values of 76.3% and 93.4%, both schemes underestimate the actual driving time, the apportionment approach significantly more than the regression. The reason for this difference between the two approaches are time-window constraints. Besides vehicle capacity constraints, time-window constraints set by the clients are the only cause why villages possibly have to be called at more than once. While the regression approach implicitly takes into account these multiple trips to a village, the apportionment approach does not. The regression coefficients were determined on the basis of historical tour data. Since these tours included time windows, cost coefficients calculated with those regression coefficients implicitly model time

windows. For the apportionment approach, however, in order to cover multiple trips to a village, it would have been necessary to model this characteristic explicitly. This has not been done since the management wanted to reduce time window restrictions in the future. Therefore, cost coefficients calculated on the basis of the apportionment approach lead to a significantly lower value compared to the regression approach.

In addition to the values for the time component "Driving", Tab. 1 gives those for the components "Standing" and "Rest". Obviously, these values have to be identical with the tour data values (see Sec. 3.3, paragraph "Loading and Unloading"). Of further interest are the values for the three components given in the column "Share". Here, the percentage of the total time according to the tour data for each component is expressed. Surprisingly, only about 42.9% of the total working time of a truck driver is spent for driving; approximately the same amount of time is needed for unloading at the client sites and the still considerable rest of 16% for the pre- and post-processing of the tour at the facility site.

These relations are not specific for the actual distribution system, but obviously also hold more or less identically for the optimized structure. Thus, estimating the cost unit rate per distance unit by dividing the annual sum of the direct cost of the fleet (fuel, oil, taxes, insurance, and maintenance), proportional overhead costs, and the wages of the drivers by the total annual kilometers—as often done in logistical practice—would lead to heavily overestimated rates.

Only taking 42.9% of the driver's wages (see Tab. 1) into account for the calculation of the cost unit rate p^{d} per distance unit (as it was done for this study), the purely time driven component $p^{\mathrm{l,u}}$ makes up more than 72% of p^{d} (see Sec. 3.3, "Driving Costs"). Thereby, drivers' wages count for ca. 75% of the time driven cost component $p^{\mathrm{l,u}}$ (see Sec. 3.3,"Loading and Unloading Costs"). Thus, 42.9% of the driver's wages already make up about 54% of the cost unit rate p^{d} per distance unit. In consequence, when also those driver's wages caused by "Standing" and "Rest" had been considered, the cost unit rate per distance would be overestimated by more than 70%.

Cost Comparison. The costs of the actual distribution system were determined in three ways: first, from cost accounting, second, by weighting the total time values given in the tour data with the specific cost unit rates, and third, on the basis of the model. For the model values, the same solution was generated as for the time comparison, but in this case on the basis of parameters as discussed in Sec. 3.3.

Comparing the model values and the weighted tour values with each other obviously leads to the same results as the time comparison since the values only differ in constant factors: the cost unit rates. However, additional information is won by comparing those values with the costs given by the cost accounting. This way, the correctness of the cost unit rates could be verified. Therefore, in Tab. 2 the corresponding values are given in percent

of the cost accounting values. The total costs are decomposed into "Bulk Transport", "Fine Distribution", and "Operating Facilities": "Bulk Transport" are the costs of the transport from the plant to the facilities conducted in full-truckload mode by contractors; fine distribution costs are caused by the company's own vehicle fleet to serve the clients in routes; and the costs of operating facilities are the periodic fixed costs. Here, especially the value of 97.2% for the fine distribution in case of the weighted tour data shows that the cost unit rates were determined correctly.

Table 2. Costs in percent of the actual cost (cost accounting)

Cost Component	Tour Data	Estimation Approach		Share (of Total)
		Apportionment	Regression	
Bulk Transport	100.0	100.0	100.0	9.6
Fine Distribution	97.2	85.6	93.9	65.6
Operating Facilities	100.0	100.0	100.0	24.8
Total	98.1	90.2	96.0	100.0

As with the time values, also for the cost values the percental partitioning of the three components as given in the column "Share" of Tab. 2 is of further interest. Basis is the total system cost according to cost accounting. From these values it can be seen that the fine distribution makes up almost two thirds of the total cost, whereas the bulk transport and the operation of the facilities only cause approximately 10% and 25%, resp. These information are of special interest since almost the same relations hold for the optimized warehouse system.

4.2 Optimization

As "Optimization" three types of quantitative analyses were conducted: First, for the actual warehouse sites an optimized allocation of clients, second, the optimization of the structure "Warehouses", and third, calculations for the structure "Transshipment-Points".

Optimized Allocation for the Actual Warehouses. The first calculation was an optimization of the allocation of clients (OA). For the actual distribution system with only regional warehouses, the clients were reallocated on the basis of the cost coefficients calculated as discussed in Sec. 3.2 and Sec. 3.3. Since capacity constraints could be neglected, every client was allocated to the warehouse j with the smallest serving cost c_{ij}, formally:

$$j(i) = \arg\min_j c_{ij} \quad \forall i\,, \tag{24}$$

where $j(i)$ is the warehouse the client i is allocated at.

In the model, this resulted in the small decrease of the total system cost of 0.5% (0.6%). Split up into the cost components, the values for the optimized allocation were 98.9% (98.8%) for the bulk transport and for the fine distribution 99.3% (99.2%) of the actual allocation values. Obviously, the costs of operating warehouses were unchanged. (Note, for each pair of values, the first is the one for the problem data based on the apportionment approach and the second—in parenthesis—for the regression approach.)

Although quite small in percentage values, the decrease reported lead to the immediate action to reconsider the assignment of clients to the current warehouses. Under consideration of additional restrictions not modeled (e.g. shape of fixed routes), clients were reassigned according to the optimized allocation.

Structure "Warehouses". For the structure "Warehouses" (WH), a Simple Plant Location Problem was solved to optimality using a branch-and-bound algorithm based on subgradient optimization (see Klose 1993). The problem parameters f_j and c_{ij} were determined according to Sec. 3.2 and Sec. 3.3 as:

$$f_j = f^{\mathrm{W}} \quad \forall j \qquad \text{and} \qquad c_{ij} = c_{ij}^{\mathrm{PW}} + c_{ij}^{\mathrm{WC}} \quad \forall i, j\,. \tag{25}$$

On a Sun Sparc Ultra 2, 166 MHz, the calculation took 0:38 min (27:13 min). Compared to the actual distribution system, the resulting optimized warehouse structure shows a reduction of the number of facilities of more than a half and has overall cost of 91.8% (92.3%). For the cost components "Bulk Transport", "Fine Distribution", and "Operating Facilities", the cost values are 99.1% (102.5%), 106.8% (109.8%), and 55.5% (44.4%), resp., of the ones for the actual system (see Tab. 3).

Structure "Transshipment-Points". Solving the structure "Transshipment-Points" (TP) was carried out as discussed in Sec. 3.2 and Sec. 3.3: First, a Simple Plant Location Problem was solved to determine the transshipment-points with problem parameters

$$f_j = f^{\mathrm{T}} \quad \forall j \qquad \text{and} \qquad c_{ij} = c_{ij}^{\mathrm{TC}} \quad \forall i, j\,, \tag{26}$$

using the same algorithm as for the structure "Warehouses" in about the same time.

Then, after aggregating for the resulting transshipment-points the demands of all clients assigned to the specific transshipment-point, a second, much smaller SPLP was solved, this time with parameters

$$f_j = f^{\mathrm{W}} \quad \forall j \qquad \text{and} \qquad c_{ij} = c_{ij}^{\mathrm{PW}} + c_{ij}^{\mathrm{WT}} \quad \forall i, j\,. \tag{27}$$

The resulting structure was characterized by three times as many transshipment-points as warehouses. The overall costs of the system were 86.8%

(87.4%) of the cost of the actual system; the three components "Bulk Transport", "Fine Distribution", and "Operating Facilities" had, compared to the actual system, costs of 121.5% (121.5%), 99.9% (99.6%), and 43.8% (43.8%), resp. (see Tab. 3).

Comparison. Based on the model values for both estimation schemes the structure "Transshipment-Points" is superior to the "Warehouses" structure. The resulting costs indicate potential savings of about 13.2% (12.6%) of the total actual distribution cost for the structure "Transshipment-Points" and of 8.2% (7.7%) for "Warehouses" (see Tab. 3). When compared to the costs which are influenced by the locational decision (about 30% of the actual distribution costs were a constant for the study, namely the costs caused by the time components "Standing" and "Rest"; see Tab. 1 and 2), the potential savings even make up about 18.8% (18.0%) and 11.7% (11.0%), resp.

Table 3. Costs in percent of the actual cost (model valuation)

Cost Component	Apportionment			Regression		
	OA	WH	TP	OA	WH	TP
Bulk Transport	98.9	99.1	121.5	98.8	102.5	121.5
Fine Distribution	99.3	106.6	99.9	99.2	109.8	99.6
Operating Facilities	100.0	55.5	43.8	100.0	44.4	43.8
Total	99.5	91.8	86.8	99.4	92.3	87.4

Comparing the solutions with regard to the number of warehouses, for both estimation approaches the optimized structure "Warehouses" consists of less than half the number of warehouses of the actual distribution system. All of them are currently used warehouse sites. As to be expected, the structure "Transshipment-Points" has slightly less warehouses than the structure "Warehouses"; their sites are not identical with actual warehouse-sites.

A comparison with respect to the two different estimation approaches shows that both calculations result in almost identical solutions. For the structure "Warehouses", the regression approach leads to a slightly smaller number of warehouses than the apportionment approach, but the chosen sites were almost the same. The solutions of the "Transshipment-Points" structure were absolutely identical with respect to the number and sites of warehouses as well as transshipment-points.

4.3 Sensitivity Analysis

Solving facility location problems does not only require to determine optimal or almost optimal solutions, but also to perform sensitivity analyses. Due to the strategic dimension of facility location decisions, first, the unavoidable

uncertainty in the data has to be taken care of, and second, the consequences of minor changes to the proposed optimal solution have to be evaluated. In order to capture the effects of data changes, it is necessary to provide insight into the behaviour of the optimal solution depending on the problem parameters. The consequences of minor changes are important to evaluate since often organizational aspects prevent a straightforward implementation of the proposed optimal solution.

Structure "Warehouses". For the structure "Warehouses", this sensitivity analysis was conducted by means of parametric integer programming. Using a tangential approximation algorithm (see Klose and Stähly 1997a, 1997b), the "Cost Curve" (i.e. its convex hull) as well as "Sensitivity Intervals" for the fixed costs of a warehouse and the cost unit rate per minute driving were computed on a Sun Sparc Ultra 2, 166 MHz, in 838:27 min (773:38 min).

The "Cost Curve" describes the relation between the number of facilities and the total cost of the optimal solution for given number w of facilities. This curve is given in Fig. 5 as percentages of the optimal solution for the problem instance "Regression". As it can be seen by the shape of the graph, the curve is very flat, meaning that a small increase or reduction of the number of warehouses does not change the total cost of the optimal solution very much. (The calculations on the basis of the apportionment approach yield almost the same results.)

"Sensitivity Intervals" are sets of possible parameter values for which a solution with a given number of warehouses is optimal, under the restriction that all other parameters are unchanged (ceteris paribus). For the parameters fixed cost f^{W} of a warehouse and cost unit rate p^{d} per minute driving, these intervals are given in Fig. 6 percental to the values used (see Sec. 3.3) for the

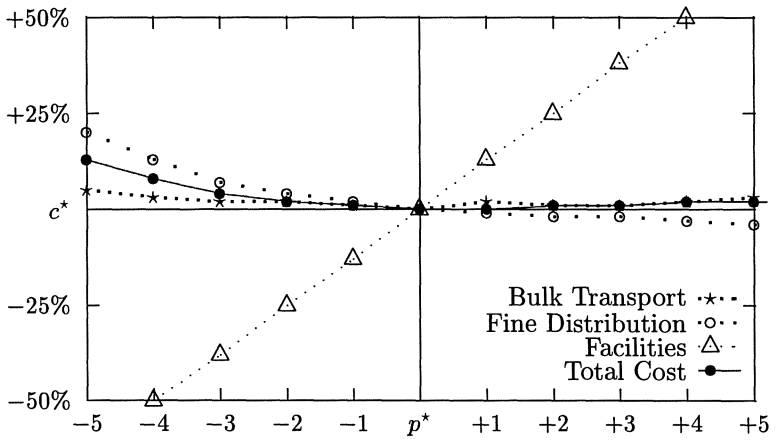
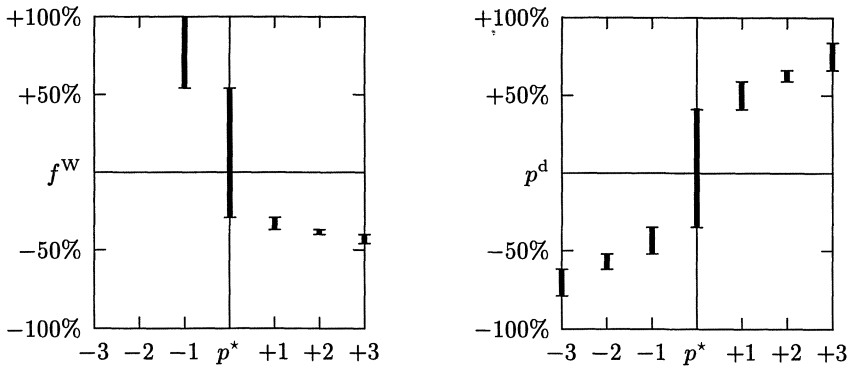


Fig. 5. Percental cost curve



(a) Warehouse fixed cost f^W (b) Distance cost unit rate p^d

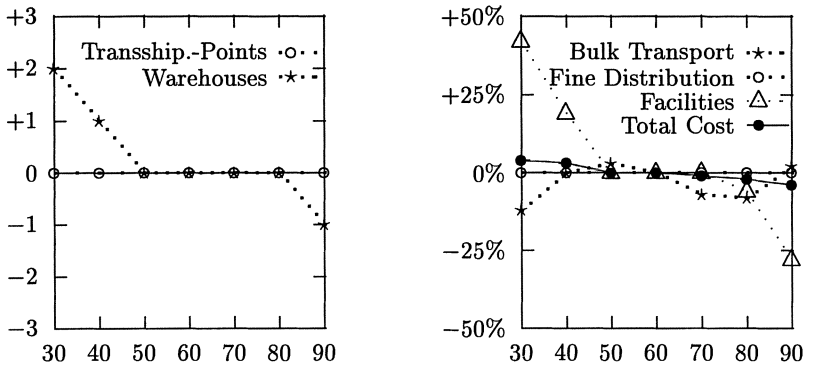
Fig. 6. Percental sensitivity intervals

regression approach. Fortunately, as expressed by Fig. 6, the results are very stable with respect to parameter changes. The used value for the fixed cost f^W can in- or decrease by up to 54% or 29%, resp., without changing the structure of the solution. For the cost unit rate p^d per minute driving, the respective values are 41% or 35%. Obviously, the intervals are mirror images of each other for the two parameters. (The apportionment approach yields almost the same results.)

Another very import information concerning the stability and robustness of the solution are the facility sites chosen, depending on the number w of facilities. In a so called "Site Using Table" the set of potential sites is listed, and for a range of values for the number w of facilities the sites used in the optimal solution for a specific w are marked. In practice, it is desirable that reducing or increasing the optimal $w^\star$ only leads to deleting or adding single facilities and not changing almost all used sites. In this sense, both problem instances were very stable and robust.

In addition to the information gained by means of parametric integer programming, further sensitivity analyses regarding the consequences of minor changes were carried out in the following way: Starting from the optimal solution with a PC-based software tool warehouses were added and/or deleted and clients were reallocated in order to track the changes in the costs. As to be expected for a combinatorial problem like the one at hand, the changes in total cost were very small. Depending on the extent of the change performed, the total cost increased not more than 1%. Thus, a variety of good solutions to choose from is available.

Structure "Transshipment-Points". Due to the two-step modeling approach of the structure "Transshipment-Points", a sensitivity analysis as for



(a) Number of facilities (b) System costs

Fig. 7. Sensitivities concerning time distances

the structure "Warehouses" was not possible. Only a separate analysis for each subproblem could be conducted. However, since the stepwise modeling neglects the relationship between the subproblems, it beard much less consequences.

Of more insight was to analyze the consequences of the time restriction concerning the allocation of transshipment-points to warehouses. On the basis of the same calculation for the first step the determination of the transshipment-points, the second step, the calculation of warehouses for given customers "Transshipment-Points", was performed with varying maximum time distances t_{ij}. In Fig. 7(a) the effect of the time restriction on the number of warehouses as well as transhipment-points is pictured for five values (40, 50, 60, 70, and 80 min) for the term t_{ij}. The resulting costs for the facilities, the bulk transport, and the fine distribution as well as their total are shown in Fig. 7(b).

Since all considered values of t_{ij} are based on the same calculation of the first step, the number of transshipment-points obviously remains unchanged. Fortunately, for the number of warehouses a small change in the time restriction (+10 or −10 min) has no effect, while larger changes show the expected result. With respect to costs, the ones caused by facilities vary according to the number of facilities, those related to the fine distribution obviously stay equal, and the bulk transport costs vary slightly more. Since the bulk transport accounts only for about 10% of the total costs (see Tab. 2), the changes in total costs are almost due to the ones in the number of facilities. Overall, the total costs decrease with a relaxation of the time restriction, as to be expected.

For the structure "Transshipment-Points", the same type of manual sensitivity analyses were conducted as for the structure "Warehouses", leading to almost the same results.

4.4 Further Analyses

Besides the optimizations, sensitivity calculations, and the manual sensitivity analyses discussed above, one further analysis was performed. For both structures, the calculations resulted in solutions with no warehouse or transshipment-point at the site of the single production plant. Although, on the basis of the defined cost structures, this result was absolutely plausible, from the viewpoint of the firm, it was undesirable. Therefore, additional optimization calculations were conducted, forcing at the site of the production plant a warehouse or a transshipment-points, resp. To this end, the fixed cost f_j of the site of the plant for the respective optimization problem was set to zero, performing the calculation with those changed fixed cost value and adding the real fixed costs afterwards. The resulting structures differed only in a single used site from the respective previous ones. Namely, the production plant site, forced to be used, replaced the closest site to it. With respect to costs, this resulted in an increase of not more than 0.5%.

5 Solution to the Problem

Although, at the beginning of the study the management was in favor of the transshipment-point structure, now they vote to stay with a two-stage warehouse system. Main argument is that for the structure "Transshipment-Points" the cost advantage is much less than expected, while at the same time requiring much higher investments in facilities. Recall, while for the structure "Warehouses" all sites used in the optimal solution are also currently in use, for the "Transshipment-Points"-structure the optimal warehouse sites are not identical with the current ones. In addition to the investments in new warehouses, when realizing the structure "Transshipment-Points" all transshipment-points have to be built new.

From the proposed optimal "Warehouse" structure a deviation seems to be made concerning the production plant site. For organizational and political reasons, the warehouse at the site of the single production plant will probably be kept. Also the optimal assignment will not be realized to full extent since some congruence between catchment areas and political areas is desired. However, the final decision is still outstanding.

6 Conclusions

In order to make the consequences of decisions clear in advance, it is necessary to model the interdependencies of the components involved in the decision problem. For strategic logistics planning, the consequences of interest are mainly costs, and the components which have to be taken into account are facilities and transport. Since the relationships between those two cost components are quite complex, it is inevitable to support strategic logistic

planning by quantitative methods. The system at hand has to be modeled and analyzed thoroughly, to be able to estimate the consequences of decisions on the overall cost.

The importance of such a quantitative modeling was also recognized by the study discussed. Using quantitative methods, the process of restructering the distribution system could be supported substantially. Besides the results of the optimization calculations, the cost insight gained during the modeling phase and the sensitivity information were of highest interest for the firm. First, the modeling of the problem required to take a deep look into the cost structure of the operations of the firm. It was necessary to allocate costs appropriately to end up with sensible cost unit rates. Second, the results of optimization calculations showed the maximum savings achievable by restructuring the system. Third, sensitivity information were of highest interest since due to organizational aspects the proposed optimal solution was difficult to implement. To know the cost consequences of minor changes to the optimal solution was therefore crucial.

From a methodical point of view the fact that both estimation schemes resulted in almost identical solutions was appreciated most. This, and the numerous sensitivity analyses validated their robustness.

Finally, the calculations show that by now real world problems as the one discussed can be solved in a reasonable amount of time. The use of quantitative methods for strategic logistic planning is therefore highly recommended.

Acknowledgement: This research has been supported by the "Swiss Federal Commission for Technology and Innovation (KTI)".

References

Fleischmann, B. (1979): Distributionsplanung. In Gaede, K.-W. / Pressmar, D. B. / Schneeweiß, C. Schuster, K.-P. / Seifert, O. editors, *Proceedings in Operations Research*, pages 293–308, Heidelberg. Deutsche Gesellschaft für Operations Research e.V. (DGOR), Physica-Verlag.

Klose, A. (1993): *Das kombinatorische p-Median-Modell und Erweiterungen zur Bestimmung optimaler Standorte.* Dissertation Nr. 1464, Hochschule St. Gallen, Dufourstrasse 50, CH-9000 St. Gallen.

Klose, A. / Stähly, P. (1997a): Parametric analysis of fixed costs in uncapacitated facility location. This volume.

Klose, A. / Stähly, P. (1997b): Sensitivity analysis in facility location applied to a depot location problem of a food producer. In Zimmermann et al. (1997).

Klose, A. / Tüshaus, U. (1995): Abstimmung zwischen strategischer und operationeller Stufe hinsichtlich der Konzipierung der logistischen Infrastruktur und Distributionsabwicklung. Arbeitsbericht, Institut für Unternehmensforschung (Operations Research), Hochschule St. Gallen, Bodanstrasse 6, CH–9000 St. Gallen.

Krarup, J. / Pruzan, P. M. (1983): The simple plant location problem: Survey and synthesis. *European Journal of Operational Research*, 12(1):36–81.

Tüshaus, U. / Wittmann, S. (1997): Locating depots for a food producer by solving uncapacitated facility location problems. In Zimmermann et al. (1997), pages 508–513.

Zimmermann, U. / Derigs, U. / Gaul, W. / Möhring, R. H. / Schuster, K.-P. (1997): editors. *Operations Research Proceedings 1996*, Berlin Heidelberg New York. Deutsche Gesellschaft für Operations Research e.V. (DGOR), Springer-Verlag. Selected Papers of the Symposium on Operations Research (SOR '96) Braunschweig, September 3 – 6, 1996.

Local Search Heuristics for the Design of Freight Carrier Networks

Helmut Wlcek

Universtität Augsburg, Lehrstuhl für Produktion und Logistik, Universitätsstraße 16, D-86135 Augsburg

Summary. The design problem of freight carrier networks is an actual problem in Germany due to the changing transport market. Such a network consist of depots, each with a pick-up and delivery area, which are linked by direct relations or via hubs. The design problem involves the decisions on the number and locations of depots and hubs. The operations in a freight carrier network and the design problem are described, a planning model is formulated and a heuristic algorithm based on a Local Search Scheme is presented. Computational tests and a real life application are reported.

1. Introduction

The framework conditions of the German transport market have strongly changed within the last years. These changes have been caused by politics on one hand and by the requirements of the customers on the other hand. Many of the small and medium sized freight carriers, which build the backbone of the transport industry in Germany, are endangered in their existence.

In the past, for reason of protection of the transport industry, the capacities of loading space and transport tariffs have been regulated by the legislator. At the beginning of 1994 these regulations expired. Furthermore, the entry barriers to the German transport market were removed for foreign freight carriers, which have competitive advantages due to their lower costs for personnel. As a consequence of the deregulation and the economically recessive situation of the German industry overcapacities of loading space arose, which caused the transport prices to fall for more than 35% (see DGM (1995)).

In the past the industry had special carriers for each region of Germany. Contrary to that, they prefer now to work closely together with just one carrier, who has to organize their transports all over Germany or even all over Europe. This requirement means an enormous organizational effort for small transport companies, which cannot afford to operate a distribution network covering all of Germany. Additionally the industry developed an increasing awareness of logistic costs in the last years. Therefore the carriers are forced to identify all possibilities to lower their costs and to improve their performance.

As a consequence, the small and medium sized transport companies for piece goods were starting to form cooperations. The carriers work closely together in carrying out the long distance transports and in delivering goods

to the receiving customers. By this organization, they achieve cost savings and are able to offer transport service all over Germany without operating the whole infrastructure on their on.

Although the German freight carriers are faced with various challenging problems, only few methods and planning software exist right now, which support them at the process of decision making.

In this paper, heuristic procedures are presented for the design of freight carrier networks, especially for piece good carriers or parcel services. The algorithms are implemented in an interactive decision support system called BOSS, which is already used in practice.

In the second section the operations in freight carrier networks are described and the planning problems of the different time horizons are derived. The third section gives a short overview of related literature. The model is described in section four and solution heuristics are presented in section five. Computational results, especially a comparison of the different methods and a case study conclude the paper.

2. Problem Description

In Germany, about 30 distribution networks for the transport of piece goods or parcels exist currently, which are operating all over Germany and parts of the European Community. Most of them provide a 24 hour service for transporting goods because of the requirements of their customers.

A direct transport from the sending to the receiving customer is economically and ecologically not reasonable due to the small weight and volume of the goods to be transported in these networks. Therefore the operations have to be organized in a special way: the transport has to be split up in several parts, where many shipments can be transported together. Usually, the transportation chain consists of three stages (see Fig. 2.1):

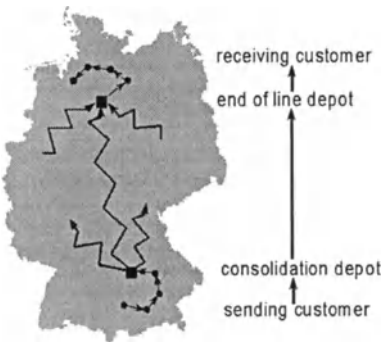


Fig. 2.1. Transport stages in freight carrier networks

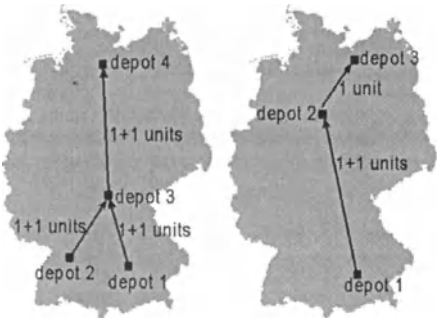


Fig. 2.2. Consolidation of loading units

at the first stage a transport from the sending customer to a consolidation depot is carried out; at the second stage, there is a transport from the consolidation terminal to an end of line terminal; the last transportation stage connects the end of line terminal with the receiving customer.

At the first transportation stage, several customers are served together in a vehicle tour. Their shipments are picked up and brought to the consolidation depot. In the depot, the goods are sorted with respect to their destinations and packed in loading units, which are used for the transport on the succeeding stage. To avoid the transshipment process, larger shipments are collected with the vehicles used for the succeeding transportation stage; at the depot the smaller shipments, heading for the same direction, are added in the same loading unit.

Even between the depots a direct transport is often not economically, too. The transports have to be consolidated, which can be done in several ways: the trucks used for the transport between the terminals can carry two loading units at a time. The transports to two depots, which are geographically close together, can be combined on one trip (see Fig. 2.2). If even one loading unit cannot be filled appropriately by the shipments to be transported on a link between depots, transports between several different depots can be consolidated in a hub facility (see Fig. 2.3).

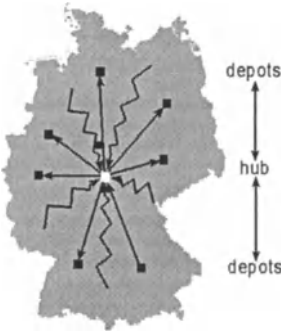


Fig. 2.3. Consolidation by hubs

Due to the additional transshipment process, this way of freight consolidation is less preferable than consolidation of loading units with respect to time as well as to costs, but the effect of bundling is higher. A further way of consolidation can be achieved by combining piece goods with shipments from other business units, which can be transported together on different parts of the transport chain, or to decrease the frequency of the transports, which increases the quantity to be transported on every trip.

The transport process from the end of line terminal to the receiving customer is the reversal of the transport process from the sender to the consolidation terminal. The pick-up and delivery tours are usually combined.

The strategic characteristics of freight networks are the number and locations of the depots, which allow to split the transport chain in the three parts, and the number and locations of the hubs, which are used to consolidate the transports between the depots.

At the tactical planning level, the transport routes between the customers have to be chosen. This problems includes the question, in which depots the shipments between customers are consolidated, as well as the issue of routing and consolidating transports between the depots. Due to organizational reasons customers are assigned to a unique depot; all shipments from and to a specific customer will be sent over this depot, even if this is not always the cheapest way. The areas, which include all customers of a specific depot, should be compact and not overlapping. A further step is the partition of the areas in segments to be served every day by the same vehicle. Other medium term planning problems are the decisions upon transport modes and frequencies. For German carriers these decisions are anticipated by the requirement of a 24 hour delivery service.

The operational planning problem is to exchange shipments between the different vehicle tour areas as far as it is necessary due to capacity restrictions. Additionally it has to be decided, whether to pick-up and deliver shipments directly or in tours (see Stumpf (1997)).

In the following, the strategic design of the freight traffic networks is considered in detail. The number and the locations of depots and hubs have to be determined so as to minimize costs of the transports without violating time restrictions. The basic dilemma of this planning problem is, that the costs for pick-up and delivery and the costs for the depot links diverge with an increasing number of depots: the transportation cost for pick-up and delivery decrease with more depots. On the other hand, the number of depot links to be served, increases quadratically with the number of depots, which causes higher costs for the intermediate transports. By consolidating these transports, especially by using a hub structure, this trend can be turned down to be linear in the best case.

3. Literature Review

The design of networks is a well known problem class in literature and considered in many different contexts. Domschke and Krispin (1997), Fleischmann (1997) and Florian (1986) give a detailed review.

Klincewicz (1991) and D. and S. Skorin-Kapov (1994) deal with p-hub-location problems. Out of a given set of fixed depots exactly p locations have to be chosen to fullfill hub functionality. Every depot is assigned to exactly one hub. All transports between depots are carried out with transshipments in the related hub locations. Transportation costs are linear and scaled by a factor smaller than 1 on links between hubs to model the economies of scale. Using tabu search techniques they solve problems with up to 52 depots and 10 hubs.

Aykin (1994,1995) considers a variant of this problem where additional routings are allowed. The transport may be carried out directly between two depots or pass just one or both related hubs. The potential hub locations are restricted to a subset of the depots. Transportation costs are also linear multiplied with different factors for links between two depots, a depot and a hub and two hubs. The maximal size of problems solved with his algorithms is 40 depots, 20 potential hub locations with 5 hubs to choose.

Crainic et al. (1993a, 1993b) investigate problems, where hub locations and the routing between depots has to be determined. Additionally, in case of imbalanced flows, there is a need for a transport of empty containers. The costs consist of fixed costs of hubs and linear transportation and transshipment costs. With various algorithms, problems with more than 200 depots and 50 potential hub locations are solved.

All mentioned articles are dealing with a subproblem of the design of freight carrier networks. Their basic assumption is, that the depot locations and the assignment of the customers to the depots are given.

The papers of Balakrishnan and Graves (1989), Khang and Fujiwara (1991), Larson et al. (1994) and Leung et al. (1990) present algorithms to solve network flow problems with non convex costs. They can be used in principle for the design of networks and deal with the assignment of customers to depots as well as with the choice of depot and hub locations. Balakrishnan and Graves consider piecewise linear costs, Khang and Fujiwara and Larson et al. fixed costs per arc; Leung et al. investigate common non convex cost functions. Their algorithms solve problems with up to 80 nodes and compute lower bounds. The gap is in most cases below 2%.

For the design of freight traffic networks much larger problems have to be solved, at least if the assignment of the customers has to be considered. Furthermore, additional constraints, as a maximum transportation time between to depots, have to be taken into account as well.

Daganzo (1991) investigates freight carrier networks with his method of Continuous Approximation (CA). Contrary to classical optimization models, which are fed with as many variables and restrictions as possible for getting

a proper image of reality, CA considers only the most important connections explicitly; as input for calculations aggregated data is used. By this technique, the optimization model stays tractable and can be solved to optimality. The solution of his models does not give an answer, where exactly to locate the depots and hubs, but it contains the values of key coefficients as the number of depots and hubs. Furthermore insights in the interactions of different design decisions and their effects on the related costs can be gained. Contrary to the other algorithms, CA works more precise with larger problem instances, because the faults made by using aggregated data are better equalized.
The method cannot be directly used to solve the considered problem, because it does not give the exact locations of the depots and hubs, but it could be very substantial for the evaluation of the quality of the proposed heuristic in this paper. This is important because neither other algorithms nor lower bounds for the considered problem are available.

4. Model

The design problem of freight carrier networks consist of determining the number and locations of depots and hubs. A discrete model is considered, where depot and hubs locations are chosen out of a set of potential locations.

Geographic information is needed as basic data in the model, especially the distances between the locations. Travel times can be calculated with additional data about the vehicles used for carrying out the transports. Furthermore, shipment data of transports between customers has to be available, which can be extracted from a past period of a transport company. In the transport business, daily as well as seasonal demand variations occur, both with a range of about 15-25% around the average. As this variations can not be easily modelled explicitly, the following procedure has to be applied: the set of data is divided in a subset for every day. For each subset, the costs of a network structure is computed separately. The total costs of a structure equal the sum of the costs of each day. This way of evaluation requires substantial computational effort. It can be reduced by analysing the subsets and grouping those with a similar demand rate together for a common evaluation.

In the model binary decision variables indicate, whether a depot or a hub is installed at specific locations.

The objective is to minimize fixed costs of depots and hubs and costs for transportation and handling. Though the decision is concerned just with determining the locations, costs for the operations in the network have to be taken into account, because they are strongly influenced by the network design.

The cost functions used for the evaluation of transport and handling processes should fit the real costs as good as possible. For this reason, standard freight tariffs, which have usually nothing in common with costs, are not suitable when optimizing the network design. A better suited cost function

can be derived from a process oriented interpretation of the transport: the main cost factors for the transport costs are the travel distance, the travel time and the number of vehicles. The weight of the vehicle has a neglectable influence. The costs per vehicle are calculated such, that both the distance and travel time are weighted with factors, modelling the costs arising due to the travel distance, as fuel consumption, and to the travel time, as the salary of the truck driver.

The vehicle tours for pick-up and delivery between depots and customer as well as round trips between the depots have to be determined for evaluating costs of the operations. For the transports between the depots, it is not sufficient to consider a maximum transport time for each shipment; a correct cost evaluation requires to compute schedules . If several shipments are consolidated on a certain link between depots, it has to be checked, if they can be transported on a single vehicle with respect to time aspects. As both of the routing problems are quite hard to solve themselves and as the cost evaluation has to be made quite often within the optimization procedure, the cost effects caused by carrying out transports in vehicle tours has to be modelled approximately by the objective function (see Sect. 5.3).

In the same way, the objective function models the cost effects of some restrictions, which have to be considered at the design of freight traffic networks. At the transportation stages between customers and depots, there is the organizational requirement of the single assignment of customers to depots. Furthermore the regions served from the different depots should be compact and non overlapping. To ensure that the planning result is operational implementable, a maximum distance between customers and depots has to be taken into account. Furthermore, very restrictive time conditions are given for the transports between the depots, which result from the need to guarantee a 24 hour transport time between customers.

In the following box, the indices, data and decision variables are described which are used in the remainder:

– Indices:
- depot and hub locations i
- customer locations k
– Data:
- transport quantity between customer locations m_{k_1,k_2}
- distance and travel time between locations $d(.,.), t(.,.)$
- maximum distance between customers and depots D
– Decision variables:
- $x_i^D, x_i^H = \begin{cases} 1 & \text{depot/hub is installed in location i} \\ 0 & \text{depot/hub is not installed in location i} \end{cases}$
– Additional variables
- $z_{i,k} = \begin{cases} 1 & \text{customer k is assigned to depot i} \\ 0 & \text{customer k is not assigned to depot i} \end{cases}$

With this notation, the model can be formulated as:

$$\begin{aligned} \min \quad & c(x^D, x^H) \\ s.T. \quad & x^D, x^H \in \{0,1\}^n \end{aligned}$$

This model for the design problem has a very simple form, since all restrictions are integrated in the objective function. The model is very well suited to be solved by use of local search algorithms.

5. Local Search Heuristics

In this section, several Local Search heuristics are presented, suitable to solve the problem described above. They are conceived to be integrated in an interactive decision support system (DSS), that means they are designed to produce results within a very limited computation time.

It is a reasonable question, whether it makes sense to deal with strategic planning problems with the help of an interactive DSS. These problems are characterized by decisions, which are binding over a long time horizon. But due to this long time horizon the uncertainty of the available data is very high – for shipment data as well as cost data. Because of this, to solve one problem instance based on a specific data set to optimality is not the right way to design a freight carrier network. Instead a lot of different data scenarios have to be investigated; the structures generated for special data sets should be analyzed with respect to their sensitivity in costs for alternative shipment and cost data. By this scenario technique the process of finding decisions for strategic problems can be supported by optimization methods. The advantage of exact solution procedures, which have usually tremendous computation times even for relatively small problems, are disappearing when scenario technique is used for decision making. With an interactive system, the user can check out many more different external influence factors and get a feeling for the relations between external parameters and their influence on the costs of different network structures. He can also weight different influence factors with their probability to become real.

One must be aware, that short computation times lead sometimes to a worse solution quality. The Local Search heuristics described in the following cannot provide lower bounds. Nevertheless, they have proved to perform very well for practical problems.

The basic idea of the algorithms is, that the number and the locations of depots and hubs has to be chosen in a way, that the costs for pick-up and delivery and the costs for transports between the depots are balanced.

The most common definition of neighbourhood in Local Search algorithms is, that two structures are neighboured, if they differ in exactly one position of their activation vectors. For real life problem sizes of 40 to 100 depot and hub locations this neighbourhood is very large. Additionally the operations

in freight carrier networks are very complex and due to the non convex cost functions neither efficient solution procedures nor optimality criteria exist. The evaluation of a given structure is a very complex problem itself. Therefore evaluating the whole neighbourhood is not possible. A single neighboured structure is chosen at a time, evaluated and perhaps accepted without considering alternatives.

The heuristics are based on the Deluge meta heuristics (see Dueck (1993)). The principal schema is the following:

1. Initialization:
1.1. Choice of an initial structure S_0;
1.2. Set the actual structure $S^a := S_0$;
Set the best known structure $S^b := S_0$;
1.3. Choice of the parameters:
Bound for acceptance of structures $P_A, (P_A \geq 1)$
Bound for search intensification around structures $P_I, (P_I \leq P_A)$
Step width for the adaptation of bounds P_S
Frequency for the adaptation of bounds P_H
1.4. Set counter for iterations without improvement $Z := 0$
2. Iteration:
2.1. Choice of an neighboured structure S_i from the actual structure S^a; **(see Sect. 5.2)**
2.2. Evaluation of the costs $K(S_i)$ of structure S_i; **(see Sect. 5.3)**
2.3. if $K(S_i) > P_A \cdot K(S^b)$ and $K(S_i) > K(S^a)$: Rejection of structure S_i; $\rightarrow$ 2.7.
2.4. If $K(S_i) < P_I \cdot K(S^b)$: Search intensification around structure S_i; **(see Sect. 5.4)**
2.5. If $K(S_i) < K(S^b)$: Update of the best known structure
2.5.1. $S^b := S^i$
2.5.2. $Z := 0$
2.6. Update of the actual structure $S^a := S_i; \rightarrow 3$.
2.7. $Z := Z + 1$;
2.8. If $Z = P_H$: Update of the parameters
2.8.1. $P_A := P_A - P_S$;
2.8.2. $Z := 0$;
3. Termination:
If $P_A \geq 1 :\rightarrow 2$.
Else stop.

The search intensification is an extension of the original procedure. It takes into account, that the generated solutions do not have to be local optimal; they can sometimes be improved easily. Therefore promising structures are considered in more detail and modified to local optimality.

In the remainder, the most important components of the heuristic procedure are described more detailed.

5.1 Choice of an Initial Structure

The choice of the depots is inspired by the heuristic knowledge, that the most favorable depot locations are those, with many customers sending or receiving large shipments nearby. Therefore, the base quantity for a depot is defined as the sum of the weight of all shipments from and to customers, for which the considered depot is the closest.

Starting with all depot locations marked as inactive, the depots are activated in order of descending base quantities. This is repeated until the sum of the base quantities of the activated depots reaches a certain percentage α of the total shipment quantity. Suitable values for α are in the range of 50 to 70%.

Afterwards all customers are examined, if they can be assigned to any activated depot. This must not be the case due to the restriction on the maximum distance between depots and customers. If a customer is found, who cannot be assigned, an additional depot has to be activated. From all depots, which could serve the specific customer the one is chosen, which has the maximum distance to any other activated depot. By this choice, the number of additional depots should be minimized, which are positioned in areas with sparse rise of shipments.

The initial choice of hub locations takes into account, that the largest effects of consolidation are realized with a single hub. By using more hubs, the transport distances can be reduced, but the effects of consolidation are decreasing. For this reason, initially only a single hub is chosen. Its geographical position should be as central as possible, that means the potential hub location is chosen, which minimizes the cumulated distance to all activated depots.

1. Initialization:
1.1. Choice of $\alpha \in (0,1)$
1.2. Deactivation of all depots: $x_i^d = 0$
1.3. Deactivation of all hubs: $x_i^h = 0$
1.4. Determination of base quantities of depots:
$B_i = \left\{ \sum_{k_1,k_2} m_{k_1,k_2} | (d(i,k_1) = \min_j d(j,k_1)) \vee \right.$
$\left. \vee (d(i,k_2) = \min_j d(j,k_2)) \right\}$
2 Choice of depots:
2.1. Sorting of B_i;
2.2. In order of descending B_i:
2.2.1. Activation of depot i: $x_i^d = 1$
2.2.2. If $\sum_i B_i \cdot x_i^d \geq \alpha \sum_i B_i$: $\rightarrow$ 2.2.
Else $\rightarrow$ 2.3.
2.3. Pass through all customers k
If $(x_i^d = 0) \wedge (d(i,k) > D)$ for all depots i:
$x_i^d = 1$ with $(d(i,k) \leq D) \wedge (\min_{j_1} d(i,j_1) = \max_{j_2} (\min_{j_1} d(j_1,j_2)))$
3. Choice of hub : $x_i^h = 1$ with $\sum_{j_1} d(i,j_1) = \min_{j_2} \sum_{j_1} d(j_1,j_2)$

5.2 Choice of a Neighboured Structure

The suitable choice of a neighboured structure and the evaluation of structures are the most important steps of the algorithm. The already mentioned magnitude of the neighbourhood and the fact, that only one of the neighbours can be chosen at a time make clear, that this choice has to be done thoroughly. During the runtime of the algorithm, only a very limited number of structures can be evaluated. Therefore, one has to try to find reasonable heuristic criteria to restrict the search on such structures, which might be possibly optimal. On the other hand a cut off any potentially good solution should be avoided.

Four different heuristics are presented. Two of them use more deterministic and two of them stochastic criteria for choosing the neighboured structure. Furthermore, two variants for the definition of the neighbourhood are considered.

For reasons of simplicity, in all heuristics the neighboured structures differ either in the depot configuration or in the hub configuration. Which configuration is modified is determined randomly with given probabilities. Furthermore, in all heuristics and in every iteration a central location is randomly chosen, to which all modifications of the configuration are related.

The first class of neighbourhood definition, the location based approach, is close to the classical definition; all criteria used to change the structure are related to the central location mentioned above. Contrary to this, in the second class, the area based approach, all depots lying in an area around the central location are considered simultaneously.

5.2.1 Location based Heuristic. Considering the stochastic variant of the location based heuristic, the activity state of the central location is simply switched. That means, a previously inactive depot or hub is turned active and inversely.

Using the deterministic procedure, for the central location, two other locations are considered, which are in close distance and have the inverse activation state. E.g. for an active depot, two inactive depots are chosen. Depending on the activity state of the central location there may be different actions executed to change the structure: in case of an active function of the central location, this function may be dropped or splited up to both of the chosen locations. Split means, that the function is deactivated in the central location, but instead activated in the two other ones. When the selected function of the central location is inactive, it can either be added or merged from the other two selected locations. Merge is the inverse function of split.

The choice of the step to take is based on two different criteria for depots and hubs. In case of variations of the depot structure the relevant key factor is the average usage rate of the transport vehicles going out and coming to the depot. Otherwise, in case of hub locations, the fill rate is not a suitable choice, because even badly used transports can yield an enormous bundling effect. A better choice is the number of depot shipments handled in the hub.

The values of the criteria determine the probability of performing the possible actions. The higher these values are, the more likely the split and add steps are executed.

1. Choice, if change of depot configuration (t = d) or hub configuration (t = h)
2. Choice of central location i
3. Choice of two neighboured locations to i: i_1 and i_2 with $(x_i^t = 1 - x_{i_1}^t) \wedge (x_i^t = 1 - x_{i_2}^t)$
4. If $x_i^t = 1$, then calculate criteria value of i $Z = Z(i)$, $(Z(i) \in (0,1))$ otherwise calculate criteria value of i_1 and i_2 : $Z = \frac{1}{2}(Z(i_1) + Z(i_2))$
5. If random number $R < Z$, $(R \in (0,1))$ Then $x_i^t = 1 - x_i^t, x_{i_1}^t = 1 - x_{i_1}^t$ and $x_{i_2}^t = 1 - x_{i_2}^t$ (split or merge) Else $x_i^t = 1 - x_i^t$ (add or drop)

5.2.2 Area based Heuristic. The second type of neighbourhood definition considers areas around the central location. Depots are part of the area, if their distance to the center location is less than its maximum radius for serving customers. The base quantity of an area is defined to be the sum of the base quantities of the depots in the area (see Sect. 5.1).

The stochastic and the deterministic procedure differ in the way to determine the number of active depots in this area. Whereas this number is defined randomly at the stochastic procedure, in the deterministic case the base quantity of the area and its concentration are used to calculate an appropriate number. The choice of the single locations is made such that each location is selected with a probability corresponding to its base quantity in relation to the base quantity of the area.

When choosing hub locations, the basic idea is, that a bundling of shipments is only possible, when several depots are served from a hub. That means, that the different hubs may not be too close together. For this reason it seems to be a reasonable rule, that in one area, there should be at most one hub. According to the number of loads with less weight than a loading unit, the necessity of a hub location in the relevant area is determined. If a hub should be placed this is done at the center of the area, that means at the central location. If already another hub exists, it is replaced by the new one.

1. Choice, if change of depot configuration (t = d) or hub configuration (t = h)
2. Choice of central location i
3. Initialization of the activity vector: if $d(i,j) < D : x_j^t = 0$
4. If t = d:

4.1. Determination of base quantities B_j of the depots

$$B_j^n = \begin{cases} 0 & \text{if } d(i,j) \geq D \\ B_i & \text{else} \end{cases}$$

4.2. Determination of the number of depots to choose:

$$A = \begin{cases} \text{random number } R & \text{at stochastic procedure} \\ \text{criteria value } Z(\sum_i B_i^n) & \text{at deterministic procedure} \end{cases}$$

4.3. Choice of the depots:

While $A > 0$:

Choose j with probability $P_j = \frac{B_j^n}{\sum_i B_i^n}$

Let $x_j^d = 1, B_i^n = 0, A = A - 1$

5. If t = h: $x_i^h = 1$

5.3 Cost Evaluation of Structures

As mentioned earlier, even the problem of evaluating the costs of the operations in a freight carrier network with a fixed structure is an extremely complex problem. For practical problem sizes, an exact solutions cannot be generated with the currently available soft- and hardware. Nevertheless this problem has to be solved very often in the Local Search heuristics. Furthermore, the maximum available computation time is limited by the requirement of being implemented in a interactive decision support system. Due to all these reasons, the evaluation process has to be done approximately, such that an acceptable solution quality can be achieved in short computation time. To evaluate the network structures, simplifying assumptions have to be made. The customers are assigned to the closest depot, which is activated. The costs for this assignment can be calculated efficiently. This rule of assignment is even local optimal in the sense, that improvements can only be realized, when a whole vehicle trips on a link between depots can be saved by the reassignment.

After assigning the customers to the depots, their shipment quantities can be aggregated in the depots which reduces the size of the considered network remarkably. The remaining problem is to estimate the costs for the transports between the depots.

For this reason, the quantities to be transported are divided in full truck loads, full loading units and the remaining part.

The full truck loads are transported directly from the sending depot to the receiving one in any case. The costs can be calculated easily; therefore, these transports have not to be considered in more detail.

The costs for full loading units are estimated without solving a routing problem. Instead, for every loading unit a fictitious tour is constructed, where other loading units may be included. The costs charged for the considered loading unit is the fair share of the costs of the combined tour.

The remaining loads are generally transported via hubs. For this reason from all active hubs the one is chosen as the central hub, which allows to handle the maximum number of shipments subject to the time restrictions. If two hubs can handle the same number, the one is chosen, which minimizes the distance to the activated depots. Additionally, for every depot a regional hub

is determined, which minimizes the cumulated distance to the other depots for all outgoing and incoming shipments. Shipments then may be carried out in two different ways: either they are transported from the sending depot via central hub to the receiving depot or via two regional hubs assigned to the depots. The second alternative is only chosen, when the transport cost is at least a certain factor cheaper, based on costs for whole vehicles. This constraint is chosen, because by splitting up the shipping quantities the bundling effect of hub transports decreases. This makes sense only if remarkable cost savings can be achieved. If both ways are not possible due to the time restrictions, they are treated like full loading units.

This organization of hub transports is implemented in a very similar manner in several parcel services in Germany.

1. Assignment of customers to depots and evaluation of costs
1.1. Initialization: let $z_{ik} = 0 \; \forall i, k$
1.2. $z_{ik} = 1$, if $d(i,k) = \min_j \{d(j,k) | x_j^d = 1\}$
2. Determination of the quantity to be transported between the pairs of depots
   $m_{ij}^D = \sum_{k_1,k_2} m_{k_1,k_2} \cdot z_{ik_1} \cdot z_{jk_2}$
3. Evaluation of costs of full vehicles
4. Evaluation of costs of full loading units (i,j):
   Choice of a loading unit on a depot link (i,j') or (i',j) with minimal distance to
   (i,i') or (j,j')
   Determination of the shortest transport route (e.g. i-i'-j)
   Calculation of the fair share of the costs of the combined transport
5. Evaluation of the remaining loads
5.1. Choice of a central hub h_z and regional hubs $h_r(i)$
5.2. Determination of the routing w_{ij} for all transports (i,j):

$$w_{ij} = \begin{cases} \text{i-}h_z\text{-j} & \text{if } k(i,h_z) + k(h_z,j) \leq (1-\alpha) \cdot (k(i,h_r(i)) + \\ & \qquad + k(h_r(i),h_r(j)) + k(h_r(j),j)) \\ \text{i-}h_r(i)\text{-}h_r(j)\text{-j} & \text{otherwise} \end{cases}$$

5.3. Evaluation of the transports

5.4 Search Intensification

The search intensification is used to investigate promising solutions in more detail. This is necessary due to the fact, that the generated solutions are usually not locally optimal. For this reason, structures, which have an objective value close to the best one known, are modified until they reach local optimality. The better the objective value is before this procedure, the more likely it is, that a new best solution can be found.

When intensifying the search around a structure, one entry of the activation vector of this structure is switched at a time. The thereby created new structure is evaluated and accepted exactly then, if the change improves

the objective value. Otherwise, the change is reversed. The different entries of the activation vector are considered in a cyclic manner. The procedure terminates, when no more improvements can be made in a whole cycle. This a classical 1-opt-procedure.

Although this intensification procedure converges within few cycles to a local optimal structure, the computation time required for this phase of the algorithm is relatively high. This is caused by the magnitude of the activation vector and the resulting large number of structures, which have to be generated and evaluated. When the parameter value P_I is chosen, which determines the bounds for starting the search intensification, the effects on computation time on one hand and the possibility of ignoring promising solutions have to be considered.

1. $K^{ori} := K^a$
2. For all depots and hubs i:
2.1. Change of the activity state $S^a : (x_i^d = 1 - x_i^d)$ or $(x_i^h = 1 - x_i^h)$
2.2. Evaluation of the costs $K(S^a)$ of the new structure
2.3. If $K(S^a) < K^a : K^a := K(S^a)$
Else reversal of the change of activity state: $(x_i^d = 1 - x_i^d)$ or $(x_i^h = 1 - x_i^h)$
3. If $K^a < K^{ori} :\rightarrow 1$.

6. Computational Results

In the following the heuristics described in the last section are compared. As already mentioned, most problems discussed in literature just deal with subproblems of the design of freight carrier networks or assume linear costs. Therefore, no other algorithms are available, which could produce comparable results.

In the tables, the stochastic and deterministic heuristics with location based and area based neighbourhood definition are compared.

Abr.	heuristic
H1	location based, stochastic heuristic
H2	location based, deterministic heuristic
H3	area based, stochastic heuristic
H4	area based, deterministic heuristic

Three parameter combinations are considered for the different search strategies:

Abr.	P_A	P_I	P_S	P_H
P1	1.03	1.005	0.01	250
P2	1.05	1.015	0.01	500
P3	1.07	1.02	0.01	1000

The evaluations are made on basis of three sets of shipment data. All of them are real data of different German cooperations of medium sized freight carriers. In the next table, some key parameters of the data are listed.

Abr.	# Shipments	# Days	# Customers	Total Weight in 1.000 t
D1	44.000	22	35.043	12.364
D2	85.000	20	41.305	23.520
D3	323.148	43	123.656	47.127

Furthermore four different location structures, i.e. sets of potential depot and hub locations, are investigated.

Abr.	# potential depots	# potential hubs
S1	60	5
S2	60	20
S3	100	10
S4	100	50

The location structures S2 and S3 contain S1, that means all potential depot and hub locations included in S1 are also included both in S2 and S3. Furthermore S4 contains all other structures.

In the following evaluations, the relative deviations in percent to the best known solution is given. Due to the lack of alternative algorithms, one of the four considered heuristics produced this best known solution (in one case, the best solution found on basis of structure S1 was better than the solutions with S3 and S4!). The three tables contain the results for the different data sets.

H	P	S1	S2	S3	S4
H1	P1	0.12	0.38	0.21	0.83
H2	P1	0.38	0.26	0.87	0.43
H3	P1	0.00	0.38	0.24	0.19
H4	P1	0.00	0.25	0.25	0.89
H1	P2	0.20	0.38	0.38	0.89
H2	P2	0.03	0.00	0.26	0.64
H3	P2	0.00	0.00	0.22	0.00
H4	P2	0.01	0.31	0.28	0.05
H1	P3	0.00	0.38	0.24	0.89
H2	P3	0.00	0.38	0.22	0.68
H3	P3	0.20	0.00	0.00	0.30
H4	P3	0.00	0.00	0.28	0.89

Data set D1

H	P	S1	S2	S3	S4
H1	P1	0.94	2.87	0.74	3.41
H2	P1	0.43	1.45	0.58	2.53
H3	P1	1.32	0.07	0.68	1.61
H4	P1	1.43	1.17	0.59	1.63
H1	P2	1.45	1.67	0.89	3.97
H2	P2	0.56	0.00	1.54	0.77
H3	P2	0.22	0.05	1.64	0.66
H4	P2	1.13	0.79	1.13	0.84
H1	P3	0.44	0.06	0.75	0.26
H2	P3	0.98	0.13	0.74	3.41
H3	P3	0.00	0.03	0.47	1.67
H4	P3	0.11	0.06	0.62	0.35

Data set D2

H	P	S1	S2	S3	S4
H1	P1	0.37	1.04	0.68	0.51
H2	P1	0.09	0.72	0.26	0.06
H3	P1	0.20	1.04	0.14	0.51
H4	P1	0.14	0.28	0.68	0.83
H1	P2	0.28	0.22	0.20	0.51
H2	P2	0.10	0.91	0.57	0.06
H3	P2	0.14	0.25	0.00	0.39
H4	P2	0.00	0.00	0.51	0.00
H1	P3	0.10	0.90	0.02	0.51
H2	P3	0.00	0.08	0.31	0.70
H3	P3	0.10	0.08	0.18	0.35
H4	P3	0.10	0.08	0.25	0.22

Data set D3

The objective function values of the solutions produced by the different heuristics are within a range of about 1% over the best known value in most cases. The reason is, that even after the initialization of the algorithm, which is the same for all the heuristics, the objective function is in average only 8.0% above the best known value; if a search intensification is started after the initialization, the average deviation is not more than 1.9% above the best known objective value.

The difference of the average deviation between stochastic and deterministic heuristics is not significant, but the average deviation of the area based heuristics is with 0.42% remarkably less than the deviation of the location based heuristics with 0.68%.

heuristic	average deviation	maximum deviation
H1	0.77	3.97
H2	0.58	3.41
H3	0.37	1.67
H4	0.47	1.62
after Initialization	8.00	10.00
after Initialization and Intensification	1.92	5.20

The number of iterations needed is for the location based heuristics remarkably higher than for the area based methods. The difference increases

with a higher number of potential locations and larger parameter values. The solution quality improves only slightly with higher parameter values. A marginal gain of solution quality is opposed to an enormous increase in computation time.

7. Practical Application

The described algorithms are implemented in the decision support system BOSS. Several case studies for carriers were done with this tool. One of them is described in the following.

A medium sized carrier operates two depots in Hamburg (HH) and Kiel (KI). From these two depots, shipments of customers in the region around the depots are picked up, consolidated and transported to 37 depots of cooperating carriers, which are spread all over Germany. The other direction of the transport process, where the cooperating partners collect goods in the regions of their depots and transport them to Hamburg and Kiel, was not considered.

In the scope of the study, the cost situation of this carrier was evaluated. A point of special interest was, whether it is more economical to operate the depot in Kiel or to serve all customers from the depot in Hamburg. Further questions concerned the number of receiving depots and the routing of the transport to these locations.

The first step of the application was to evaluate the current network structure and the operations. Starting from Kiel, there were eight depots delivered directly; the remaining transports were carried out via Hamburg. From Hamburg all depots were served directly.
The second scenario examined, if a cost reduction could be achieved by reducing the number of receiving depots and enlarging the destiny areas of the remaining terminals. BOSS had to choose, which depots were to be served. Concerning the routing, again, Kiel was allowed to consolidate freight in Hamburg. Starting from Hamburg, all transports had to go directly to the other depots.
In a further step, not only the number of depots was optimized, but also the routing. There were no more limitations how to carry out the transport, especially it was allowed to combine loading units on one vehicle.
These three scenarios were also evaluated with just a single sending depot in Hamburg.

For the evaluation, shipment data was available covering one month, containing 11.000 single shipments from customers of the region of Hamburg and Kiel to customers all over Germany. Cost data for a suitable evaluation of transports was available as well.

The results of the scenarios are shown in the following table. The cost values are divided in costs for pick-up and delivery on one hand and the cost

for serving the depot links on the other hand. The unit of the values is DM per day.

scenario	sending depots	pick-up & delivery	depot links	total costs
current structure	HH,KI	41.959	46.797	88.756
free no. of receiving depots	HH,KI	42.288	36.225	78.513
free optimization	HH,KI	42.150	32.619	74.769
current structure without KI	HH	43.963	38.646	82.609
free no. of receiving depots	HH	44.221	32.563	76.784
free optimization	HH	44.000	30.365	74.364

When the number of depots to be delivered is optimized, seven (scenarios with depot Kiel) respectively six (without Kiel) depots are closed, which are located in areas receiving only few shipments. As a consequence, the costs for the delivery are increased, because some shipments have to be transported over larger distances. The costs for the depot links can be reduced dramatically, because the transport of small amounts over long distances to the closed depots are not longer necessary. Without the prescription of rules for the routing, one more depot can be served in an economical manner. Due to this fact, the delivery costs are slightly reduced. By combining the links with only one loading unit to vehicle tours, a further saving of costs for the depot links can be achieved. Without the depot in Kiel, the costs for picking up the goods at the sending customers are remarkably higher, due to the longer distances. This additional costs are overcompensated by the reduction of cost for the depot links, which need not to be operated any more. The savings for the depot links get smaller, when no rules for the routing are prescribed.

As a result, the carrier decided not to close the depot Kiel, because the expected cost savings did not justify the political consequences within the company, which are caused by shutting down a location; instead he decided to reorganize the routings.

Fig. 7.1. Graphical interface of the planning tool BOSS

In Figure 11, the results produced by BOSS for the scenario 'free optimization' for sending depots in Hamburg and Kiel are depicted. The locations of depots and customers with their assignments (depicted by different colors) as well as the routing of the transports are shown.

8. Conclusion

The different heuristics for the design of freight carrier networks, which are described in this paper, produce results with objective values within a very small range. Although no lower bounds are available, the small difference between the results give the impression, that the solution values are close to the optimum.

Currently four large German carriers are using the planning system BOSS. According to their statements, the algorithms have proved to perform well in real life applications.

With respect to solution quality and number of iterations it can be observed, that deterministic methods for the choice of neighboured structures within the Local Search algorithms perform not better than stochastic criteria. With different neighbourhood definitions larger effects can be achieved. Algorithms, which allow larger changes of the structure behave in both respects more advantageous.

The variation of the objective function values produced with the different heuristics is smaller than the gap between the estimated and a more detailed evaluation of the operations in the network. For this reasons, the next steps of research will not focus on improving the heuristics, but to develop a fast and better method for the evaluation of the costs related with a freight traffic network.

References

Aykin, T. (1994): Lagrangean relaxation based approaches to capacitated hub-and-spoke network design problem. in: European Journal of Operational Research 79, 501 - 523

Aykin, T. (1995): The hub location and routing problem. in: European Journal of Operational Research 83, 200 - 219

Balakrishnan, A ./ Graves, P.C. (1989): A composite algorithm for a concave-cost network flow problem. in: Networks 19, 175 - 202

Crainic, T.G. / Delorme, L. (1993a): Dual-ascent procedures for multicommodity location- allocation problems with balancing requirements. in: Transportation Science 27/2, 90 - 101

Crainic, T.G. / Gendreau, M. / Soriano, P. / Toulouse, M. (1993b): A tabu search procedure for multicommodity location/allocation with balancing requirements. in: Annals of Operations Research 41, 359 -383

Daganzo, C.F. (1991): Logistic System Analysis. (Springer Verlag) Berlin

Deutsche Gesellschaft fur Mittelstandsberatung mbH (Ed.) (1995): Transport- und Speditionswesen: Positionen. Perspektiven. Strategien. Neu-Isenburg

Domschke, W., Krispin, G. (1997): Location and layout planning - a survey. in: OR Spektrum 19/3, 181-194

Dueck, G. (1993): New optimization heuristics: the great deluge algorithm and the record-to-record travel. in: Journal of Computational Physics 104, 86-92

Fleischmann, B. (1979): Distributionsplanung. in: Proceedings in Operations Research 8, 293 - 308

Fleischmann, B. (1997): Design of freight traffic networks. in: Stähly, P. et al. (Eds): Advances in distribution logistics. (Springer Verlag) Berlin

Florian, M. (1986): Nonlinear cost network models in transportation analysis. in: Mathematical Programming Study 26, 167 - 196

Khang, D.B. / Fujiwara, O. (1991): Approximate solutions of capacitated fixed-charge minimum cost network flow problems. in: Networks 21, 689 - 704

Klincewicz, J.G. (1991): Heuristics for the p-hub location problem. in: European Journal of Operational Research 53, 25 - 37

Larsson, T. / Migdalas, A. / Ronnqvist, M. (1994): A lagrangean heuristic for the capacitated concave minimum cost network flow problem. in: European Journal of Operational Research 78, 116 - 129

Leung, J.M.Y. / Magnanti, T.L. / Singhal, V. (1990): Routing in point-to-point delivery systems: formulations and solution heuristics. in: Transportation Science 24/4, 245 - 260

Skorin-Kapov, D. / Skorin-Kapov, J. (1994): On tabu search for the location of interacting hub facilities. in: European Journal of Operational Research 73, 502 - 509

Stumpf, P. (1997): Vehicle routing and scheduling for trunk haulage. in: Stähly, P. et al. (Eds): Advances in distribution logistics, (Springer Verlag) Berlin

Chapter 3

Transport Planning and Scheduling

The Minimization of the Logistic Costs on Sequences of Links with Given Shipping Frequencies

Luca Bertazzi and Maria Grazia Speranza

Dept. of Quantitative Methods, University of Brescia, Italy

Abstract. In this paper we consider a sequence of links in which a set of products has to be shipped from a common origin to a common destination through one or several intermediate nodes. Our aim is to determine periodic shipping strategies that minimize the sum of inventory and transportation costs when a set of shipping frequencies is given. From the theoretical point of view, the main question is to derive a formulation of the total inventory cost on the network. From the computational point of view, given that the simpler single link problem is known to be NP–hard, we present several heuristic algorithms based either on the decomposition of the sequence in links or on the dynamic programming techniques and compare them on a set of randomly generated problem instances.

Keywords. Sequences of links, Inventory and transportation costs, Shipping frequencies, Heuristic algorithms

Introduction

One of the most important problems in multi–products logistic networks is to determine the shipping frequencies that minimize the sum of inventory and transportation costs. This problem has been studied by Blumenfeld *et al.* (1985), Burns *et al.* (1985) and Anily and Federgruen (1990) using EOQ–based models. In these papers some important assumptions are made: Only one frequency, say f, can be selected to ship all the products and the corresponding time between shipments $t = 1/f$ can be any positive real number. A closed optimal solution is derived for the single link case and more complex networks, such as sequences of links and one origin–multiple destinations cases, are solved by a decomposition of the networks in links. The main drawback of this approach is that the obtained solution can be infeasible from a practical point of view, as discussed in Hall (1985), Maxwell and

Muckstadt (1985), Jackson *et al.* (1988), Muckstadt and Roundy (1993); for instance, it is very unrealistic to assume that products can be shipped every $\sqrt{2}$ time instants. In Speranza and Ukovich (1994b) a new approach to this problem has been proposed on the basis of the motivations given in Speranza and Ukovich (1992), where a Decision Support System for logistic managers has been presented. In this case the following main assumptions are made: A set of shipping frequencies is given and the time between shipments is a multiple of a chosen time unit. These assumptions take the real problem more carefully into account. In Speranza and Ukovich (1994b) a mixed integer linear programming problem has been formulated for the single link case and in Speranza and Ukovich (1996) some properties and a branch–and–bound algorithm have been presented; in particular, the problem has been proved to be NP–hard. In Bertazzi, Speranza and Ukovich (1995) an improved version of the branch–and–bound algorithm and several heuristic procedures with very good performance have been proposed. The approach with given frequencies has been also applied to the one origin–multiple destinations case in Bertazzi, Speranza and Ukovich (1997); finally, in Speranza and Ukovich (1994a) and in Bertazzi and Speranza (1996) a first analysis of the sequences of links case has been carried out.

In this paper our attention is focused on the sequences of links case; in particular, we consider multi–products sequences in which a unique decision–maker, given a set of shipping frequencies, has to determine for each link the percentage of each product to ship at each frequency and the number of vehicles to use at each frequency in order to minimize the sum of inventory and transportation costs. A typical application is when different types of vehicles are used on different links; an example is the case of overseas shipments in which products must be shipped first from the producer to a deposit by trucks or train, then from there to overseas by ship or plane and then to the destination by trucks or train again. Our aim is two fold: To analyze this problem from the theoretical point of view and to propose efficient computational methods. From the theoretical point of view, the main question is the formulation of the total inventory cost. From the computational point of view, given that the simpler single link case has been proved to be NP–hard, we propose several heuristic algorithms and compare them on the basis of a set of randomly generated problem instances.

The paper is organized as follows. In Section 1 the problem is described and formulated. In Section 2 properties of the inventory cost, which simplify the computation of the inventory cost itself, are presented. In Section 3 several heuristic algorithms, based either on the decomposition of the network or on the dynamic programming tecniques, are proposed; in Section 4 computational results obtained by comparing these algorithms on a set of randomly generated problem instances are shown and discussed; finally, in Section 5 some conclusions are drawn.

1 Problem description and formulation

In the sequences of links we are interested in, a set of products has to be shipped from a common origin to a common destination through one or several intermediate nodes. Each product is made available at the origin at a given constant rate and absorbed at the destination at the same given rate. Shipments on each link are performed on the basis of a set of frequencies. Each frequency is such that the corresponding time between shipments is integer. As a consequence, the transportation plan is periodic with time horizon equal to the minimum common multiple of all times between shipments. A set of vehicles with equal transportation capacity and cost is associated to each frequency; each vehicle in this set can be used only in the time instants (shipping times) which are multiples of the corresponding time between shipments and, if it is used, the corresponding transportation cost per journey is charged independently of the quantity loaded on. The inventory cost is charged for each product on the basis of the idle time in each node; the total idle time depends on the frequencies at which each product is shipped. The selected shipping frequencies also determine the level of the starting inventory to make available at time 0 in order to avoid stock–out during the time horizon.

For this situation, our aim is to determine a periodic shipping strategy that minimizes the sum of the transportation and inventory costs. More specifically, we have to decide for each link:

1. The quantities of each product to be shipped at each frequency;
2. the number of vehicles to use at each frequency;
3. the starting inventory of each product to make available at time 0 in each node;

on the basis of the following constraints:

1. The total quantity of each product made available at the origin must be shipped at the destination during the time horizon;
2. the number of vehicles used at each frequency must be sufficient to load the corresponding shipping quantity;
3. no stock–out can occur during the time horizon.

More formally, this problem can be described as follows.

Let $M = \{1, 2, \ldots, m\}$, $m \geq 3$, be the set of nodes in the sequence and $L = \{(l, l+1)\colon\ l = 1, 2, \ldots, m-1\}$, be the set of links; for the sake of simplicity, we denote the link $(l, l+1)$ by l, that is the starting node of the link. We will use $l \in M$ to identify a node and $l \in L$ to identify a link.

A set I of products has to be shipped from 1 to m through $m-2$ intermediate nodes. Each product $i \in I$ is made available in 1 and absorbed

in m in a continuous way (*continuity assumption*) at a given constant rate (*steady–state assumption*); for each product $i \in I$ the supply and demand rates are equal (*equilibrium assumption*) and denoted by q_i. Each product $i \in I$ has an inventory cost h_i and a volume v_i.

A set F^l of shipping frequencies f_j^l, $j \in J^l$, is given for each link $l \in L$; without loss of generality, we assume that each frequency f_j^l is such that the corresponding time between shipments $t_j^l = 1/f_j^l$ is integer (*given discrete frequencies assumption*). Each product $i \in I$ can be partially shipped on each link $l \in L$ with each of the frequencies $f_j^l \in F^l$ (*multiple frequencies assumption*).

Let $H = mcm\{t_j^l, j \in J^l, l \in L\}$ be the time horizon on the sequence and $T = \{0, 1, \ldots, H-1\}$ be the set of time instants in H. If frequency f_j^l is selected, the corresponding number of shipments in H is H/t_j^l. In this paper, we assume that the first shipment for all selected frequencies is made in 0 (*phasing all frequencies in 0 assumption*). This assumption corresponds to the situations in which the decision–maker has no degrees of freedom in the phasing of the frequencies; an example is when shipments are performed by carriers or public transportation systems. Then the following shipments will be made at the time instants which are multiples of the period t_j^l, $\forall j$. For each frequency f_j^l, the quantity of each product $i \in I$ shipped at any shipping time is constant and proportional to t_j^l (*fixed shipping quantity assumption*).

On each link $l \in L$ shipments can be performed at each frequency f_j^l by any number of vehicles with transportation capacity r_j^l and transportation cost c_j^l. The transportation cost is charged for each journey independently of the quantity loaded on the vehicles. The total transportation time is supposed to be constant and not greater than the minimum time between shipments; situations with greater transportation time can be carried on by considering that they imply only a different formulation of the stock–out constraints. Finally, we assume that shipments from any node are performed in the time instants $t \in T$ and that in the intermediate nodes a possible shipment at time t takes place after the arrival of the vehicles from the previous node of the sequence.

Our aim is to derive a periodic shipping strategy, with period H, that minimizes the sum of the inventory and transportation costs on the sequence. The periodic shipping strategy we are interested in is defined by the following elements:

1. The percentage x_{ij}^l of product $i \in I$ to ship on link $l \in L$ with frequency f_j^l, $j \in J^l$;
2. the number of vehicles y_j^l to use on link $l \in L$ with frequency f_j^l, $j \in J^l$;
3. the starting inventory d_i^l of product $i \in I$ to make available at time 0 at node $l \in M$ in order to avoid stock–out during the time horizon H.

The above variables must satisfy the following constraints:

1. *Demand constraints*: For each link $l \in L$ and for each product $i \in I$, the total quantity made available at node l must be shipped to the node $l+1$ during the time horizon H. These constraints can be expressed as:

$$\sum_{j \in J^l} x_{ij}^l = 1 \qquad i \in I \quad l \in L. \tag{1}$$

2. *Capacity constraints*: For each link $l \in L$, the number of vehicles y_j^l used with each frequency f_j^l must be sufficient to load the corresponding shipping quantity that is, given the fixed shipping quantity assumption, $\sum_{i \in I} t_j^l q_i x_{ij}^l$. These constraints can be expressed as:

$$t_j^l \sum_{i \in I} v_i q_i x_{ij}^l \leq r_j^l y_j^l \qquad j \in J^l \quad l \in L. \tag{2}$$

3. *Stock–out constraints*: For each node $l \in M$ and for each product $i \in I$, the inventory in each time instant must be non negative.

   In particular, at the origin 1, each product $i \in I$ is made available in a continuous way and possible shipments are performed in t; therefore, the stock–out constraints are satisfied if the level of the inventory in each time instant t, immediately after a shipment is performed, is non-negative. The stock–out constraints can be formulated as:

$$d_i^1 + q_i t - \sum_{j \in J^1} (1 + \lfloor t/t_j^1 \rfloor) t_j^1 q_i x_{ij}^1 \geq 0 \qquad t \in T \quad i \in I \tag{3}$$

   as the summation gives the total quantity of product i which is shipped from node 1 up to time t.

   At each intermediate node $l \in \{2, 3, \ldots, m-1\}$, each product arrives from the previous node with the frequencies f_j^{l-1} and is shipped to the following node with the frequencies f_j^l; therefore, the stock–out constraints at each time instant t immediately after a possible shipment can be expressed as:

$$d_i^l + \sum_{j \in J^{l-1}} (1 + \lfloor t/t_j^{l-1} \rfloor) t_j^{l-1} q_i x_{ij}^{l-1} - \sum_{j \in J^l} (1 + \lfloor t/t_j^l \rfloor) t_j^l q_i x_{ij}^l \geq 0 \quad t \in T, i \in I \tag{4}$$

   Finally, at the destination m, each product $i \in I$ is absorbed in a continuous way; therefore the stock–out constraints are satisfied if the level of the inventory is non-negative at each time instant t, immediately

before the arrival of products from the previous node. Then, the stock–out constraints can be formulated as:

$$d_i^m + \sum_{j \in J^{m-1}} (1 + \lfloor (t-1)/t_j^{m-1} \rfloor) t_j^{m-1} q_i x_{ij}^{m-1} - q_i t \geq 0 \qquad t \in T, i \in I \quad (5)$$

where the summation does not appear when $t = 0$.

We can now formulate the optimization problem for the sequence, which is referred to as Problem $\mathcal{F}$.

Let S be the set of feasible solutions (x, y, d) of Problem $\mathcal{F}$, that is the set of values for the variables which satisfy the constraints (1)-(5). The variables $x \in [0,1]^{|M||I||F|}$ represent the percentages at which each product is shipped at each frequency on each link, $y \in Z^{|M||F|}$ the number of vehicles used with each frequency on each link and $d \in \Re_+^{|M||I|}$ the starting inventory at time 0 of each product on each node. Problem $\mathcal{F}$ can be expressed as follows.

Problem $\mathcal{F}$

$$\min_{(x,y,d) \in S} z'(x,d) + z''(y)$$

where $z'(x,d)$ is a function expressing the total inventory cost during the time horizon H and $z''(y)$ is a function expressing the total transportation cost in H.

Let Q_{it}^l be the level of the inventory of product $i \in I$ at node $l \in M$ at the end of the time interval $[t, t+1)$, $t \in T$; then the inventory cost $z'(x,d)$ has the following form:

$$z'(x,d) = \sum_{i \in I} \sum_{l \in M} \sum_{t \in T} h_i Q_{it}^l.$$

The quantities Q_{it}^l obviously depend on the variables x and d and can be defined through the following recursive formula in order to completely define a mathematical programming model:

$$Q_{it}^l = Q_{it-1}^l + e_{it}^l - u_{it}^l$$

where $Q_{i0}^l = d_i^l$, e_{it}^l is the quantity of product i made available or entering in l during the time interval $[t, t+1)$ and u_{it}^l is the quantity absorbed or shipped from node l during the time interval $[t, t+1)$. The quantities e_{it}^l and u_{it}^l can be expressed as functions of x and d and take different expressions when l is the origin, an intermediate node or the destination. However, given the complexity of the mathematical programming problem and our consequent interest to investigate heuristics, we are more interested in a simple way to compute the function $z'(x,d)$, given the variables x and d. This issue will be discussed in the following section.

The transportation cost $z''(y)$ can be expressed as:

$$z''(y) = \sum_{l \in L} \sum_{j \in J^l} c_j^l H / t_j^l y_j^l.$$

As $z''(y)$ is a function increasing in the variables y, if feasible values are given for the variables x, then the optimum number of vehicles y^* is obtained from the capacity constraints (2) as:

$$y_j^{l*} = \lceil t_j^l / r_j^l \sum_{i \in I} v_i q_i x_{ij}^l \rceil. \tag{6}$$

2 The inventory cost

We first show that the total inventory cost of each product $i \in I$ on the sequence is constant over time and then how to compute it on the basis of the starting inventory d only.

Theorem 1 $\sum_{l \in M} Q_{it}^l = \sum_{l \in M} d_i^l$, $\forall t \in T$.

Proof For each time instant $t \in T$ we have

$$\sum_{l \in M} Q_{it}^l = \sum_{l \in M} (d_i^l + \sum_{k=0}^{t} e_{ik}^l - \sum_{k=0}^{t} u_{ik}^l).$$

Since $e_{it}^1 = u_{it}^m$ and $e_{it}^{l+1} = u_{it}^l$, $l = 1, 2, \ldots, m-1$, then

$$\sum_{l \in M} Q_{it}^l = \sum_{l \in M} d_i^l.$$

□

Therefore, the total inventory cost takes the form:

$$z'(x, d) = \sum_{i \in I} \sum_{l \in M} h_i d_i^l H,$$

where the variables d must obviously satisfy the stock-out constraints. As $z'(x, d)$ is a function increasing in the variables d, if feasible values are given for the variables x, then the optimal values d^* of the variables d can be computed as the minimum values that satisfy the stock–out constraints. Therefore, the set of solutions to explore in order to obtain the minimum cost of Problem $\mathcal{F}$ is $X_S = \{x : (x, y^*, d^*) \in S\}$; therefore, Problem $\mathcal{F}$ can be reduced to the following equivalent problem, referred to as Problem $\mathcal{F}'$, in which the optimization is taken only on the percentages x.

Problem $\mathcal{F}'$

$$\min_{x \in X_S} z'(x, d^*) + z''(y^*).$$

Computation of the minimum starting inventory

In this section we show some results which make the computation of d^* simpler, once the values for the variables x are given. In particular, we derive for the origin and the destination a closed formulation of the minimum starting inventory.

For the origin 1 we can state the following proposition.

Proposition 1 *The minimum starting inventory at the origin* 1 *at time* 0 *which satisfies the stock–out constraints (3) is*

$$d_i^{1*} = \sum_{j \in J^1} t_j^1 q_i x_{ij}^1 \qquad i \in I.$$

Proof The stock-out constraints (3) with starting inventory d_i^{1*} become:

$$t - \sum_{j \in J^1} \lfloor t/t_j^1 \rfloor t_j^1 x_{ij}^1 \geq 0.$$

We verify that the constraints are satisfied, since

$$t - \sum_{j \in J^1} \lfloor t/t_j^1 \rfloor t_j^1 x_{ij}^1 \geq t - \sum_{j \in J^1} t/t_j^1 t_j^1 x_{ij}^1 = 0.$$

Moreover, these constraints are satisfied at equality in 0. □

For the destination m we can state the following proposition.

Proposition 2 *The minimum starting inventory* d_i^{m*}, $i \in I$, *in* m *at time* 0 *which satisfies the stock–out constraints (5) is zero.*

Proof The stock–out constraints (5) with $d_i^{m*} = 0$ are obiouvsly satisfied in 0. For $t > 0$, given that $\sum_{j \in J^{m-1}} x_{ij}^{m-1} = 1$, they can be expressed as follows:

$$\sum_{j \in J^{m-1}} (t_j^{m-1} + \lfloor (t-1)/t_j^{m-1} \rfloor t_j^{m-1} - t) x_{ij}^{m-1} \geq 0.$$

These constraints are satisfied if

$$t_j^{m-1} + \lfloor (t-1)/t_j^{m-1} \rfloor t_j^{m-1} - t \geq 0 \qquad j \in J^{m-1}.$$

These inequalities are always satisfied during the time horizon H because the left hand side assumes periodically, with period t_j^{m-1}, the following values:

$$t_j^{m-1} - k \qquad k = 1, 2, \ldots, t_j^{m-1}.$$

□

In conclusion, given the variables x, the optimal values of y can be directly computed as in (6). The optimal values of d_i^1 can be computed as shown in Proposition 1 and the optimal values of d_i^m are 0 (see Proposition 2). The optimal values of d_i^l, $l \neq 0, m$, can be obtained by evaluating the stock-out constraints (4) and taking the minimum value of d_i^l which satisfies all the constraints.

3 Algorithms

Given that the simpler single link case has been proved to be NP–hard in Speranza and Ukovich (1996), a heuristic approach is required for the sequences of links. We present four classes of heuristic algorithms, based either on the decomposition of the sequence in links or on the solution of a simpler problem through dynamic programming techniques.

In each of these algorithms the approximation is in the selection of the percentages at which the products are shipped. In other words, once the approximate percentages $\bar{x}$ are heuristically computed, the solution we consider is the following:

$$(\bar{x}, y^*(\bar{x}), d^*(\bar{x})) \tag{7}$$

and the corresponding total cost is

$$z'(\bar{x}, d^*(\bar{x})) + z''(y^*(\bar{x})), \tag{8}$$

where the inventory cost is computed by means of the results presented in Section 2. The approximate percentages are obtained in the first three classes of heuristic by optimizing separately each link of the sequence and in the fourth one by assuming that only one frequency can be selected on each link.

In the following we assume that all vehicles on each link $l \in L$ have the same transportation capacity, independently of the frequency, that is $r_j^l = r^l$, $\forall j \in J^l$, and therefore the same transportation cost c^l. Moreover, we assume, without loss of generality, that the frequencies on each link are ordered in the decreasing order, that is $f_1^l > f_2^l > \ldots > f_{|J^l|}^l$, $\forall l \in L$.

Heuristic *Dec*

The first class of heuristic procedures we propose, referred to as *Dec*, is based upon the idea to decompose the sequence in links and to optimize each link separately. In other words, the approximate percentages $\bar{x}$ are obtained by solving, for each link $l \in L$ separately, the following Problem $\mathcal{P}^l$, proposed by Speranza and Ukovich (1994b) for the single link case.

Problem $\mathcal{P}^l$

$$\min \sum_{i \in I} \sum_{j \in J^l} h_i t_j^l q_i H x_{ij}^l + \sum_{j \in J^l} c^l H / t_j^l y_j^l \tag{9}$$

$$t_j^l \sum_{i \in I} v_i q_i x_{ij}^l \leq r^l y_j^l \qquad j \in J^l \tag{10}$$

$$\sum_{j \in J^l} x_{ij}^l = 1 \qquad i \in I \tag{11}$$

$$0 \leq x_{ij}^l \leq 1 \qquad i \in I, j \in J^l \tag{12}$$

$$y_j^l \geq 0 \qquad \text{integer} \qquad j \in J^l. \tag{13}$$

The objective function (9) expresses the minimization of the sum of the inventory cost and the transportation cost on the time horizon H. Constraints (10) are the capacity constraints, which state that the number of vehicles y_j^l must be sufficient to load all the products assigned to frequency f_j^l. Constraints (11) are the demand constraints, which impose that all the quantity of product i made available in l must be shipped to $l+1$ at some of the given frequencies.

Problem $\mathcal{P}^l$ is solved by using the heuristic *Best* proposed by Bertazzi, Speranza and Ukovich (1995), for which it was shown an average error with respect to the optimal solution less than 0.4% and a computational time of a few seconds on instances with up to 10,000 products and 15 frequencies.

Given $\bar{x}$, the corresponding total cost is computed as described in (8).

Stationary frequencies heuristics

The second class of heuristic algorithms, referred to as *Stationary frequencies heuristics*, is based upon the idea to apply the same percentages $\bar{x}$ to all links in the network; the rationale is that, in this case, the inventory cost in the intermediate nodes is zero. Obviously, these heuristics can be used only if $F^l = F$, $\forall l \in L$.

These heuristics can be described as follows:

1. Identify a representative link l^* and compute the approximate percentages by solving Problem $\mathcal{P}^{l^*}$ using heuristic *Best*;
2. apply the approximate percentages to all links $l \in L$;
3. compute the total cost as described in (8).

The selected link l^* can belong to L or can be a pseudo–link which captures the important characteristics of the sequence. We propose two different heuristics in this class: The *SF-r* heuristic, in which the link l^* is randomly selected in L and the *SF-a* heuristic, in which the link l^* is a pseudo–link with transportation cost $c^{l^*} = 1/(m-1)\sum_{l \in L} c^l$.

EOQ–based heuristics

The third class of heuristics, referred to as *EOQ–based heuristics*, is composed of two discretized versions of the EOQ–based algorithms by Blumenfeld *et al.* (1985). In these heuristics the main assumption is that only one frequency can be selected for each link and that the corresponding time between shipments can assume any positive real value.

These heuristics can be described as follows:

1. For each link $l \in L$:

   (a) determine the unique optimal continuous time between shipments

   $$\bar{t}^l = \min\left(\sqrt{\frac{c^l}{\sum_{i\in I} h_i q_i}}, \frac{r^l}{\sum_{i\in I} v_i q_i}\right).$$

   In this formula the first quantity over which the minimum is taken is the classical "Wilson's formula" for the EOQ, while the second one takes into account the finite capacity of the vehicles;

   (b) discretize $\bar{t}^l$ in order to obtain a feasible solution for Problem $\mathcal{F}$;

2. on the basis of the times between shipments selected in Step 1, determine the approximate percentages $\bar{x}$ for Problem $\mathcal{F}$ and then compute the total cost as in (8).

We propose two different heuristics in this class. In the first one, referred to as EOQ-s, for each link $l \in L$ the optimal continuous time between shipments $\bar{t}^l$ is rounded off to $\hat{t}^l$ equal to the nearest period t^l_j, $j \in J^l$, smaller than $\bar{t}^l$; then all the products are shipped at the corresponding frequency $\hat{f}^l = 1/\hat{t}^l$. If $\bar{t}^l < t^l_1$, then t^l_1 is selected.

In the second EOQ–based heuristic, referred to as EOQ-l, for each link $l \in L$ the optimal continuous time between shipments $\bar{t}^l$ is rounded off to $\hat{t}^l$ equal to the nearest period t^l_j, $j \in J^l$, larger than $\bar{t}^l$; then the maximum number of full load vehicles $\lfloor \hat{t}^l \sum_{i\in I} v_i q_i \rfloor$ is shipped at the corresponding frequency $\hat{f}^l = 1/\hat{t}^l$ and the remaining volume is shipped at the lowest available frequency $f^l_{|J^l|}$. The products are loaded on these vehicles on the basis of the non–increasing ratio h_i/v_i.

Dynamic programming–based heuristic

The fourth heuristic algorithm we present is based upon the following idea: To solve optimally a simpler problem. In particular, the problem we consider is based upon the assumption that only one shipping frequency can be selected on each link for all the products $i \in I$ and the tecnique we use in order to obtain an optimal solution for this problem is the deterministic dynamic programming algorithm.

The motivations for this heuristic are the following: Firstly, the problem with just one frequency for each link is very easy to implement from an operational point of view; secondly, given that the same problem is approximately solved by the heuristic EOQ–s, doing so we can evaluate the performance of this type of EOQ–based algorithm; finally, we can evaluate if the assumption of multiple frequencies for each product is critical in order to obtain good solutions.

In our problem we have $m-1$ stages where each stage l corresponds to the link $l \in L$. At each stage l, the state space S_l is the set of shipping frequencies F^{l-1} on the previous link $l-1$; therefore, the state x_l is the frequency f^{l-1*} selected on the previous link. The control space C_l is the set of shipping frequencies F^l on the link l; therefore, the control u_l is one of the frequencies that can be selected on l.

Finally, the cost $g_l(x_l, u_l)$ is the sum of inventory cost on node l and transportation cost on link l. As described in Bertsekas (1995), the corresponding deterministic dynamic programming algorithm takes the following form:

$$J_m(x_m) = \text{inventory cost at the destination } m \qquad x_m \in S_m$$

$$J_l(x_l) = \min_{u_l \in C_l} [g_l(x_l, u_l) + J_{l+1}(x_{l+1})] \quad x_l \in S_l \quad l = 1, \ldots, m-1.$$

Once the optimal percentages x^{DP} for the problem with single frequency on each link are computed, we set the approximate percentages $\bar{x}$ of Problem $\mathcal{F}$ equal to x^{DP} and then we compute the total cost as in (8).

4 Computational results

The heuristic algorithms proposed in the previous section have been implemented in Fortran on a personal computer with an Intel 80486 processor and tested on a large set of randomly generated problem instances.

In particular, 36 specific situations have been tested, corresponding to different number of links in the sequence (2, 3, 4, 5), different number of products (10, 50, 100) and different range of the inventory cost per unit time ([1, 3], [3, 6], [1, 6]).

For each specific situation, 5 problem instances have been generated with the following data:

- Set of shipping frequencies $F^l = \{1, 1/2, 1/4, 1/5, 1/10\}$, $\forall l \in L$;
- Transportation capacity $r^l = 1$, $\forall l \in L$;
- Unit volume v_i of each product $i \in I$: Randomly selected from 10^{-3} to 10^{-2};
- Quantity q_i of each product $i \in I$ made available in 1 and absorbed in m in the unit time: Randomly selected from 50 to 100;
- Transportation cost c^l, $l \in L$: Randomly selected from 100 to 500.

In all cases, random selections are performed according to a uniform distribution.

For each of the specific situations the minimum cost among all the heuristic algorithms has been identified and the percent increase provided by each algorithm has been computed and shown in the Tables I-IV. Table I shows

Prod.	Inv.	EOQ-l	EOQ-s	Dec	SF-r	SF-a	DP
10	1-3	27.8767 (0)	5.0826 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	5.0826 (2)
10	3-6	44.2768 (0)	2.5320 (3)	0.0000 (5)	0.0000 (5)	0.0000 (5)	2.5320 (3)
10	1-6	28.1602 (0)	3.7345 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	3.7345 (2)
50	1-3	2.8749 (1)	0.8336 (2)	0.0199 (4)	0.0199 (4)	0.0199 (4)	0.8336 (2)
50	3-6	5.0117 (0)	0.2560 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.2560 (4)
50	1-6	3.5893 (0)	0.4273 (2)	0.0799 (4)	0.0013 (4)	0.0013 (4)	0.4273 (2)
100	1-3	2.4372 (0)	0.1430 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.1430 (4)
100	3-6	4.0461 (0)	0.0224 (4)	0.0000 (5)	0.0224 (4)	0.0224 (4)	0.0224 (4)
100	1-6	3.1239 (0)	0.1030 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.1030 (4)
		13.4885 (1)	1.4594 (27)	0.0111 (43)	0.0048 (42)	0.0048 (42)	1.4594 (27)

Table I: Average percent increase on 5 instances with 2 links

Prod.	Inv.	EOQ-l	EOQ-s	Dec	SF-r	SF-a	DP
10	1-3	21.3561 (0)	6.0546 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	6.0546 (2)
10	3-6	36.9157 (0)	3.3610 (3)	0.0000 (5)	0.0000 (5)	0.0000 (5)	3.3610 (3)
10	1-6	22.7169 (0)	4.7897 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	4.7897 (2)
50	1-3	1.9806 (1)	1.0727 (2)	0.0392 (4)	0.0392 (4)	0.0392 (4)	1.0727 (2)
50	3-6	3.7570 (1)	0.3549 (4)	0.0128 (4)	0.0128 (4)	0.0128 (4)	0.3549 (4)
50	1-6	2.7621 (1)	0.7637 (2)	0.2391 (2)	0.1780 (3)	0.1882 (3)	0.7637 (2)
100	1-3	1.7569 (0)	0.1738 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.1738 (4)
100	3-6	3.2769 (0)	0.0749 (4)	0.0000 (5)	0.0749 (4)	0.0749 (4)	0.0749 (4)
100	1-6	2.4289 (0)	0.1403 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.1403 (4)
		10.7723 (3)	1.8651 (27)	0.0323 (40)	0.0339 (40)	0.0350 (40)	1.8651 (27)

Table II: Average percent increase on 5 instances with 3 links

the results obtained for sequences with 2 links, Table II for sequences with 3 links, Table III for sequences with 4 links and, finally, Table IV for sequences with 5 links.

Each table is organized as follows. Column 1 gives the number of products, Column 2 the range of the inventory cost and the other columns the average percent increase on 5 instances of each heuristic and, in parentheses, the number of times each heuristic reaches the best solution.

The computational time of each heuristic algorithm has been lower than 30 seconds on each of the instances.

Prod.	Inv.	EOQ-l	EOQ-s	Dec	SF-r	SF-a	DP
10	1-3	17.2398 (0)	6.6292 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	6.6292 (2)
10	3-6	31.6026 (0)	3.8964 (3)	0.0000 (5)	0.0000 (5)	0.0000 (5)	3.8964 (3)
10	1-6	18.9958 (0)	5.4650 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	5.4650 (2)
50	1-3	1.4619 (1)	1.2278 (2)	0.0510 (4)	0.0510 (4)	0.0510 (4)	1.2278 (2)
50	3-6	2.9300 (1)	0.4358 (4)	0.0332 (4)	0.0332 (4)	0.0332 (4)	0.4358 (4)
50	1-6	2.1780 (1)	0.9346 (2)	0.2869 (2)	0.2348 (3)	0.2530 (3)	0.9346 (2)
100	1-3	1.3435 (0)	0.1906 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.1906 (4)
100	3-6	2.7742 (0)	0.1386 (4)	0.0310 (4)	0.0000 (5)	0.1386 (4)	0.1386 (4)
100	1-6	1.9692 (0)	0.1622 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.1622 (4)
		8.9439 (3)	2.1200 (27)	0.0447 (39)	0.0354 (41)	0.0529 (40)	2.1200 (27)

Table III: Average percent increase on 5 instances with 4 links

Prod.	Inv.	EOQ-l	EOQ-s	Dec	SF-r	SF-a	DP
10	1-3	14.3851 (0)	7.0094 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	7.0094 (2)
10	3-6	27.5589 (0)	4.2708 (3)	0.0000 (5)	0.0000 (5)	0.0000 (5)	4.2708 (3)
10	1-6	16.2727 (0)	5.9356 (2)	0.0000 (5)	0.0000 (5)	0.0000 (5)	5.9356 (2)
50	1-3	1.1558 (2)	1.3734 (1)	0.0954 (3)	0.0954 (3)	0.0954 (3)	1.3734 (1)
50	3-6	2.3326 (1)	0.4956 (4)	0.0483 (4)	0.0483 (4)	0.0483 (4)	0.4956 (4)
50	1-6	1.7720 (1)	1.0625 (2)	0.3243 (2)	0.2790 (3)	0.3015 (3)	1.0625 (2)
100	1-3	1.0656 (1)	0.2034 (3)	0.0023 (4)	0.0023 (4)	0.0023 (4)	0.2034 (3)
100	3-6	2.3736 (0)	0.1558 (4)	0.0260 (4)	0.0000 (5)	0.1558 (4)	0.1558 (4)
100	1-6	1.6397 (0)	0.1767 (4)	0.0000 (5)	0.0000 (5)	0.0000 (5)	0.1767 (4)
		7.6173 (5)	2.2981 (25)	0.0551 (37)	0.0472 (39)	0.0670 (38)	2.2981 (25)

Table IV: Average percent increase on 5 instances with 5 links

The following conclusions can be drawn from these results. The heuristics *Dec*, *SF-r* and *SF-a* had a similar performance: They reached the best solution in about 90% of the instances and the average error on all the instances was always less than 0.07%. Among these heuristics, the one which performed the best was the heuristic *SF-r*. The EOQ–based heuristics had a poor performance; in particular, the *EOQ-l* reached the best solution only in about 7% of the instances and the average error on all the instances was about 10%; moreover, in the instances with 10 products the performance of this heuristic was very poor with an average error of about 26%. The heuristic *EOQ-s* performed better: It reached the best solution in about 60% of the instances with an average error on all the instances of about 2%. Finally, the heuristic *DP* performed as the *EOQ-s*: This means that the *EOQ-s* is an excellent heuristic for the problem in which only one frequency can be selected on each link; secondly, given that the heuristics *Dec*, *SF-r* and *SF-a* performed better than the *DP*, we can conclude that better solutions can be obtained by allowing multiple frequencies for each product.

5 Conclusions

In this paper the logistic networks referred to as sequences of links have been studied from both the theoretical and the computational point of view.

From the theoretical point of view, particular attention has been focused on the formulation and evaluation of the total inventory cost.

From the computational point of view, given that the simpler single link case is known to be NP–hard, several heuristic algorithms have been proposed. In particular, four classes of algorithms based either on the decomposition of the sequence or on the dynamic programming techniques, have been presented. The performance of each heuristic algorithm has been evaluated on the basis of several randomly generated problem instances. The computational results showed that the EOQ–based heuristics have been outperformed by the other ones.

Acknowledgement

This work has been partially supported by Progetto Finalizzato Trasporti 2 of the CNR (National Research Council of Italy) with Contract 96.00120.PF74.

References

Anily, S. / Federgruen, A. (1990): One Warehouse Multiple Retailer Systems with Vehicle Routing Costs. in: Management Science 36, 92–114

Bertazzi, L. / Speranza, M.G. (1996): Minimization of Logistics Costs on Sequences of Links. in: Proceedings of the 4th IFIP WG7.6 Working Conference, Noisy–le–Grand, May 28–30, 1996

Bertazzi, L. / Speranza, M.G./ Ukovich, W. (1995): Exact and Heuristic Solutions for a Shipment Problem with Given Frequencies. Technical Report 101, Department of Quantitative Methods, University of Brescia (submitted)

Bertazzi, L. / Speranza, M.G./ Ukovich, W. (1997): Minimization of Logistic Costs with Given Frequencies. in: Transportation Research B 31, 327–340

Bertsekas, D.P. (1995): Dynamic Programming and Optimal Control. (Athena Scientific) Belmont

Blumenfeld, D.E. / Burns, L.D. / Diltz, J.D. / Daganzo, C.F. (1985): Analyzing Trade-offs between Transportation, Inventory and Production Costs on Freight Networks. in: Transportation Research 19B, 361–380

Burns, L.D. / Hall, R.W. / Blumenfeld, D.E. / Daganzo, C.F. (1985): Distribution Strategies that Minimize Transportation and Inventory Cost. in: Operations Research 33, 469–490

Hall, R.W. (1985): Determining Vehicle Dispatch Frequency when Shipping Frequency Differs among Suppliers. in: Transportation Research 19B, 421–431

Jackson, P.L. / Maxwell, W.L. / Muckstadt, J.A. (1988): Determining Optimal Reorder Intervals in Capacitated Production–Distribution Systems. in: Management Science 34, 938–958

Maxwell, W.L. / Muckstadt, J.A. (1985): Establishing Consistent and Realistic Reorder Intervals in Production–Distribution Systems. in: Operations Research 33, 1316–1341

Muckstadt, J.A. / Roundy, R.O. (1993): Analysis of Multistage Production Systems. in: Graves, S.C., Rinnooy Kan, A.H.G. and Zipkin, P.H. (eds.) Handbooks in Operations Research and Management Science, Vol. 4: Logistics of Production and Inventory, North–Holland, 59–131

Speranza, M.G. / Ukovich, W. (1992): A Decision Support System for Materials Management. in: International Journal of Production Economics 26, 229–236

Speranza, M.G. / Ukovich, W. (1994a): Analysis and Integration of Optimization Models for Logistic Systems. in: International Journal of Production Economics 35, 183–190

Speranza, M.G. / Ukovich, W. (1994b): Minimizing Transportation and Inventory Costs for Several Products on a Single Link. in: Operations Research 42, 879-894

Speranza, M.G. / Ukovich, W. (1996): An Algorithm for Optimal Shipments with Given Frequencies. in: Naval Research Logistics 43, 655–671

Using break quantities for tactical optimisation in multi-stage distribution systems

Marcel J. Kleijn[1] and Rommert Dekker[1]

[1] Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Abstract. In this chapter we discuss a tactical optimisation problem that arises in a multistage distribution system where customer orders can be delivered from any stockpoint. A simple rule to allocate orders to locations is a break quantity rule, which routes large orders to higher-stage stockpoints and small orders to end-stockpoints. A so-called break quantity determines whether an order is small or large. We present a qualitative discussion on the implications of this rule for the marketing process, and a qualitative and quantitative analysis on the implications for the transportation and inventory costs. Furthermore, we present a case study for a company that implemented a break quantity rule. Finally, in the last section the main results are summarised.

Keywords. Distribution systems, inventory, transportation, marketing, break quantity rule

1 Introduction

Distribution systems are concerned with the effective management of the delivery of finished goods to the final customers. Since in general there are many complex interactions between the components of such a system, the decision process can be extremely difficult. Therefore, in most cases the decisions are decomposed into *strategical*, *tactical* and *operational* decisions. Strategical decisions include the determination of the location and size of factories and warehouses, the design of transportation facilities, and so on. Tactical decisions cover the problem of how to use the resources in such a way that customer demand is met at minimum cost or maximum service. Operational decisions involve all day-to-day operational and scheduling decisions. Since an integral approach of these categories is impossible (see e.g. Tüshaus & Wahl (1997)), one often uses an hierarchical approach, where first the strategical decisions are made, followed by the tactical and operational decisions.

In this chapter we discuss a tactical problem that occurs when at the strategical level the decision is made that customer orders can be delivered both from end-stage stockpoints (say, warehouses) and from higher-stage stockpoints (e.g. central warehouses or distribution centres) or factories. Deliveries from a higher-stage stockpoint or factory will henceforth be referred to as *direct deliveries* (Fleischmann (1993, 1997)). Direct deliveries can be advantageous because bypassing a warehouse results in shorter distances and saves warehousing (handling, storage) costs. However, there may also be a loss in transportation economies of scale, thus raising the transportation costs. If direct deliveries are allowed, then upon arrival of a customer order a decision has to be made from which location to deliver this order. In principle, an optimal decision will depend on the locations of the customer, the factories and the warehouses, on the size of the order, on the stock levels at the factories and warehouses, on the amount in transit to the warehouses and on the maximum delivery lead time quoted by the customer. Furthermore, if it is possible to combine the delivery of orders, an optimal decision will also depend on orders by other customers. In practice, there is a need for rules that are easy to understand and implement. A simple way to allocate orders to stockpoints is to route large orders to the nearest higher-stage stockpoint or factory, and small orders to the nearest warehouse (see e.g. Ballou (1992), Fleischmann (1993, 1997)). This rule will be called a *break quantity rule*, where a so-called break quantity determines whether an order is small or large. The implementation of such a rule in logistics software is very simple. Some standard packages already include the option of setting a maximum issue quantity. The large orders may then be allocated automatically or by the logistics manager.

Applying a break quantity rule will have a number of opposite effects on the costs and service level in a distribution system, thus making the determination of a good break quantity a difficult task. In most distribution networks a weight limit of 1 or 2 tons is used (Fleischmann (1997)). However, this number is usually based on experience and intuition, rather than on a quantitive analysis. In this chapter we will discuss how to determine a good break quantity by carefully examining the relevant costs. In the next sections we analyse the implications of a break quantity rule on the performance of a distribution system, and we present a discussion on the determination of the break quantity. In Section 4 a case study is discussed, in which some additional complexities that may arise in practice are addressed. The main results are summarised in the last section.

2 Implications of a break quantity rule

In this section the influence of a break quantity rule on the performance of a general distribution system is analysed. In particular, we discuss the effects on the marketing process, the transportation and handling costs, and the inventory costs.

An important motivation for using warehouses in a distribution system is the improved customer service that is caused by shorter delivery lead times. Fleischmann (1993, 1997) observed that the high level of competition in many consumer goods markets has caused the customers to claim a better distribution service, in particular shorter delivery lead times and more frequent deliveries of smaller amounts. Nevertheless, the introduction of a break quantity rule may lead to longer delivery lead times for large orders, for example if small orders are delivered from stock and large orders are handled on a produce-to-order basis. This situation was described in a case study (Nass, Dekker & Sonderen-Huisman (1997)) for a company that applied a break quantity rule. However, Kok & Janssen (1996) argue that a major reason for the occurrence of occasional large orders is the discount structure used by companies to increase sales, and the need for immediate delivery is much less for these large orders. Also, a customer placing a large order may be the manager of another warehouse, and the delivery lead time can be negotiated upon. Therefore, we conjecture that the possible increase in delivery lead time has the least negative marketing effects for large orders. Finally, in many practical situations, the arrival of an unexpected large order causes the management to make an ad-hoc decision whether or not to deliver the order from stock. Applying a break quantity rule on the one hand means less flexibility for the management, but on the other hand it creates a consistent view towards customers. Upon order entry, direct feedback can be given about the delivery of the order. If a customer does not accept the break quantity rule and considers placing his large order at a competitor (e.g. because of an increase in delivery lead time for the large order), he may be convinced by offering a price rebate (Kasturi Rangan & Jaikumar (1991), Kok & Janssen (1996)), which can be financed by the reduction in transportation, handling and inventory costs. If the customer still can not be convinced, then an exception can be made for him.

Transportation costs are mainly depending on distance and shipment size. Due to economies of scale, these costs are typically an increasing and concave function of the distance or the shipment size (Ballou (1992)). The influence of a break quantity rule on the transportation costs is very difficult to predict, because there are opposite effects. On the one hand, since direct deliveries are always shorter than deliveries via a warehouse, the transportation costs will decrease. On the other hand, some economies of scale are lost because ship-

ments that used to be consolidated in replenishment orders to the warehouse are now shipped directly. This causes the transportation costs to increase. If customers are located near the warehouse, then it is likely that direct deliveries are more expensive than deliveries via the warehouse. However, this cost increase is minimised if direct deliveries are preserved for large orders. A perfect situation arises if a large order implies a full truck load, since in this case a break quantity rule has no increasing effect on the transportation costs. Finally, we observe that a break quantity rule implies that part of the demand is bypassing the warehouse, and thus the handling costs are reduced.

If a break quantity rule is applied, then at the warehouses the peaks in demand are filtered out, which results in a reduction of the average demand at the warehouses, a reduction of the average stock in transit to the warehouses, and a reduction of the demand variability at the warehouses. Hence, the inventory holding costs at the warehouses will decrease. The greatest reduction is obtained for items having an erratic demand pattern, i.e. items which have occasional very large demand transactions interspersed among a majority of small transactions (Silver (1970)). Safety stock levels for such items tend to be quite large in order to meet certain service requirements. Orders that are not allocated to the warehouses must be delivered from another location. If this location is e.g. a factory that produces to order and does not keep inventory, a break quantity rule will have no inventory effect here. But, if large orders have priority over replenishment orders, a break quantity rule may cause an increase in production costs. If the other location holds inventory, say a central warehouse which supplies regional warehouses, there will be a negative effect on the inventory costs, i.e. they will increase. The central warehouse will now face the occasionally occurring large orders, which increases the demand variability and thus leads to higher inventory costs. However, if large orders from several regional warehouses are allocated to the same central warehouse, the centralisation effect (Eppen (1979)) at this warehouse induces the inventory costs to decrease (see also Dekker, Kleijn & Kok (1997)). As was the case for the transportation costs, the net effect on the central warehouse inventory costs is difficult to predict.

To conclude this section, we summarise in Table 2.1 the main advantages and disadvantages of using a break quantity rule instead of a traditional policy where all demand is delivered from the warehouse.

Table 2.1: Main advantages/disadvantages of break quantity rule

advantages	disadvantages
1. total transportation distance decreases	1. longer delivery lead times
2. less stock needed at warehouse	2. less transportation ec. of scale
3. less handling	3. more stock needed at central warehouse or factory

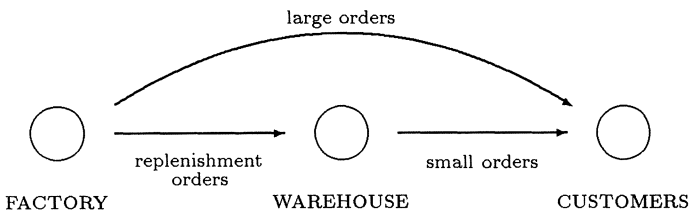
3 Determining the break quantity

When a break quantity rule is applied, some important decisions have to be made. For example, a company having customers from different regions can set a different break quantity for each region, or one break quantity for all regions. Although the first option slightly decreases the consistency towards the customers, it allows for a better trade-off between inventory and transportation costs, since transportation costs are in most cases region dependent. Another important issue is the determination of the break quantity. Such a decision will usually be made using a qualitative analysis with respect to the marketing process and a quantitative analysis with respect to the transportation and inventory costs. As far as the marketing process is concerned, we assume that the break quantity rule is accepted by the customers and thus will have no effect on the total demand. In the next subsections the quantitative analysis with respect to the transportation and inventory costs is discussed, for a simple distribution system consisting of one factory, one warehouse and one customer region.

3.1 Notation and assumptions

Consider a simple distribution system consisting of a factory, a warehouse, and some customers located in the same region. It is assumed that a break quantity rule is applied, with a break quantity equal to q. The system is illustrated by Figure 3.1.

Figure 3.1: The distribution system



It is also assumed that the demand process can be described by a compound Poisson process, with arrival rate λ. Upon arrival, a customer places an order for j units with probability $a(j)$, $j = 1, \ldots, \infty$. The main reason for modelling demand in this way is because it allows us to distinguish between customers based on their order sizes. The distribution of the demand during a period of t time units, given a break quantity q, has a density function denoted by f_q^t. Using Adelson's recursion scheme (Adelson (1966)), it can be shown that

this density function satisfies the following recursive relations

$$f_q^t(j) = \begin{cases} e^{-\lambda t(1-a_q(0))} & \text{if } j = 0 \\ (\lambda t/j) \sum_{i=0}^{j-1} (j-i) a_q(j-i) f_q(i) & \text{if } j = 1, 2, \ldots \end{cases}$$

where

$$a_q(j) := \begin{cases} \sum_{i=q+1}^{\infty} a(i) & \text{if } j = 0 \\ a(j) & \text{if } 1 \leq j \leq q \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Large orders are produced to order at the factory and shipped to the customer as soon as possible. Small orders are delivered from stock on hand if possible, and demand which can not be satisfied directly from stock on hand is backlogged. The scheduling of the replenishment orders is assumed to be determined by the inventory policy.

3.2 Transportation costs

In many situations, the transportation of finished goods to customers is contracted out (Ballou (1992), Fleischmann (1993)). In this case we define the following relevant transportation tariffs:

$T_{fw}(i)$: the costs of shipping i units from the factory to the warehouse

$T_{wc}(i)$: the costs of shipping i units from the warehouse to a customer

$T_{fc}(i)$: the costs of shipping i units from the factory to a customer

A typical transportation tariff consists of a minimum charge, and (decreasing) transport rates for several weight classes (Ballou (1992), Fleischmann (1993)).

For a given break quantity q, the expected transportation costs per time unit for shipments to the customers are given by

$$\lambda \sum_{j=1}^{q} T_{wc}(j) a(j) + \lambda \sum_{j=q+1}^{\infty} T_{fc}(j) a(j)$$

The average cost for the transportation of replenishment orders depends on the inventory policy. For example, if a replenishment order is placed every R time units, the expected transportation costs per time unit are given by

$$(1/R) \sum_{j=1}^{\infty} T_{fw}(j) f_q^R(j)$$

If replenishment orders are always shipped in batch sizes equal to Q, these expected costs become

$$(1/Q)T_{fw}(Q)\sum_{j=1}^{\infty} jf_q(j) = (1/Q)T_{fw}(Q)\lambda\sum_{j=1}^{q} ja(j)$$

Observe that $\sum_{j=1}^{\infty} jf_q(j) = \lambda\sum_{j=1}^{q} ja(j)$ denotes the average demand for the warehouse per time unit, if the break quantity equals q. Many times in practice replenishment orders are transported in full truck loads, implying that the transportation tariff for these orders is linear in the shipment size (e.g. Nass, Dekker & Sonderen-Huisman (1997), Fleischmann (1997)). This is possible if for example replenishment orders for different items are combined in one shipment. In this case the expected transportation costs are given by

$$TC(q) = \lambda\sum_{j=1}^{q} T_{wc}(j)a(j) + \lambda\sum_{j=q+1}^{\infty} T_{fc}(j)a(j) + \lambda\sum_{j=1}^{q} t_{fw}ja(j)$$

with t_{fw} the transportation rate for replenishments orders. Observe that these costs are independent of the inventory policy at the warehouse, since the scheduling of the replenishment orders no longer affects the transportation costs. It is now relatively easy to analyse the effect of a break quantity rule on the transportation costs.

3.3 Inventory costs

A break quantity rule can be combined with any inventory policy at the warehouse, since it only influences the demand distribution. If the inventory policy parameters were first determined based on a compound Poisson distribution with arrival rate λ and order size distribution $a(j)$, $j = 1, \ldots, \infty$, the new parameters can be set in a similar way for a compound Poisson demand distribution with the same arrival rate and order size distribution $a_q(j)$, $j = 1, \ldots, q$ (see (1)). Examples of such approaches are given in Silver (1970), Hollier, Mak & Lam (1995a, 1995b) and Mak & Lai (1995a, 1995b).

Most inventory control systems operate with approximative models, which are reasonable since the total cost curve usually has a flat bottom, so that slight deviations from optimum values of the policy parameters result in only small changes to the total costs. Therefore, as an example, we now discuss an approximative inventory model, where the only relevant demand information is contained in the mean and variance of the demand per time unit. For a given break quantity q, it can be shown (see e.g. Tijms (1994)) that this mean μ_q and variance σ_q^2 are equal to

$$\mu_q = \lambda\sum_{j=1}^{q} ja(j)$$

$$\sigma_q^2 = \lambda \sum_{j=1}^{q} j^2 a(j)$$

Suppose we have an inventory system where every R time units a replenishment order is placed which arrives L time units later, and management has implied the restriction that the probability of a stockout during the lead time plus review time is less than α, with $0 < \alpha < 1$. With K the fixed cost for placing a replenishment order and h the unit holding cost, we obtain that (see e.g. Ballou (1992)) an approximation for the average costs is given by

$$IC(q) = (1/R)\left(K + \frac{1}{2}h\mu_q R + hz\sigma_q\sqrt{R+L}\right)$$

with $z := \Phi^{-1}(\alpha)$ and Φ the standard normal distribution. If the pipeline inventory is also taken into account, the average costs become

$$IC(q) = (1/R)\left(K + h\mu_q(\frac{1}{2}R + L) + hz\sigma_q\sqrt{R+L}\right)$$

One can observe that the inventory costs are increasing with μ_q and σ_q, and thus with the break quantity q. For higher service levels, the value of z tends to be higher, and the inventory costs become more sensitive to the standard deviation of the demand. Also, the lead time for replenishment orders has a significant influence on the inventory holding costs.

Optimising the inventory costs with respect to the break quantity always leads to a break quantity of zero, because no inventory is maintained at the factory. However, in this case all orders are shipped directly from the factory, which may lead to very high transportation costs. Therefore, the optimisation of the break quantity should be based on both the inventory and the transportation costs. In Dekker et al. (1997) it is shown that in general the average cost function does not have a shape that allows for the design of a straight optimisation algorithm. However, using enumeration only over values of q satisfying $a(q) > 0$ it is possible to determine the break quantity that minimises the average transportation and inventory costs.

4 A case study

Recently, we analysed a company in Western-Europe that applied a break quantity rule. The company produces technical thermoplasts in many different grades and colours. About 50% of the total volume is produced to order, which corresponds to 90% of the product varieties. The remaining volume is produced to stock. Customers are located all over the world, but most of them are located in Europe. The company has four production plants, located in different countries (Spain, France, Scotland and the Netherlands),

and in each of these plants different products are manufactured. Furthermore, the company has one distribution centre, located in the Netherlands, in which different orders for the same customer are consolidated and shipped to the customer at the end of every week. Orders for produce-to-order products can either be delivered directly to the customer, or they can first be shipped to the distribution centre, where they are consolidated with other orders from that customer. Orders for produce-to-stock products are delivered from stock on hand at the distribution centre, and in case of a stockout the order will be handled as a produce-to-order product. If a customer places an order for a product, the company immediately promises a certain delivery date. Reliability of the delivery date is very important for the company, more important than the actual lead time. The break quantity rule is implemented such that an order for a produce-to-order product is shipped directly to the customer if the size of the order exceeds the break quantity, and it is shipped via the distribution centre otherwise. Since the location of the customer has a large impact on the transportation costs, the company asked us to determine a break quantity for each region. The transportation to the customers is contracted out, while the replenishment orders are shipped to the distribution centre with full truck loads. Hence, the costs for replenishment orders are proportional to the shipment size. Also, the handling and inventory costs were assumed to be proportional to the shipment size. The inventory costs for the produce-to-order products were low compared to the transportation costs, since on average the goods were kept in stock at the distribution centre for only three days.

As in Section 3.2, the transportation tariff for replenishment orders is denoted by t_{fw}, where it is assumed that this rate includes the handling and inventory costs at the distribution centre. Moreover, the transportation costs for shipping j units from the production plant to a customer in region r, resp. from the distribution centre to a customer in region r, are denoted by $T^r_{fc}(j)$ and $T^r_{wc}(j)$ respectively.

The main problem we encountered was the following: whenever a customer places an order there is no information available on other outstanding orders from the same customers. Hence, it is impossible to determine the transportation costs from the distribution centre to the customer, because the total shipment size is not known. However, it was possible to determine an upperbound on the optimal break quantity, such that direct deliveries of orders with sizes exceeding this upperbound are cheaper than deliveries via the distribution centre. To obtain this upperbound, we observe that the transportation cost rate for shipments from the distribution centre to customers in region r is bounded from below by the full truck load cost rate, i.e. $t^r_{wc} := \lim_{j\to\infty} T^r_{wc}(j)/j$. Hence, the costs of delivering an order of j units via the distribution centre can never be smaller than $(t_{fw} + t^r_{wc})j$, so

an upperbound for the optimal break quantity in region r is given by

$$\min\{j \geq 0 : T^r_{fc}(i) \leq (t_{fw} + t^r_{wc})i \text{ for all } i \geq j\} \tag{2}$$

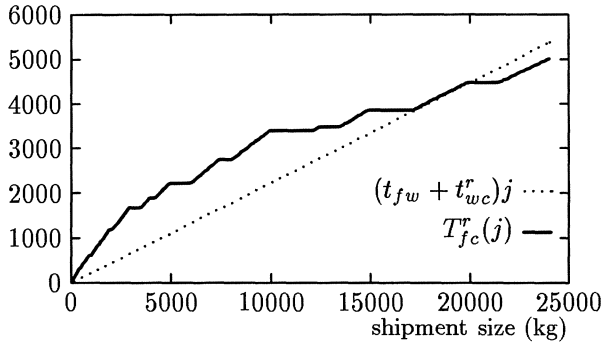
As an example we consider a customer region in Germany and the production plant in Spain. The tariffs for transportation from the plant in Spain to the customers in Germany and from the distribution centre in the Netherlands to the customers in Germany are given in Table 4.1.

Table 4.1: Transportation rates (Dfl/100 kg)

	shipment size (tons)						
	minimum	0-0.1	0.1-0.5	0.5-1	1-2	2-3	3-4
Spain-Germ.	0	89.32	75.72	69.19	62.26	57.89	48.31
Neth.-Germ.	0	65.00	32.50	24.20	16.60	13.00	11.20
	4-5	5-7.5	7.5-10	10-12.5	12.5-15	15-20	>20
Spain-Germ.	45.36	37.20	34.29	28.12	25.95	22.53	20.88
Neth.-Germ.	9.50	7.90	6.60	5.60	5.10	4.70	4.35

The costs for replenishment orders were 18.04 Dfl/100 kg (including 3.00 holding and handling costs). The lower bound on the cost rate of transportation via the distribution centre equals 22.39 (4.35+18.04) Dfl/100 kg. In Figure 4.1 both the transportation tariff $T^r_{fc}(j)$ and the lower bound $(t_{fw} + t^r_{wc})j$ are plotted, illustrating how the upper bound given in (2) can be determined. The flat parts in the tariff $T^r_{fc}(j)$ are caused by blanketing back (Ballou (1992)). For example, the cost $T^r_{fc}(21000)$ of transportating 21 tons from the factory in Spain to the customer in Germany is determined by $\min\{T^r_{fc}(20000), 20.88 * 21000/100\} = T^r_{fc}(20000) = 4506$. It can be verified that the upper bound given in (2) equals 20125 kg, i.e. $T^r_{fc}(20125) = 4506 = 22.39 * 20125/100$.

Figure 4.1: Ilustration of determination upper bound on break quantity



The 1994 demand of an arbitrary single customer, located in the region in Germany, is given in Table 4.2.

Table 4.2: Demand (kg) of a customer during 1994

week	1	8	11	14	22	27
size (kg)	500	19800	4000	500	20	18750
week	27	28	29	30	39	
size (kg)	21875	10420	7920	21525	19540	

For this particular customer we present the costs of the following different policies:

1. *all direct*, **costs 31072**. With this policy all orders are delivered directly from the production plant in Spain.

2. *all via dc*, **costs 29113**. Here all the orders are delivered via the distribution centre. Observe that in week 27 two orders are consolidated.

3. *upperbound*, **costs 28530**. In this policy a break quantity rule is applied, with a break quantity equal to the upperbound 20125 kg.

4. *optimal break quantity*, **costs 28408**. This policy uses a break quantity rule with the optimal break quantity (any size between 10420 and 18749 kg), obtained by evaluating the total costs for all possible break quantities.

The cost reduction obtained by using a break quantity rule with $q = 20125$ instead of a policy where all orders are delivered via the distribution centre was 2%. The cost difference between using the upper bound (2) instead of the optimal break quantity was only 0.4%. With these observations we conclude our discussion of the case study.

5 Conclusions

In this chapter we discussed a tactical optimisation problem that arises when orders can be delivered from any stockpoint in the distribution system. A simple rule to allocate orders to locations is a break quantity rule, which routes large orders to higher-stage stockpoints (central warehouses, factories) and small orders to end-stockpoints (warehouses). A so-called break quantity determines whether an order is small or large.

The implications of a break quantity rule for the marketing process and the transportation, handling and inventory costs were described in Section 2, and in Section 3 a quantitative analysis of the impact on the transportation and inventory costs was presented. Summarising, the analysis for the

transportation costs consists of comparing transportation tariffs for different break quantities, while the analysis for the inventory costs mainly focusses on the determination of the demand parameters, as a function of the break quantity, from which the average costs can be determined. The aggregate effect on both the transportation and the inventory costs should determine whether it is worthwhile to implement a break quantity rule, and if so, how large the break quantity should be. Finally, in Section 4 a case study was presented, illustrating an additional complexity that may arise in practice.

It is difficult to say if a break quantity rule will lead to a better performance of a distribution system, without looking closely at the marketing process and the transportation/inventory costs. However, in general it seems worthwhile to consider the implementation of a break quantity rule in distribution systems where demand is erratic (i.e. occasional very large demand transactions interspersed among a majority of small transactions) and the sizes of these large orders approach the full truck load size. For this situation, the reduction of the inventory and transportation costs will be significant if a break quantity rule is applied.

References

Adelson, R.M. (1996): Compound Poisson distributions, in: Operations Research Quarterly 17, 73–75.

Ballou, R.H. (1992): Business Logistics Management, 3rd edition. (Prentice Hall) Englewood Cliffs, NJ.

Dekker, R./Frenk, J.B.G./Kleijn, M.J./Kok, A.G. de (1997): On the newsboy model with a cutoff transaction size. Technical Report 9736/A, Econometric Institute, Erasmus University Rotterdam, The Netherlands.

Dekker, R./Kleijn, M.J./Kok, A.G. de (1997): The break quantity rule's effect on inventory costs in a 1–warehouse, N–retailers distribution system, to appear in: International Journal of Production Economics.

Eppen, G.D. (1979): Effect of centralization on expected costs in multi–location newsboy problem, in: Management Science 25, 498–501.

Fleischmann, B. (1993): Designing distribution systems with transport economies of scale, in: European Journal of Operational Research 70, 31–42.

Fleischmann, B. (1997): Design of freight traffic networks, in: Advances in Distribution Logistics, P. Stähly et al (editors).

Hollier, R.H./Mak, K.L./Lam, C.L. (1995a): Continuous review (s, S) policies for inventory systems incorporating a cutoff transaction size, in: International Journal of Production Research 33, 2855–2865.

Hollier, R.H./Mak, K.L./Lam, C.L. (1995b): An inventory model for items with demands satisfied from stock or by special deliveries, in: International Journal of Production Economics 42, 229–236.

Kasturi Rangan, V./Jaikumar, R. (1991): Integrating distribution strategy and tactics: a model and an application, in: Management Science 37, 1377–1389.

Kok, A.G. de/Janssen, F.B.S.L.P. (1996): Demand management in multi-stage distribution chain. Technical Report 9639, Center for Economic Research, Tilburg University, The Netherlands.

Mak, K.L./Lai, K.K. (1995a): The determination of optimal partial fill policy for an inventory system with lumpy demand items, in: Applied Mathematical Modelling 19, 724–737.

Mak, K.L./Lai, K.K. (1995b): Optimal (s, S) policy for an inventory system when the demand can be partially satisfied, in: International Journal of Systems Science 26, 213–231.

Nass, R./Dekker, R./Sonderen-Huisman, W. van (1997): Distribution optimization by means of break quantities: A case study, to appear in: Journal of the Operational Research Society.

Silver, E.A. (1970): Some ideas related to the inventory control of items having erratic demand patterns, in: Journal of the Canadian Operations Research Society 8, 87–100.

Tijms, H.C. (1984): Stochastic Models: An Algorithmic Approach. (Wiley) New York.

Tüshaus, U./Wahl, C. (1997): Inventory positioning in a two-stage distribution system with servicel level constraints, in: Advances in Distribution Logistics, P. Stähly et al (editors).

Estimating the length of trunk tours for environmental and cost evaluation of distribution systems

Stefan Kraus

Universität Augsburg, Lehrstuhl für Produktion und Logistik, Universitätsstraße 16,
D-86135 Augsburg

Summary. During the last 10 years the road freight traffic in Germany, measured by vehicle miles, increased dramatically by about 41 %. About 40-50 % of this mileage is estimated to be caused by the distribution of consumer goods. Thus, the decisions of the single industrial companies on their distribution systems have a strong influence on the total freight traffic and on the environment. In a distribution network the majority of the traffic (about 70-80 %) is caused by the delivery from the warehouses to the customers, commonly served in tours. Therefore, we distinguish between local delivery tours for small order sizes over short distances and direct deliveries of orders with a large size in trunk tours. Usually, in the framework of strategical distribution network design the assignment of the customers to the warehouses or transshipment points is optimized on the base of direct distances between warehouse and customers; the real length of the tours are not regarded. Considering that most of the environmental parameters for evaluating the traffic in distribution systems (e.g. fuel consumption, emissions) behave nearly proportional to the vehicle miles, the estimation of the tour length is very important. Futhermore transportation costs are strongly dependent on the tour length. The investigation aims at estimating the length of direct deliveries in trunk tours. We focus on the presentation of an estimation model, and first computational results are shown comparing the accuracy of the model with real tour data of a consumer goods manufacturer. In addition, we show numerical examples for evaluating the trunk traffic by environmental measures and by costs.

1. Introduction

During the last 10 years the road freight traffic in Germany increased dramatically. The official statistics show that, during this period, the mileage of the freight traffic (measured in kilometers) grew by 41 % and the freight traffic result (measured in ton-kilometers) by 61 % (see Bundesminister für Verkehr (1995)). This development points out the increasing ecological impact of the freight traffic on the environment, which entailed a strong sensibility of the population in Germany (see Ihde/Eckart/ Stieglitz (1994); Klotz (1992); Kreitmair/Kraus (1995)). With respect to the shipper, the growth of the traffic goes hand in hand with negative effects on the logistics operations, as congestion and traffic jams reduce his service level. Further, the political reactions in Germany result in negative economic effects on the transportation business, e.g. road pricing or increasing taxes.

Estimations show that in Germany about 40-50 % of the vehicle miles in the freight traffic are caused by the distribution of consumer goods (see Bundesamt für Güterverkehr (1993)). This highlights the influence of the consumer goods distribution on the total road traffic and the need for investigating the distribution traffic under environmental aspects. Many studies in the consumer goods industry have shown that

the amount of traffic required for the distribution of a certain producer strongly depends on the structure of the distribution system, i.e. the number of transport stages, the number of warehouses and the distribution paths. However, various trends in the last 10 years - in particular the concentration of stocks in a few warehouses, the replacement of warehouses by stockless transshipment points and the increase of delivery frequencies - have reduced the total distribution costs (see Fleischmann (1993)), but induced remarkably more traffic. While the cost effects of the distribution system design have been investigated intensively (see Kunz (1977); Konen (1982); Konen (1985); Krieger (1984)), the impact on traffic and environment has been mostly neglected so far.

The undesired effects of distribution traffic have motivated serveral new global concepts of logistics, e.g. Freight Traffic Centers (see Bahn AG (1995); Eckstein (1995); Heinrich (1995)) and the so-called "City Logistics" (see Hautau (1995); Köhler (1995)), but a clear quantitative evaluation of these concepts is not available. The objective of our investigation is to quantify the freight traffic induced by a distribution system. It should contribute to an objective theoretical foundation for the current intensive discussion on ecological aspects of traffic in Germany.

In several case studies we found that in a distribution system the majority of the traffic (about 70-80 %) is caused by the delivery from the warehouses to the customers, which are usually served in tours. Considering that most of the environmental parameters for evaluating the traffic in distribution systems (e.g. fuel consumption, emissions) behave nearly proportional to the mileage (see Kraus (1996)), the determination of the tour length is very important. Furthermore, transportation costs are strongly dependent on the tour length. For that reason we focus in our paper on investigating the delivery tours to the customers. In Section 2 the structure and the logistics operations of a distribution network are described. Section 3 presents a new model for estimating the length of trunk tours in distribution networks. The numerical results in Section 4 show the accuracy of our estimation model, testing it with real tour data of a consumer goods manufacturer. Additionally, we present numerical examples for evaluating trunk tours by environmental measures (Section 5) and by transportation costs (Section 6).

2. Distribution network

In a distribution network different types of transport relations have to be distinguished. Typically, a distribution network in Germany comprises 3 transportation stages (see Figure 2.1 and Fleischmann (1997)).

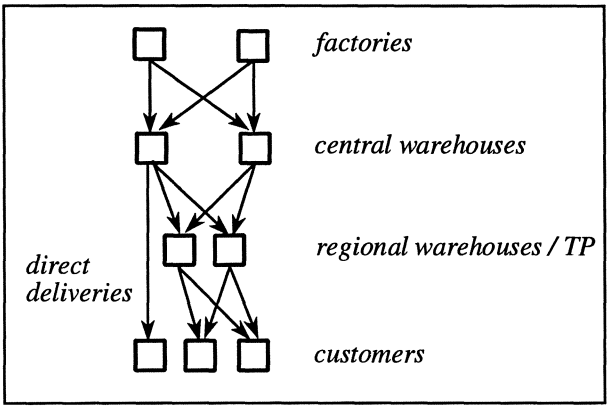


Fig. 2.1 Distribution network

The different products of the factories (F) are delivered to central warehouses (CW). In trunk transports the commodities are shipped from the CW to the regional warehouses (RW) or to stockless transshipment points (TP). The customers (wholesalers, retailers, department stores, etc.) are commonly supplied by the RW or TP. For customers, who have large order sizes, it may be profitable to serve them directly from the CW in trunk transports. Subsequently, we will call this kind of transportation direct deliveries, and analogeous these customers are named direct customers. A distribution network in Germany has the typical size of 1-8 factories, 0-5 CW, 5-40 RW or TP and 2,000-100,000 customers. More detailed information about the structure of a distribution system is given in Fleischmann (1993) and Paraschis (1989).

The stockage in the warehouses enables a bundling of transports, hence on the relations F/CW and CW/RW usually full truck loads are shipped. Since a TP keeps no stocks, there is the need to supply it daily from the CW. If the shipment quantity to a TP comprises less than a truck load, then for using the truck capacity efficiently, the deliveries are combined with direct deliveries to trunk tours. Typically orders with a large size (usually more than 1 or 2 tons) are delivered over long distances (in average about 300 km) in trunk tours. These trunk tours are investigated in detail in Section 3. Customers with small order sizes are served over short distances (commonly less than 100 km) from the RW or TP in local delivery tours. In the context of strategical distribution network design local delivery tours are already investigated by Fleischmann (1979), (1997) and Tüshaus/Wittmann (1997). Subsequently we focus on regarding trunk tours.

3. Estimation model for trunk tours

In the framework of strategical distribution network planning, only the direct distances between warehouse and customer, resp. TP, are considered for optimization. This proceeding is reasonable and typical, because the solution with a multidepot vehicle routing procedure would mix strategic planning aspects, e.g. location of warehouses and facilities, with operational questions, for example the daily scheduling of the

vehicles. Furthermore, transportation tariffs used in optimization models for strategical distribution network planning are often designed in a way that the price for a delivery is calculated by the direct transportation distance and the weight of the freight. For ecological and cost evaluation, it is too inaccurate to only consider the direct distances. We found that regarding the direct distances only may overestimate the mileage of trunk tours up to 100 %, if the transportation is executed by a carrier (see also Figure 3.1). As already discussed above, for ecological evaluation of distribution networks the accurate determination of the mileage required for transportation is very important. Since no data of routing is available in the framework of strategical distribution planning, the length of the tours has to be estimated.

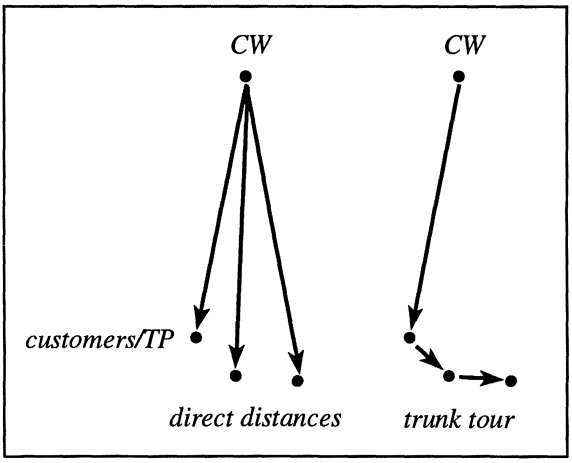


Fig. 3.1 Direct distances and trunk tours

The estimation of delivery tours is generally investigated in the literature (see Daganzo (1991); Fleischmann (1979); Newell/Daganzo (1986); Ohseok/Golden/Wasil (1995)). These models do not consider the special characteristics of trunk tours, such as the small number of customers (2 to 3 customers in average) and the large order size (commonly > 1 ton), hence they will lead to inaccurate results. In addition these models often use input-data, which is not available or which is difficult to determine in the framework of strategical distribution planning, such as the shape or the surface of delivery areas.

We present a new model for estimating the length of trunk tours in distribution networks, which enables the assignment of mileage, environmental meassures and transportation costs to a single order. This model only uses input-data, which is available directly from the distribution network (direct distance and order size) and additional parameters, which commonly in practice are easy to be determined. For modelling, the following assumptions are taken into account:

a) The CW is the starting point of the tour.
b) The transportation is performed by a carrier. For the return of the vehicle from the

supplied customers we assume that the carrier has the possibility to take freight for the way back from the delivery area. Hence the way for returning from the last customers of a tour does not have to be taken into account for estimating the required tour length for distribution. In conclusion, complete vehicle cycles are not to be regarded.

c) Orders, which size is greater than the vehicle capacity, are served in several deliveries.

d) The distances between subsequent customers in a tour are constant.

e) The orders, which are combined to a tour with the regarded order k, have an equal order size. As will be shown later, this assumption has only formal character and does not influence on the model results. It is only important for the deduction of the total truck load (see Section 3.1 and equation (13)).

The assumptions a) to e) lead to the following model (see Figure 3.2):

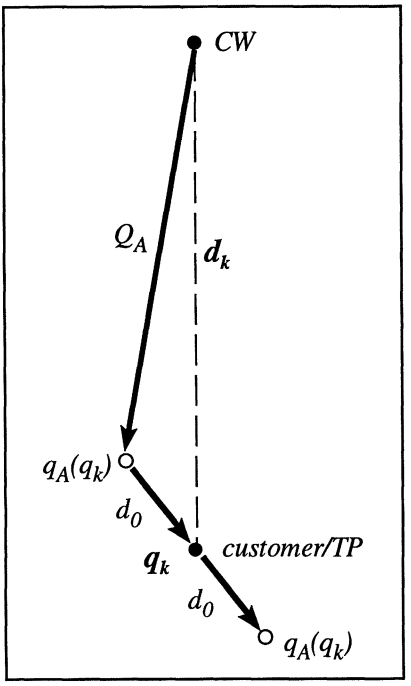


Fig. 3.2 Modelling a trunk tour

Index:

k = {1, ..., K} order of a customer/TP

Data directly available from the distribution network:

q'_k size of order k [kg]

d_k distance from CW of the customer/TP with order k [km]

Data usually known by the carrier:

Q	vehicle capacity [kg]
Q_A	average vehicle load on the way to the first customer in the tour ($Q_A \leq Q$) [kg]
d_0	average distance between the customers [km]
n_A	average number of customers in a tour

Parameter to determine:

$q_A(q'_k)$ size of an order combined with order k (function of q'_k) [kg]

The Parameters Q, Q_A, d_0 and n_A are usually known by the carrier responsible for the transportation or are easy to be estimated by a (small) sample of real tour data. The value for $q_A(q'_k)$ is to be calculated as shown subsequently.

If the customer with order k is served first in the tour, the (mean) estimated mileage D_k of a tour, in which the order k is delivered, is

$$D_k = d_k + d_0 \cdot (n_A - 1) . \quad (1)$$

Thereby, d_k denotes the distance to the first customer and the term $d_0 \cdot (n_A - 1)$ quantifies the distance between the customers in the tour. For a precise estimation of the mileage for delivering an order k, three additional items have to be regarded.

i) Since several orders are delivered on a tour, only a part of the mileage D_k is to assign to the order k. The basis for an assignment may be e.g. the distance of the customers from the CW, the number of the orders delivered on a tour or the order size. The distance-dependent assignment will allocate nearly the same amount of mileage to each order, because usually the customers in a tour are located in the same region. In the same manner, the assignment based on the number of orders has the effect that to every order on a tour the same amount of mileage is allocated. This proceeding neglects the different use of the vehicle capacity by the combined orders on a tour. Thus, the assignment based on the order size fits better with the causal principle.

ii) As an order of size $q_k = Q$ can not be combined with other orders, an assignment of the mileage with the amount $d_k + d_0 \cdot (n_A - 1)$ would overestimate the real driven mileage d_k in this case. For that reason, we later introduce a factor $\gamma(q_k)$ for correcting the term $d_0 \cdot (n_A - 1)$ dependent on the order size.

iii) Equation (1) is based on the assumption that the customer with order k is served first in the tour. As real data shows, this assumption leads to an overestimation of the transportation distance to the first delivery point in the tour. Hence for a good estimation, we later correct d_k with a function $\mu(q_k)$.

Generally the estimation model comprises 2 steps. The first step refers to item i), where based on the parameter $q_A(q'_k)$ the total truck load Q_k on the tour for delivering the order k is estimated (Section 3.1). As shown later in detail, the quotient of q'_k resp. q_k and Q_k is the base to assign a part of the tour length to a single order k. The second step is to estimate the tour length for delivering the order k and assign it to the regarded order k (Section 3.2). This refers mainly to the items ii) and iii).

3.1 Total truck load

The number of full truck loads for serving the order k is

$$a_k = \left\lfloor \frac{q_k'}{Q} \right\rfloor . \tag{1}$$

The relevant size q_k of the order k, which may be combined with other orders to a tour is given by equation (2), because full truck loads can not be combined, they are delivered directly. Thus, q_k takes the values $0 \leq q_k < Q$, where

$$q_k = q_k' - a_k \cdot Q . \tag{2}$$

If the carrier only combines orders, which belong to the regarded distribution system, Q_A becomes

$$Q_A = n_A \cdot \frac{\sum_{k=1}^{K} q_k'}{K + \sum_{k=1}^{K} a_k} = n_A \cdot \frac{\sum_{k=1}^{K} (q_k + a_k \cdot Q)}{K + \sum_{k=1}^{K} a_k} , \tag{3}$$

where the quotient represents the average size of an order delivered on a tour. A simple estimator for $q_A(q_k)$ may be given by

$$\frac{Q_A - q_k}{n_A - 1} . \tag{4}$$

This implies that the size of the combined orders is decreasing with an increasing size of the regarded order k. This assumption is reasonable considering that Q_A is a constant parameter given by the order structure. The quotient (4) has the negative effect that for $q_k > Q_A$ the result becomes negative, which is unreasonable (see Figure 3.3). For that reason, q_k is corrected by a factor $\delta(q_k)$ so that $q_A(q_k)$ is positive for every q_k:

$$q_A(q_k) = \frac{Q_A - \delta(q_k) q_k}{n_A - 1} \tag{5}$$

Subsequently we will only use equation (5) for determing q_A. Figure 3.3 shows the graph of q_A.

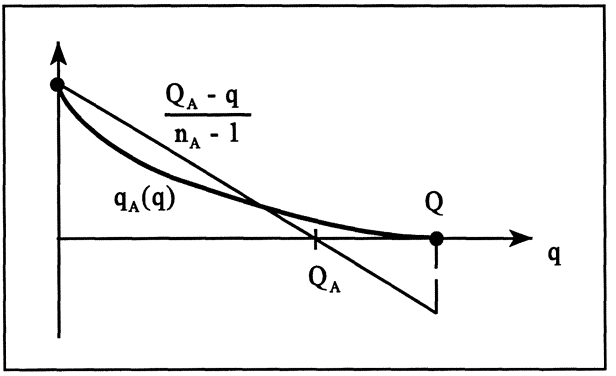


Fig. 3.3 Size of the combined orders

For the function type of the correction function $\delta(q)$ we choose a quadratic polynomial

$$\delta(q) = r_{q1} + r_{q2}q + r_{q3}q^2 \quad . \tag{6}$$

In principle the use of every other function type is possible (e.g. exponential functions), which leads to the desired monotonous and convex shape of $q_A(q)$ in the intervall $0 \leq q < Q$. The quadratic polymon has the advantage that the coefficients r_q can be calculated analytically in an easy way considering the subsequent 3 constraints:

$$\lim_{q \to Q} \delta(q) = \frac{Q_A}{Q} \tag{7}$$

$$\lim_{q \to Q} d\frac{q_A(q)}{dq} = 0 \tag{8}$$

$$\sum_{k=1}^{K} \left(Q_A - q_k + a_k \cdot (Q_A - Q)\right) = \sum_{k=1}^{K} \left(Q_A - \delta(q_k) q_k + a_k \cdot (Q_A - \lim_{q_k \to Q} \delta(q_k) \cdot Q) \right) \tag{9}$$

Equation (7) implies that $q_A(q)$ becomes 0 for $q \to Q$, because in this case the vehicle capacity is completely used and no other orders can be combined with q. Equation (8) expresses that the function $q_A(q)$ approximates this point horizontally. This ensures the continuity of the first derivation for $q = Q$. As tests with practical data had shown, it prevents $q_A(q)$ from the occurence of an unreasonable extreme point (minimum) in the intervall $0 \leq q < Q$. Equation (9) ensures that the estimation of $q_A(q)$ leads in average to the same value as without the factor for correction $\delta(q)$ (see equation (4)). Due to the 3 contraints $q_A(q)$ must have a convex shape (see Figure 3.3). Solving the equations (6) to (9) we obtain the coefficients:

$$r_{q1} = \frac{Q}{Q_A} - r_{q2}Q - r_{q3}Q^2 \tag{10}$$

$$r_{q2} = \frac{1}{Q}\left(-\frac{Q_A}{Q} - 2r_{q3}Q^2 \right) \tag{11}$$

$$r_{q3} = \frac{\sum_{k=1}^{K} q_k\left(1 + \frac{Q_A}{Q}\left(\frac{q_k}{Q} - 2 \right) \right) + \sum_{k=1}^{K} a_k(Q - Q_A)}{\sum_{k=1}^{K} q_k(Q - q_k)^2} \; . \tag{12}$$

The estimated total truck load Q_k of the tour for delivering the order k is

$$Q_k = q_k + (n_A - 1)\cdot q_A(q_k) = q_k(1 - \delta(q_k)) + Q_A \; . \tag{13}$$

As shown in Section 3.2, Q_k is used to assign a part of the total estimated tour length to a single order k delivered on a tour. The term $(n_A - 1) \cdot q_A(q_k) = Q_A - q_k \cdot \delta(q_k)$ represents the total size of the combined orders, with is only dependent on Q and Q_A and q_k. This shows that assumption e) at the beginning of Section 3 does not influence the approach.

The capacity constraint of the vehicle ($Q_k \leq Q$) is satisfied, because

$$\frac{Q - q_k}{n_A - 1} \geq \frac{Q_A - \delta(q_k)\cdot q_k}{n_A - 1} \tag{14}$$

is valid for every $Q_A \leq Q$ and $0 \leq q_k < Q$ and hence

$$Q \geq q_k\cdot(1 - \delta(q_k)) + Q_A = Q_k \; . \tag{15}$$

3.2 Assigned tour length

The items i) to iii) described in the beginning of Section 3 lead to the following estimation of the mileage d_k^z assigned to the delivery of an order k:

$$d_k^z = \frac{q_k}{Q_k}(d_k\cdot\mu(q_k) + d_0(n_A - 1)\cdot\gamma(q_k)) + a_k d_k \; , \tag{16}$$

where the Term $d_k\cdot\mu(q_k) + d_0(n_A - 1)\cdot\gamma(q_k)$ denotes the estimated tour length and the quotient q_k/Q_k is the base for assigning a part of the tour length to a single order k. The product $a_k \cdot d_k$ quantifies the distance driven in full truck loads for serving the order k. The total estimated mileage for the direct deliveries and the supply of the TP in a distribution network is

$$D^T = \sum_k^K d_k^z \,. \tag{17}$$

In order to obtain a correct estimation of the total mileage, the (mean) mileage to be assigned to an order must be

$$d_k^{z*} = \frac{1}{Q_A}\left(q_k + a_k Q\right)\cdot\left(d_k\cdot\mu_A + (n_A - 1)\cdot d_0\right), \tag{18}$$

where, μ_A denotes the factor for the average reduction of d_k for an estimation of the distance to the first customer in a tour. If a (small) sample T of real tours is known, which comprises |T| orders, μ_A can be estimated by the quotient of the average distance to the first customer in a tour d_I (obtained from the sample) and the average distance to every customer from the CW.

$$\mu_A = \frac{d_I}{\frac{1}{|T|}\sum_{k\in T}^K d_k} \tag{19}$$

If no sample is available, tests with real tour data have shown that the value $\mu_A = 0.97$ leads to good results. For correcting d_k we use similar to $\delta(q)$ the function type of a quadratic polynomial:

$$\mu(q) = r_{m1} + r_{m2}q + r_{m3}q^2 \tag{20}$$

With the concave function shape of $\mu(q)$ (see Figure 3.4), we take into account that with a growing order size q the probability for a combination with other orders decreases. This implies that the probability that a customer with order k is served first in the tour increases with an increasing order size. In principle other function types, which lead to the desired monotonous and concave shape in the intervall $0 \le q < Q$, may be used too. The polynomial has the advantage that the coefficients r_m can be determined analytically by the constraints (21), (22) and (23).

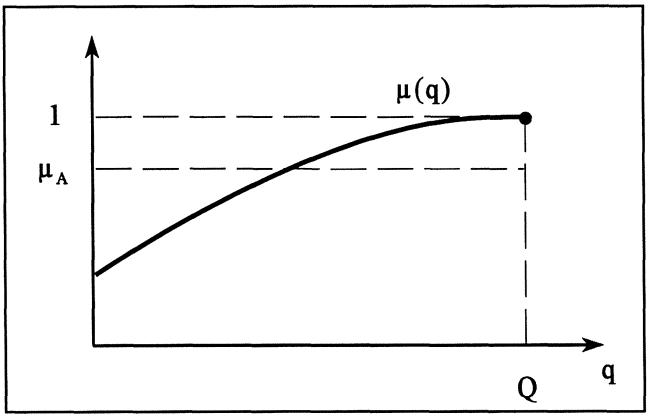


Fig. 3.4 Function for correcting d_k

$$\lim_{q \to Q} \mu(q) = 1 \tag{21}$$

$$\lim_{q \to Q} d\frac{\mu(q)}{dq} = 0 \tag{22}$$

$$\frac{\mu_A}{Q_A} \sum_{k=1}^{K} \left(q_k + a_k Q\right) \cdot d_k = \sum_{k=1}^{K} \left(\frac{q_k}{Q_k} \cdot d_k \cdot \mu(q_k) + a_k d_k \right) \tag{23}$$

Equation (21) implies that orders of a size Q can not be combined with other orders to a tour, because the vehicle capacity is already used. To serve such a customer, the distance d_k is driven. Equation (22) is analogeous to equation (8) and prevents $\mu(q)$ from the occurance of an exteme point (maximum) in the interval $0 \leq q < Q$. Equation (23) ensures that the sum of the (assigned) distances driven to the first customer in the estimated tours before and after the correction are identical (see the equations (16) and (18)). This guarantees for the right level of the estimated tour length in average. As tests with real data have shown, equation (23) leads to good estimates referring to the total mileage driven to the first customer. The solution of the equations (20) to (23) leads to the coefficients

$$r_{m1} = 1 - r_{m2} Q - r_{m3} Q^2 \tag{24}$$

$$r_{m2} = -2 r_{m3} Q \tag{25}$$

$$r_{m3} = \frac{\frac{\mu_A}{Q_A} \sum_{k=1}^{K} \left(q_k + a_k Q\right) \cdot d_k - \sum_{k=1}^{K} \left(\frac{q_k}{Q_k} + a_k \right) \cdot d_k}{\sum_{k=1}^{K} \frac{q_k}{Q_k} \cdot d_k \cdot (Q - q_k)^2} \quad . \tag{26}$$

Similar to $\mu(q)$, the function $\gamma(q)$ for correcting the distance between the customers in a tour considers the decreasing probability of finding orders for the combination to a tour, if the order size q increases. Analogeous to $\mu(q)$ it is expressed by a quadratic polynomial:

$$\gamma(q) = r_{d1} + r_{d2} q + r_{d3} q^2 \quad . \tag{27}$$

The coefficients r_d are calculated from the three equations

$$\lim_{q \to Q} \gamma(q) = 0 \tag{28}$$

$$\lim_{q \to Q} d\frac{\gamma(q)}{dq} = 0 \tag{29}$$

$$\sum_{k=1}^{K} \frac{1}{Q_A}\left(q_k + a_k Q\right)\cdot(n_A - 1)\cdot d_0 = \sum_{k=1}^{K} \frac{q_k}{Q_k}\cdot(n_A - 1)\cdot d_0 \cdot \gamma(q_k) \tag{30}$$

Equation (28) considers that it is impossible to combine an order with a size q = Q with other orders, because in this case the vehicle capacity is used completely. Then, only the mileage d_k is driven by the vehicle. Equation (29) has the similar meaning as (22) and (8). Analogeous to equation (23), equation (30) ensures that the sum of the assigned estimated distances between the customers are equal before and after the correction (see the equations (16) and (18)). The solution of equations (27) to (30) leads to the coefficients

$$r_{d1} = -r_{d2} Q - r_{d3} Q^2 \tag{31}$$

$$r_{d2} = -2 r_{d3} Q \tag{32}$$

$$r_{d3} = \frac{\sum_{k=1}^{K} \frac{1}{Q_A}\left(q_k + a_k Q\right)}{\sum_{k=1}^{K} \frac{q_k}{Q_k}\cdot(Q - q_k)^2} \quad . \tag{33}$$

Figure 3.5 shows for a given distance d_k the graphs of $\gamma(q)$ and d_k^z, which is illustrated standardsized by A · B. The parameter A denotes the estimated mileage of a tour related to the mean expected mileage. As tests with practical problems had shown, the shape of A is monotonous decreasing, because the influence of $\gamma(q)$ is dominating $\mu(q)$. The parameter B represents the part of the estimated mileage, which is assigned to the order k. The product of A and B results in a monotonous increasing function over q.

For short distances d_k (approx. $d_k < 1.5\ d_0$) the function A · B may have an extreme point. In this case the mileage assigned to an order q < Q would be greater than the mileage assigned to an order q = Q, which is not reasonable. This effect is caused by the correction functions, specially equations (21) and (28), and the assumption that all orders q < Q are combined with other orders to a tour. But in reality, orders of size q < Q of customers located close to the CW (approx. $d_k < 1.5\ d_0$), are rather delivered without combination by a vehicle with an appropriate capacity. Therefore, we modify equation (16) as follows and rule out an extreme point.

$$d_k^z = \frac{q_k}{Q_k} \min\left(d_k\ ,\ d_k \cdot \mu(q_k) + d_0 (n_A - 1)\cdot\gamma(q_k)\right) + a_k d_k \tag{34}$$

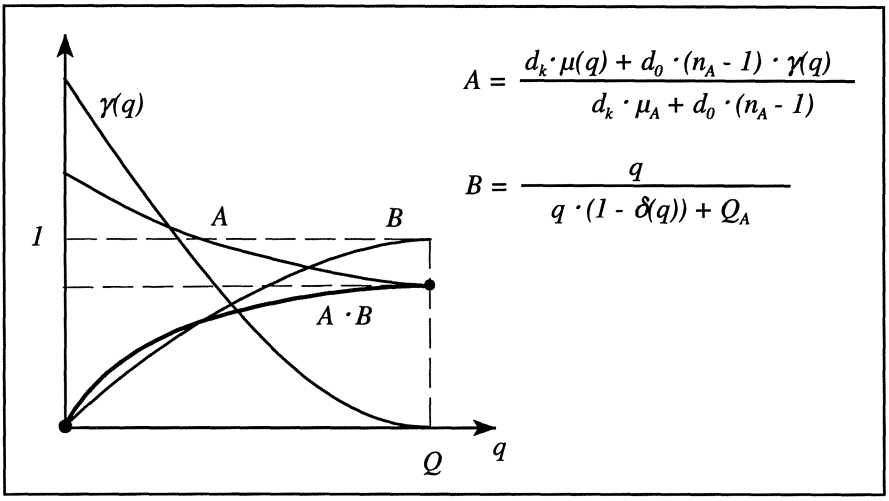


Fig. 3.5 Shape of γ(q) and the assigned distance A · B

4. Computational results

We tested the model with real order data of a consumer goods shipper. The data consists of real trunk tours in Germany for a period of 4 days. Starting point of the tours is a CW located in the south of Germany. Table 1 shows the results, where the above model (equation (34)) is called 'model 1'.

To test the efficacy of the correction functions δ, μ and γ we further introduce a 'model 2'

$$d_k^{z*} = \frac{1}{Q_A}\left(q_k + a_k Q\right)\cdot\left(d_k + (n_A - 1)\cdot d_0\right), \tag{35}$$

which is a simplified version of model 1, considering no correction functions. In addition we show the results for an estimation by direct distances d_k (see Figure 3.1).

In Table 1 the real mileage is compared with the estimated mileage of the model. The absolute deviation denotes the deviation of the total estimated mileage from the real mileage related to the real mileage. Mostly, the errors for model 1 are less than 1 %, for model 2 about 3 %. The last column of Table 1 shows that the calculation by direct distances leads to very high overestimations for the tour length. The mean absolute deviation (MAD) is calculated by

$$MAD = \frac{\frac{1}{T}\sum_{t=1}^{T} | d_t^z - d_t^r |}{\frac{1}{T}\sum_{t=1}^{T} d_t^r}, \tag{36}$$

where T denotes the number of tours in the period, d_t^r the real length of tour t and

$$d_t^z = \sum_{k \in t} d_k^z \,. \tag{37}$$

Table 4.1 Numerical results

period (day)	no. of tours	no. of orders	mileage real data [km]	**model 1**				model 2		direct distance
				mileage [km]	**total dev. [%]**	**no. of best results [%]**	**MAD [%]**	total dev. [%]	MAD [%]	total dev. [%]
1	45	81	14,806	**14,735**	**-0.48**	**66.7**	**11.1**	2.75	21.6	61,8
2	57	127	22,341	**22,612**	**1.21**	**64.9**	**16.3**	2.73	21.8	83,6
3	26	52	8,156	**8,226**	**0.85**	**71.3**	**13.7**	5.13	22.7	67,6
4	56	111	20,660	**20,500**	**-0.78**	**67.9**	**13.1**	3.04	18.8	77,7
1-4	184	371	65,963	**66,066**	**0.16**	**71.2**	**14.2**	3.32	21.4	74,9

The results show that the MAD of model 1 is for the most part less than 15 %. The estimation with model 2 leads to a MAD mostly greater than 20 %. We further compare our models with respect to the accuracy of the estimated tour length. Comparing the accuracy between the real mileage of the single tours and the estimated mileage for each tour, model 1 leads to better results than model 2 (see the column 'no. of best results' in Table 1). For instance the value in the last line signifies that in 71.2 % the estimation for the tour length of model 1 is better than the calculation by model 2. The comparison of the estimated tour length with real tour data shows that our model 1 leads to very precise estimations referring to the total mileage and the length of a single tour. The results of model 1 are much better for every considered period (day) than the estimations by the simplified model 2.

Figure 4.1 shows the deviation of the estimated tour length (model 1) from the real tour length

$$AD_t = \frac{d_t^z - d_t^r}{d_t^r} \tag{38}$$

over the total truck load on the tour

$$Q_t = \sum_{k \in t} q_k \,. \tag{39}$$

Each point in the diagram represents a tour. The results show, that the deviation is mostly less than 20 %. Database for Figure 4.1 are all 184 tours of Table 1. Further we considered a truck capacity of Q = 18,400 kg, an average truck load of q_A = 14,600 kg, an average number of customers on a tour of n_A = 2.02 and an average distance between subsequent customers of d_0 = 56.9 km. This is also the database for all the further investigations.

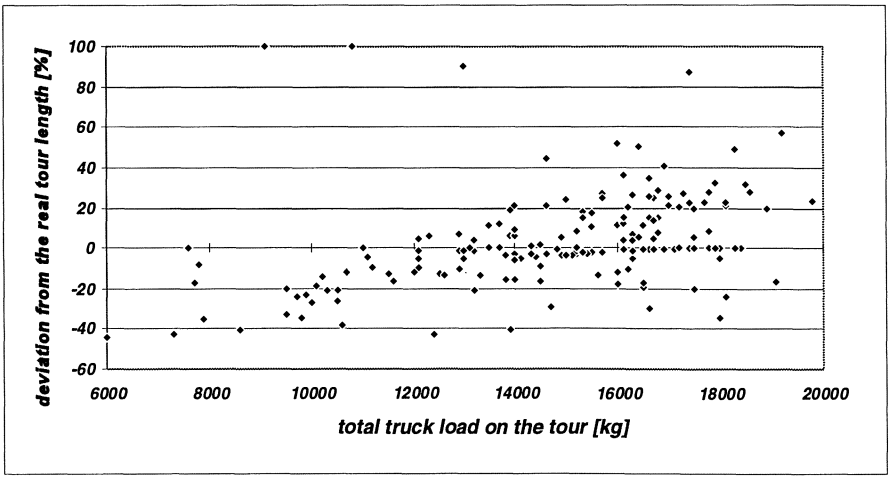


Fig. 4.1 Deviation of model 1 from the real tour length

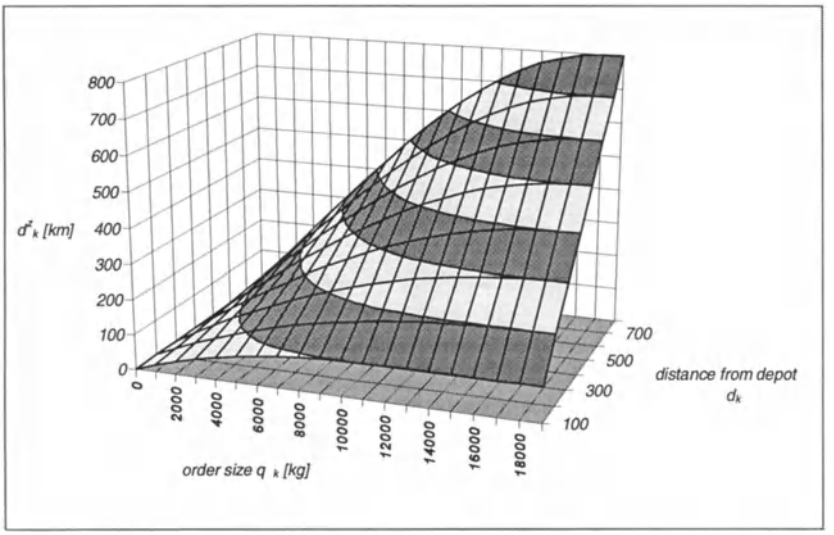


Fig. 4.2 Assigned tour length

Figure 4.2 illustrates the shape of the assigned tour length d_k^z (equation (34)) over the order size q_k and the distance d_k of the customer location from the CW. We obtain a monotonous increasing function. For instance the assigned distance to an order with the size q_k = 5,000 kg located in the distance of d_k = 300 km from the CW is d_k^z = 140.7 km.

Based on equation (34) the vehicle miles for the trunk tours in a distribution system can be calculated. Therefore, the correction functions δ, μ and γ have to be determined once, and then for calculating the (assigned) distances only the distance d_k from CW and the order size q_k of each order are needed. The mileage is an important meassure for the environmental evaluation of the traffic, because fuel-, energy-consumption and emissions behave nearly proportionally to the vehicle miles. The equation (34) builds the foundation for the environmental evaluation of the traffic in distribtuion systems, which is explained in the next Section. In Section 6 it is shown that our estimation model may also be used for the calculation of transportation tariffs.

5. Environmental keyfigures

As already discussed in Section 1, the environmental impact of transportation is strongly dependent on the mileage. Thus, our estimation model may build the base for calculating ecological measures quantifying the environmental impact of transportation. In the context of transportation, important ecological measures are the energy- and the fuel-consumption such as emissions (CO , CO_2, NO_x, SO_2, H_yC_z and particles). These ecological measures are calculated as follows, using the notations:

u index for the environmental measure (energy, fuel, CO , CO_2, NO_x, SO_2, H_yC_z and particles)

l index for the vehicle class

v (average) vehicle speed [km/h]

e_{ul} environmental factor for measure u and vehicle class l [kWh/km], [l/km], [g/km] (dependent on speed and vehicle load)

$$e_{ku}^{z} = \frac{q_k}{Q_k}\left(d_k \cdot e_{ul}(v, Q_k) \cdot \mu(q_k) + d_0 \cdot (n_A - 1) \cdot e_{ul}\left(v, \frac{n_A}{2} q_A(q_k) \right) \cdot \gamma(q_k) \right) + a_k \cdot d_k \cdot e_{ul}(v, Q) \quad (40)$$

The result of equation (40) represents the energy-, fuel-consumption or the emissions, which can be assigned to the delivery of a single order k. The environmental factor e_{ul} for the determination of e^{z}_{ku} is based on data of the German authority of environment (Umweltbundesamt) (see Hassel (1983); Hassel (1995); INFRAS AG (1995)). If we sum e^{z}_{ku} over all orders k = {1, ..., K}, the total amount of energy- and fuel-consumption such as emissions may be calculated for the trunk tours in a distribution network.

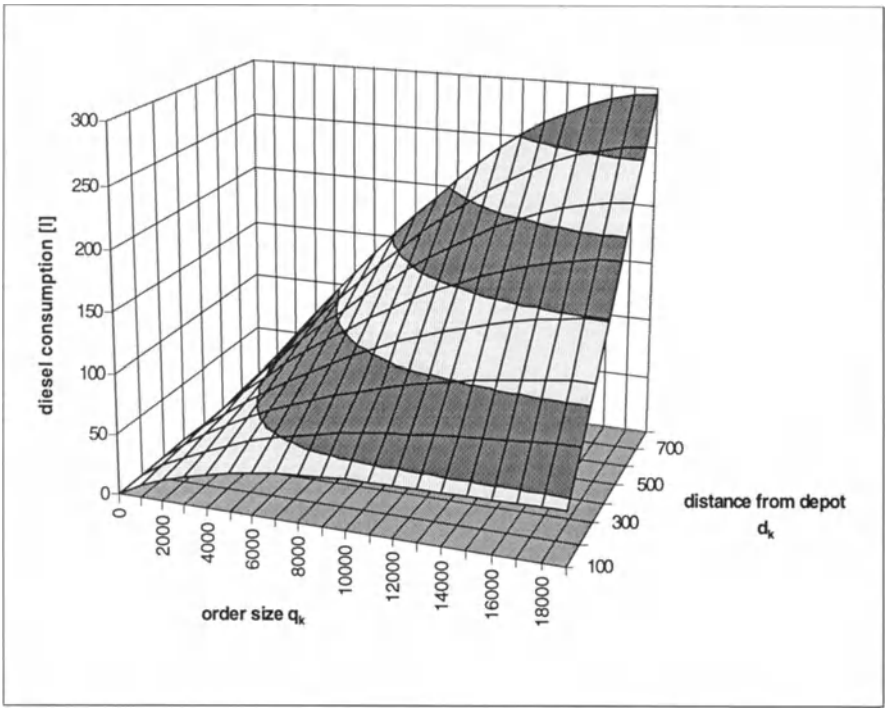


Fig. 5.1 Fuel consumption

Figure 5.1 shows the fuel consumption $e^z_{kDiesel}(q_k, d_k)$ assigned to an order q_k located in the distance d_k from CW. We considered a vehicle-type of 40 tons allowed total truck load, which is usually used in the German trunk haulage. Further the (average) vehicle speed used in equation (40) variies dependent from the distance d_k between 35 und 80 km/h, in a way that with an increasing distance the (average) vehicle speed increases too. Analogeous to Figure 4.2 we obtain a monotonous increasing function $e^z_{kDiesel}(q_k, d_k)$. For instance, if we evaluate the environmental impact by the fuel consuption for a delivery of an order with size q_k = 5,000 kg located in the distance of d_k = 300 km from the CW, the assigned consumption $e^z_{kDiesel}$(5,000 kg; 300 km) is 53.7 l. In the same way the energy-consumption and the emissions for the transportation of each order can be calculated.

6. Transportation tariff

Moreover, our model can be used for calculating transportation tariffs, which is a very important question in Germany since the deregulation of tariffs (see Fleischmann (1997)). In principle transportation costs may be calculated dependent on the order size q_k and the distance d_k of the customer from the CW by

$$c(q_k, d_k) = c_d \cdot d_k^z + c_t \cdot \left(\frac{d_k^z}{v} + s(q_k) \right) \tag{41}$$

where

q_k order size [kg]
d_k^z assigned distance to the customer with order k [km]
c_d vehicle cost per km [DM/km]
c_t vehicle cost per hour [DM/h]
v (average) vehicle speed [km/h]
s stop time at the customer (dependent on the order size) [h].

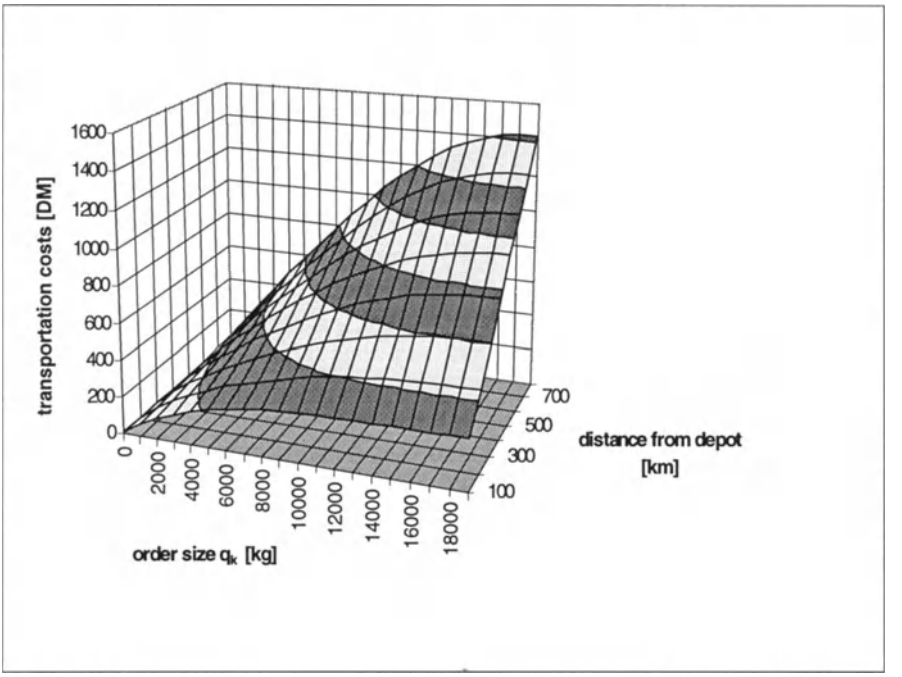


Fig. 6.1 Transportation costs

With $c_d = 0.70$ DM/km and $c_t = 75$ DM/h we obtain a tariff, which is degressive in the order size and the distance, illustrated by Figure 6.1. In this context degressive means that the cost rates decreases with an increasing order size q_k resp. distance d_k. Based on these assumptions the tariff for transportation of an order with the size $q_k = 5{,}000$ kg located in the distance of $d_k = 300$ km from the CW is $c(q_k, d_k) = 304.26$ DM.

7. Conclusions

For analysing the environmental impact of distribution systems the calculation of the mileage is important, because fuel and energy consumption such as emissions behave nearly proportional to this measure. An essential part of the mileage in distribution systems is driven in trunk tours, which length has to be estimated in the framework of strategical distribution network design. The model we presented shows good results for the estimated tour length. This model can build the foundation for an environmental evaluation of the trunk traffic in distribution systems. It can be used for calculating fuel and energy consumption such as emissions. In addition it can also be used for the calculation of transportation tariffs dependend on the order size and the distance.

References

Bahn AG (1995): Masterplan II. Frankfurt

Bundesamt für Güterverkehr (1993): Fernverkehr deutscher Lastkraftfahrzeuge 1993. Köln

Bundesminister für Verkehr (1995): Verkehr in Zahlen. Bonn

CORINAIR (1993): Commission of the European Communities/CORINAIR Working Group on Emission Factors for Calculating 1990 Emissions from Road Traffic, Volume 1, "Methodology and Emission Factors". Final Report, Brüssel

Daganzo, C. F. (1991): Logistics Systems Analysis, Lecture Notes in Economics and Mathematical Systems. (Springer) Berlin et al.

Eckstein, W. E. (1995): Ökonomische Folgen von Güterverkehrszentren. in: Bayrisch Schwäbische Wirtschaft 9, 24-26

Fleischmann, B. (1979): Distributionsplanung. in: Proceedings in Operations Research 8, (Springer) Würzburg, Wien, 293-308

Fleischmann, B. (1993): Designing distribution systems with transport economies of scale. in: European Journal of Operational Research 70, 31-42.

Fleischmann, B. (1995): Design of Freight Traffic Networks: Models, Methods and Applications. SVOR/ASRO Tutorial on New OR Technologies for Transportation and Facility Location, Thun October 12-13

Fleischmann, B. (1997): Design of Freight Traffic Networks. in this volume

Hassel, D. et al. (1983): Das Abgas- und Emissionsverhalten von Nutzfahrzeugen in der Bundesrepublik Deutschland im Bezugsjahr 1980. Forschungsbericht des TÜV-Rheinland. Umweltbundesamt(UBA)-Berichte 11/83, (Erich Schmidt) Berlin

Hassel, D. et al. (1995): Abgasemissionsfaktoren von Nutzfahrzeugen in der Bundesrepublik Deutschland für das Bezugsjahr 1990, Umweltforschungsplan Nr. 104 05 151/02, im Auftrag des Umweltbundesamtes, Berlin

Hautau, H. (1995): Gewichtige Argumente. in: Logistik Heute 8, (Huss) München, 40

Heinrich, J. (1995): Güterverkehrszentren verringern die Umweltbelastung. VDI (Verein Deutscher Ingenieure) - Nachrichten, Juni 9, Nr. 23, 14

Ihde, G. B. / Eckart, D. / Stieglitz, A. (1994): Möglichkeiten und Probleme einer umweltorientierten Konsumgüterdistribution. in: Marketing 3, 199-208

INFRAS AG (1995): Handbuch für Emissionsfaktoren des Straßenverkehrs. Bern

Klotz, H. (1992): Öko-Distribution wird zum wichtigen Wettbewerbsfaktor. in: Deutsche Verkehrszeitung, Nr. 86, Juli 21, 3

Kunz, D. (1977): Untersuchungen über den Einfluß der Struktur von Warenverteilungsnetzen auf die Distributionskosten. (Westdeutscher) Opladen

Konen, W. (1985): Kennzahlen in der Distribution. (Springer) Berlin et al.

Konen, W. (1982): Analyse und Reorganisation von Distributionssystemen. in: Baumgarten H. et al. (eds.), RKW-Handbuch Logistik, Nr. 7340, (Erich Schmidt) Berlin

Köhler, U. (1995): Traffic and transport planning in German Cities. in: Transportation research 29A, no. 4, 253-261

Kraus, S. (1996): Ökologische Bewertung von Gütertransporten und deren Anwendung in Distributionsnetzen. in: Keller, H. B./Grützner, R./Hohmann, R. (Hrsg.): Tagungsband zum 6. Treffen des Arbeitskreises "Werkzeuge für Simulation und Modellbildung in Umweltanwendungen" der GI Fachgruppe 4.6.1. Wissenschaftliche Berichte des Forschungszentrums Karlsruhe, FZKA 5829, Karlsruhe, 178-198.

Kreitmair, G. / Kraus, S. (1995): Mehr als nur gute PR. in: Logistik Heute 8, (Huss) München, 8

Krieger, W. (1984): Computergestützte Auswahl interkontinentaler Distributionsverfahren. (Dunker & Humbolt) Berlin

Newell, G. F. / Daganzo, C. F. (1986): Design of multiple-vehicle delivery tours-I. A ring-radial network. in: Transportation research 20B, no. 5, 345-363

OECD (1994): Environmental impact assessment of roads. Road transport research, Paris

Ohseok, K. / Golden, B. / Wasil, E. (1995): Estimating the length of the optimal TSP: An tour empirical study using regression and neural networks. in: Computers Operations Research 22, no. 10, 1039-1046

Paraschis, I. N. (1989): Optimale Gestaltung von Mehrprodukt-Distributionssystemen. (Physica) Heidelberg

Spelthahn, S. / Schlossberger, U. / Steger, U. (1993): Umweltbewußtes Transportmanagement. (Paul Haupt) Bern, Stuttgart, Wien

Tüshaus, U. / Wittmann, S. (1997): Strategic Logistic Planning by Means of Simple Plant Location: A Case Study. in this volume

Vehicle Routing and Scheduling for Trunk Haulage

Petra Stumpf

Universität Augsburg, Lehrstuhl für Produktion und Logistik, D-86135 Augsburg

Summary. The paper deals with the daily control of the vehicles in a network of less-than-truckload carriers. The operations in such a network consist of the regional pick-up and delivery traffic around the depots, regular line service between the depots, possibly through one or several hubs, and additional transports of partial loads between customers. The regional traffic planning, which is closely related to the classical vehicle scheduling problem, is not considered here. However, scheduling the trunk haulage tours is a quite different problem which has found little attention in literature, so far. We analyze various problem settings encountered in practice, present a new heuristic algorithm for a basic problem and report on computational tests with real-life data.

1. Introduction

Since the beginning of the nineties the German transportation business, which is characterized especially by small and medium companies, is confronted with a lot of new challenges. The deregulation of freight traffic markets, which came simultaneously with the beginning of the Common Market, resulted in a repeal of the political reglementations of quantities and prices which suppressed a real competition in the past.

Simultaneously the liberalization of regulations enabled the domestic as well as the foreign competitors to enter easily into the German traffic market. The advance of foreign competitors in combination with the release of tariffs caused an aggressive fight in the carrier sector about the new partition of the market, and consequently a drastic decay of freight charges of up to 50% could be recognized on the transportation market.

Therefore many, especially small and medium carrier companies are acutely endangered in their existence. In order to compete against big carrier companies and foreign competitors they consistently and continuously have to improve their service structures as well as their cost structures.

When rethinking the service offer the orientation towards the wishes of customers becomes more and more important. In recent years the loaders demand, besides a high service quality (punctuality, reliability, flexibility), also individual logistic concepts, which surpass the actual transport of goods. Therefore the carriers have to offer more and more complete logistic services. Furthermore, small and medium carrier companies have to cooperate with other carriers and to operate on the transport market as a unique alliance, if they want to profit from the liberalization of traffic markets. Thus they

can offer their services covering all of Germany or Europe and and therefore reduce their size disadvantage compared to big carrier companies.

Next to the improvement of services a reasonable composition of prices is the second prerequisite for being accepted on the transport market. Endangered carriers have to calculate their prices cost oriented if they want to secure their existence. The low price level, which is caused by the surplus of transport capacity and hence the tightened competition, can only be kept stable permanently if the intern organization is improved continuously and possible cost reducing potentials are used consequently.

For this reason in the last years intensive efforts can be recognized especially by the threatened medium-sized companies to process the transports more efficiently and reliably. Because of the complex problems the companies are often dependent on external support. Considering the **operations of those carriers transporting piece goods and partial loads** the following strategic, tactical and operational problems can be identified:
The strategic to medium-term problem of designing freight traffic networks (see Fleischmann (1997), Wlcek (1997)) is especially important in times of continuously enlarging transport markets and consists of the following decisions:

- Selection of suitable locations for consolidation points (depots, hubs)
- Assignment of customers to sending and receiving depots
- Routing of depot-depot transports ("line-haul shipments") in trunk haulage, i.e. the determination of shipment paths between depots

Given the results of the strategic optimization, the short-term processing of freight transports has to be planned (see AIF (1996)). The short-term planning of transports can be divided into the following subproblems:

- Planning of pickup and delivery of piece goods
  Essential elements are the rather medium-term planning of standard tour areas for pickup and delivery from and to one depot. Within the scope of daily disposition the tour areas are adjusted to the actual transport quantities on this day. If necessary, i.e. if the standard tours are infeasible with respect to capacity or time, the areas have to be changed slightly.
- Disposition of vehicles for trunk haulage
  Considering the daily planning of trunk haulage tours, an overlapping between piece goods and partial loads transport can be noted. Indeed most carriers would like to transport partial loads and piece goods completely separate. Often this is impossible, however, due to the low amount of shipments in both sectors. Because of capacity and cost aspects the joint transport is therefore necessary.
  A dispatcher of trunk haulage therefore has to combine partial loads and consolidated piece goods (line-haul shipments) to tours. Furthermore he has to assign the tours and full loads to vehicles.

Within the scope of this paper we will focus on the problem of **daily disposition of trunk haulage**. Whereas a multiplicity of solution procedures, literature and implementations is available for traditional vehicle routing problems, there is a research deficit concerning special problems of trunk haulage planning. In practice, very few decision supporting systems are installed. Therefore, the disposition of trunk haulage tours and trucks is often done manually without or with only little computer support.

The problem of vehicle routing and scheduling of trunk haulage is part of a recent research project on the disposition of the vehicles in a carrier network, supported by AIF (Arbeitsgemeinschaft industrieller Forschungsvereinigungen e. V.). In this project which included several carrier companies as partners, the focus was first on identifying the planning and control problems arising in practice and then on working out appropriate solutions. Accordingly, this paper starts with a rather detailed description of these practical problems (Section 2.), because they are new and differ remarkably from the well-known vehicle scheduling problems. Then we give an overview over the available literature in Section 3.. Subsequently, we formulate a model for the basic problem of trunk haulage planning (Section 4.), for which in Section 5. a starting heuristic and several strategies of an improvement heuristic are presented. In Section 6. we report on computational tests with practical data. Finally, Section 7. states the conclusions.

2. Problem

In this section the processes of partial loads and piece goods transport are described first. The problems of trunk haulage disposition can be derived therefrom.

Piece goods transport

It is not economical to transport shipments of small shipment size, so-called piece goods shipments, directly from their origins to their destinations. Therefore, they are picked up with vans in local tours and brought to their corresponding sending depot. At sending depots the piece goods shipments are consolidated to line-haul shipments. Subsequently they are transported to the corresponding receiving depots. The shipments are either transported directly between the sending and receiving depots or are newly consolidated at hubs or regional hubs. At the receiving depots the consolidated shipments are resolved. Afterwards in local tours (again with vans) the shipments are delivered to their destinations.

Considering the trunk haulage planning, the transport of piece goods is relevant only insofar, as the consolidated line-haul shipments have to be transported between the depots with trucks planned by the trunk haulage

dispatcher. Additionally, partial loads are usually transported together with line-haul shipments in order to increase the use of capacity of the trucks.

Transport of partial loads

Shipments with greater shipment size (partial loads) are not picked up and delivered by vans like piece goods shipments, due to the high pickup and delivery costs and the time and cost intensive transshipments at depots. Partial loads are generally picked up and delivered by trucks. Because one partial load normally does not use the full capacity of a truck, up to six partial loads are combined to one tour. Furthermore, in order to increase the use of capacity, line-haul shipments can be loaded onto trucks which transport partial loads. Also partial loads can be, but do not have to be consolidated and break bulked at depots. In practice one can identify the following transportation modes for the partial loads transport:

1. Direct pickup and delivery (see Figure 2.1)
The partial load is picked up and delivered by this truck which also performs the long distance route in trunk haulage. The partial load therefore does not change the truck between its origin and destination. All full loads are transported in this mode.

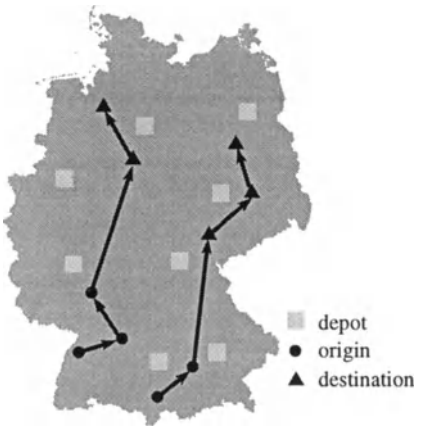


Fig. 2.1. Direct pickup and direct delivery

2. Indirect pickup, direct delivery (see Figure 2.2)
This mode often is selected if the transports are strongly destination focused, e.g. in the car supply industry. The partial load is picked up from a so-called carrier vehicle which brings it to the corresponding sending depot. The partial load is then loaded onto a truck, which directly delivers the partial load to its destination after the long distance transport.

Carrier vehicles are trucks, which operate during the night in trunk haulage. During the day they pickup partial loads and bring them to the depot or they deliver partial loads from the receiving depot to their destinations.

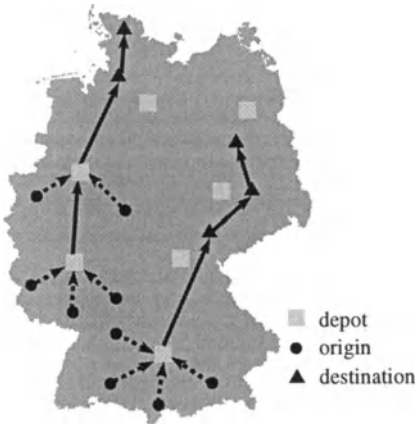


Fig. 2.2. Indirect pickup and direct delivery

3. Direct pickup, indirect delivery (see Figure 2.3)
The partial load is picked up by the same truck, which after picking up other shipments performs the long distance tour also. However, the truck does not directly deliver the partial load to its destination but only to its corresponding receiving depot. From this depot the delivery of the partial load is done by a carrier vehicle.

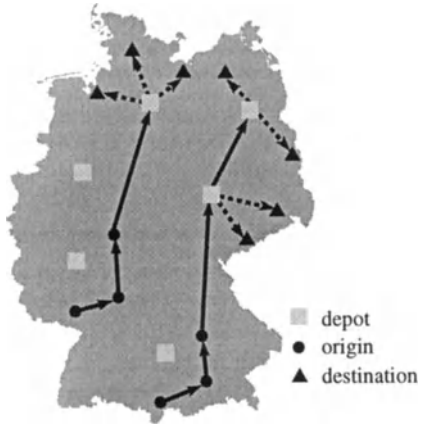


Fig. 2.3. Direct pickup and indirect delivery

4. Indirect pickup and delivery (see Figure 2.4)
In this transportation mode the pickup as well as the delivery is processed by carrier vehicles. Therefore, partial loads have to be handled twice in depots. A further transshipment of consolidated partial loads, e.g. at a hub, is not possible (contrary to the piece goods transport).

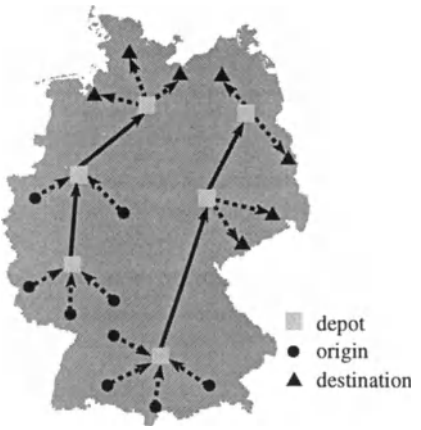


Fig. 2.4. Indirect pickup and indirect delivery

The so-called "broken" piece goods transport and the broken partial loads transport differ on one hand by the different vehicles, which pickup and deliver the shipments. On the other hand the handling of partial loads is essentially easier and faster than the handling of piece goods: Due to the larger shipment size, the partial load transshipment often consists only of a change of loading units, so that the partial load changes its vehicle and driver, but not the loading unit. Furthermore, it is possible, that a partial load stays on the same truck, which changes from carrier traffic to trunk haulage or vice versa. In this case the partial load will not be handled at all. Finally, because of the small number of partial loads per truck the organizational effort of handling is essentially smaller than when transshipping piece goods.

The different modes in the partial loads transport could be optionally combined by the trunk haulage dispatcher. For instance a trunk tour, which contains several directly picked up or delivered partial loads, can have one or more depot loading locations. At the depots line-haul shipments or indirectly picked up or delivered partial loads can be loaded or unloaded.

Functions of the short term disposition of trucks

The functions of the short term disposition of trucks are varying in different carrier companies. However, the following essential main subjects can be identified:

- *Determination of transportation mode*
  For each order arriving at the the short or trunk haulage disposition either per mail, telephone or telefax, it has to be decided whether it is treated as piece goods shipment or as a partial load.
  In practice, the demarcation of partial loads and piece goods shipments is handled as follows: "Clear" piece goods shipments, e.g. shipments with a relatively small shipment size, normally are processed automatically by the dispatcher for short-haul traffic. Only for borderline cases and shipments which come in directly at the trunk haulage department, the dispatcher has to decide whether to assign them to piece goods or to partial loads.
  All piece goods shipments are picked up and delivered by vans via short-haul tours. The planning of short-haul tours is done by the corresponding department. At the depot the consolidated shipments are assigned to line-haul shipments. These line-haul shipments have to be considered when planning the trunk haulage tours. If a shipment is treated as a partial load, it has additionally to be decided upon the transportation mode of this partial load (see Figures 2.1 - 2.4).
- *Construction of trunk haulage tours*
  The dispatcher has to combine the known and expected partial loads and line-haul shipments to tours. On a tour usually up to six partial loads or line-haul shipments, respectively, are loaded. In order to estimate the temporal course of a tour, the sequence of loading locations has to be determined as well as the planned arrival times per each loading location. When planning the tours the dispatcher usually considers the daily unchanged line-haul shipments first. The transport of line-haul shipments usually is organized according to a schedule, because the cooperation between several depots of different carrier companies is necessary. The line-haul shipments are daily transported in the same tours, but it is possible to transport additional partial loads on these tours in order to increase the use of capacity.
- *Procurement of trucks*
  Subsequently, the constructed tours as well as the full loads have to be assigned to trucks. There exist several possibilities to obtain suitable trucks. If the carrier company uses its own fleet of trucks, usually the "best" tours are assigned to its own trucks first. Tours are good, if they are well used with respect to capacity. On the other hand they should yield a high freight price. Also such tours are favored by its own trucks, for which a connecting or back load exists for the following day.
  As an alternative to own trucks, it is also possible to book complete subcontractor trucks at the transport market for one or several tours.
  Shipments which cannot be inserted efficiently in one of the planned tours can be individually placed to another carrier which has a tour of its own, in which the shipment fits respecting the direction, the time and the capacity.

However, the purchase of a small loading space is chosen rarely because of the higher costs.

- *Other functions*
  Besides the above described main functions trunk haulage dispatchers have other functions in the praxis, e.g. the planning of carrier vehicles, the administration of loading units or the employment planning for own drivers, which are not discussed in detail in this paper.

When combining shipments to tours, several time and capacity restrictions have to be considered:

- Consideration of truck capacities (weight, loading meters, pallets)
- Consideration of truck starting locations and possible operating times
- Adherence to a maximum duration of tours and a maximum driving time
  Concerning legal constraints, subcontractors are responsible on their own to keep the maximum operating and driving times, if they take over a tour. However, if the dispatcher plans tours with a maximum driving time of 10 or 11 hours, he creates the precondition, that the tours can be managed by one driver. For the own fleet of trucks and drivers the legal driving and recreation times have to be planned more precisely. In this case the dispatcher has to supervise the operations and the rest of the drivers during the last days.
- Consideration of customer time windows
- Consideration of time restrictions resulting from the medium-term fixed schedule for piece goods transport

Usually the shipment orders which come into the trunk haulage department are not rejected. Therefore, the disposition of tours and trucks has the objective to transport all shipments at minimal cost from their origins to their destinations. As on the transport market the prices for trucks fluctuate dependent on the weekday and on the season, the dispatcher has to overview the offer of loading space and the corresponding prices daily. Only then he can assign each tour to its best suitable, cheapest truck.

Temporal course of trunk haulage disposition

The temporal course when dispatching trunk haulage tours can be described as follows:

The planning of trunk haulage tours usually takes place in the early morning for shipments, which are picked up in the afternoon and delivered to their destinations on the following morning. Only few partial loads are declared the day before. Rather the shipments come in during the night and mostly during the morning via telephone, telefax or mail. In the morning, the dispatcher has to begin with planning even though not all shipments of the day are available. Usually also there exist only estimations for the line-haul shipments. Therefore, on basis of the presently known and expected shipments

the tours are planned in a dynamic process. If new shipments come in, the old plan possibly will be discarded. However, already during the planning process tours can be released, e.g. the corresponding trucks begin their tours with loading at the origins. These released tours thereby are fixed.

All in all the trunk haulage disposition of shipments is a dynamic process. With the beginning of planning, trunk haulage tours are constructed, which can be updated or enlarged if new information comes in (concept of rolling horizon).

This dynamic planning can be extended to several days. Indeed the main function of the dispatcher consists in the disposition of partial loads for the present day. But considering cost aspects it is reasonable to plan the present tours so that during the following day suitable connecting/back loads exist. Ideally, for company owned trucks complete vehicle round trips should be planned. However, as mentioned above, only a small percentage of the shipments is usually known for the following days. Therefore it is not possible to plan and fix all tours for more than one day in advance.

3. Literature Review

The problem of trunk haulage planning became more and more important in the past few years. Caused by the hard competition many carrier companies are faced with an increasing cost pressure. As the dispatchers of trunk haulage have to decide every day about expenses of several thousand marks, the companies try to better control and possibly reduce the costs in this sector. Cost reducing potentials are expected from a centralized disposition for several depots. Furthermore, the quality of disposition should be independent from the dispatcher, if possible.

Though the improvement of trunk haulage planning is an important problem in the practice, this topic has found little attention in the literature up to now. For traditional one-depot vehicle routing problems a multiplicity of literature is available. In contrast to these problems, within the scope of trunk haulage planning point-to-point-shipments have to be planned, for which, in addition, the transportation mode has to be decided upon. Furthermore, the tours do not have to be related to one or more depots. However, it is possible to take ideas and concepts from procedures for the traditional one-depot-vehicle routing problem or its extensions, respectively. For instance, special trunk haulage problems, like the planning of vehicle round trips when considering full loads, can be modeled as vehicle routing problem (VRP) (see Schmidt (1994)). Furthermore, there exists a close relationship between the combination of partial loads to tours and special pickup&delivery- as well as dial-a-ride problems, which also consider point-to-point-shipments (see Dumas et al. (1991), Ioachim et al. (1995), Savelsbergh et al. (1995)).

In the following we concentrate on papers which especially treat the disposition of tours and vehicles in trunk haulage. These are only a few papers, the

most of which are in German. The reason seems to be that shipping **partial loads** is characteristic for the German transportation market. Considering the international (and especially the US) literature, in almost all cases there is only a difference between the truckload (TL) or full loads transport and the less-than-truckload (LTL) transport.

Powell (1987, 1996), Dejax/Crainic (1987) and Rego et al. (1995) deal with the problem of (dynamic) allocation of truckloads to trucks. It has to be decided whether to accept arriving truckload orders or not. Accepted truckloads have to be assigned to trucks and trucks which have no load at the moment, either have to wait at their current location or are repositioned to neighbouring regions. The proposed models either are deterministic or contain - in case of uncertain demand - stochastic elements.

The transportation of LTL-shipments described e.g. by Powell (1983, 1986) and Leung et al. (1990) is similar to the piece goods transport described in Section 2.. Therefore, the main problems are either strategic, e.g. the location of hubs or depots, or tactical like the problem of organizing the transport of line-haul shipments. The pickup and delivery of single shipments is not considered but only the already consolidated line-haul shipments. A truck which is used for the long distance transport does not visit multiple origins and destinations on a single trip.

Feige (1983) considers the transport coordination of point-to-point transport orders. By combination of single transport orders the transport distance has to be minimized. The problem of transport coordination consists of two subproblems, which are closely related to each other. The function of the **parallel coordination problem** is to combine transport orders, which can be transported together on a truck, to trips so that all time windows for pickup and delivery are kept and a maximum reduction of vehicle kilometers can be reached. Within the scope of **sequence coordination**, trips which can be transported one after the other are combined to tours and assigned to vehicles, so that the empty kilometers of the vehicles are minimized. Indeed the vehicles have identical capacities, but different starting locations for each vehicle have to be considered. Ideally complete vehicle round trips should be planned.

The parallel coordination problem is similar to the problem of planning trunk tours described in Section 2.. If the planning horizon is extended to two or three days, the problem of planning vehicle round trips for own vehicles approximately corresponds to the sequence coordination problem.

Feige points out, that for the exact solution of the transport coordination problem both subproblems have to be considered simultaneously. However, because this problem is mathematically too complex, both subproblems are solved one after the other. For the first problem, an insertion procedure is

used. For the second problem which is modeled as an assignment problem two regret heuristics as well as an exact solution procedure are presented.

Also Söllig (1989) deals with the transport coordination within the scope of disposition of trucks. In contrast to Feige (1983) beside own trucks also subcontractor trucks are available. Given transport orders have to be combined to tours and the tours have to be assigned to trucks, so that the resulting total transportation costs are minimal. The tour costs are the kilometer dependent variable costs for own trucks and the freight price for subcontractor trucks. Because of the cost structure the own trucks are used with priority.

In a first, two-stage model the assignment of transport orders to tours is considered separately from the vehicle assignment. Solution procedures, especially for the first stage are mentioned only briefly. In a one-stage consideration the transport orders directly are assigned to trucks. By setting simplistic assumptions the problem can be modeled as quadratic assignment problem for which Söllig (1989) describes a heuristic solution procedure.

A case study considering the disposition of full loads, is described by Dargel (1983). He formulates the multi-period-problem as a set-partitioning problem and describes a heuristic solution procedure. Although the problem is different from that described in Section 2. the work is interesting because of the consideration of legal restrictions of the driver working time.

Recent work concerning the trunk haulage planning is done by Erdmann (1993), Schmidt (1994) and Brown/Ronen (1997). Erdmann (1993) considers the transport of shipments within one day. Each shipment can be transported either via the broken or via the non-broken transportation mode. The broken mode corresponds to the already described piece goods transport or broken transport of partial loads (see Section 2.). The transportation modes "indirect pickup, direct delivery" and "direct pickup, indirect delivery", which are important for partial load transports, are not considered here. Shipments, which are transported via the non-broken transportation mode, do not leave their loading unit between their origins and their destinations. However, they can be part of a direct or an indirect tour. On a direct tour the truck first picks up all shipments at their origins. From the last origin the truck directly goes to the corresponding destinations. Therefore, a direct tour has no depot contact. An indirect tour has contact to two depots. At the depots additional shipments can be loaded or unloaded and the drivers of the truck change. Because of these driver changes more total driving time is available than on direct tours.

In addition to the decision about the transportation mode and the assignment of shipments to tours also the sequence of loading locations as well as the temporal course of each tour has to be determined. Furthermore, all tours have to be assigned to vehicles, which all have the same capacity. However,

the vehicles have individual starting and ending locations and an earliest starting and a latest ending time. For all vehicles it is known whether they operate on direct or on indirect tours. The tours have to be planned and assigned to vehicles, so that the costs, depending on total driving and stop times, are minimized. As simplification, the local pickup and delivery tours in the broken mode are note planned. However, the costs are approximately considered by a cost factor which depends on the customer and on the depot.

Erdmann presents a heuristic procedure in which first shipments for the non-broken transportation mode are selected and combined to tours. The selection of shipments for the non-broken transportation mode is supported by an approximative "advantage criterion" which can be determined by solving a Traveling Salesman Problem with time windows. The broken transportation mode then is used for the remaining shipments. First each shipment is assigned to the next sending and receiving depot. In a postoptimization step the assignment of shipments to their sending or receiving depot is changed successively with the objective to decrease the number of necessary vehicles.

The problem discussed by Erdmann is very similar to the problem described in Section 2., but there are main differences concerning the following aspects: Whereas in the problem focused by Erdmann the sending and receiving depots of a shipment have to be determined during the planning procedure, in carrier companies which transport piece goods and partial loads the areas of each depot are fixed on a long- or medium-term run. A fixed schedule for line-haul shipments is not considered also. Considering the operational aspects, Erdmann only assumes homogenous time windows for all origins and all destinations, respectively. Furthermore the possibilities of partial load transports are restricted as mentioned above. For the planning horizon of one day (13 h - 13 h) it is assumed, that all data are available in advance and that no back or connecting loads have to be considered.

Opposite to Erdmann (1993), Schmidt (1994) concentrates in his work on the disposition of directly picked up and delivered partial loads. Line-haul shipments with consolidated piece goods are considered as special partial loads between two depots. Similar to Feige (1983) the problem of planning vehicles in trunk haulage consists in the problem of building tours, which corresponds to the parallel coordination problem, and the problem of planning vehicle round trips, which corresponds the sequence coordination problem. While in the problem of Feige (1983) the transport orders are given, here an additional decision on the acceptance of the transportation orders has to be taken. Therefore, the objective is not to transport all orders at minimum cost, but rather to select "good" orders.

Schmidt also decomposes the problem into the two subproblems. In the tour building problem basic orders are selected for each tour which have to be transported anyway, but for which a single transport is not economical. Then, in a savings-based heuristic the transport orders are assigned to these basic

orders according to increasing distance measures. During the tour building process the tours are not yet assigned to vehicles. But as a heterogeneous fleet of vehicles has to be considered, it is checked approximately if the assignment of transport orders to tours is feasible with respect to capacity.

The objective of the subsequent vehicle round trip planning is to combine the tours to round trips and to assign the round trips to vehicles so as to minimze the transportation cost. The problem can be modeled as multicommodity network flow problem (MCNFP) or in case of homogeneous vehicles as assignment problem (AP), if the check of time windows is simplified. As the transportation orders are completely known only for the present planning period, Schmidt describes a procedure based on an AP-Formulation, in which only the first tours on a vehicle are fixed, but where possible round trips are taken into account.

A case study considering the consolidation of orders between several plants of a large manufacturer and its customers is presented by Brown/Ronen (1997). The combination of orders to multiple day trunk tours is part of an interactive process similar to the disposition process described in Section 2.. They present a column generation procedure, which first generates possible consolidated tours, considering different restrictions as time windows, legal constraints or a maximal number of delivery stops on a route. In the subsequent optimization step a set of tours is selected which causes the minimal total mileage and in which any order is not consolidated more than once. The optimization step is done by solving an "elastic" set partitioning problem.

The presented works each deal with some parts of the problems described in Section 2.. Erdmann (1993) defines the problem of one-day planning. But he neglects some practical relevant requests, such as the consideration of connecting tours or the dynamically incoming transport orders. The last aspect is not considered by Feige (1983) also. The planning of vehicle round trips are treated by Feige (1983) and by Schmidt (1994). Especially Schmidt (1994) realizes the vehicle round trip planning within the concept of rolling horizon. Therefore, he can process the uncertainty and the incompleteness concerning the shipment data of the following periods. But he does not deal with the dynamic incoming orders during one period. Brown/Ronen (1997) consider a problem of practical relevance for the distribution of goods from few plants to many customers. But they do not deal with different transportation modes. The problem of legal driving and recreation times is approximately considered by Erdmann (1993) by the prescribed changes of drivers on indirect tours and the maximum operating time on direct tours. A relative detailed planning of recreation times is presented by Dargel (1983) for the special problem of full load disposition.

Because of the problem complexity all authors have to make additional assumptions for simplification. Furthermore, they use almost exclusively heuristic solution heuristics.

4. Model

The problem described in Section 2. is a large extension of traditional vehicle routing and scheduling problems. The standard VRP is contained as the special case, where the origins of all shipments coincide with the depot. For this complex problem, it is necessary, on one hand, to make additional assumptions for modelling the problem. On the other hand heuristic procedures for optimization (see Section 5.) have to be used. In this Section a standard problem, which includes the basic functions of the disposition of trunk haulage, will be modelled. First we describe the main assumptions of the model. In Section 4.2 the basic problem with required data, restrictions, decisions and objectives is formulated. Section 4.3 discusses the criteria of evaluating the disposition result in greater detail.

4.1 Assumptions

The complex practical problem necessitates structuring and simplification. First of all the planning situation should reflect the disposition functions in a simplified, but practically still relevant standard case. Therefore, the following assumptions are made:

1. One-day planning of trunk haulage
Only tours for the present day are planned. We will not consider information about possible back and connecting loads on the following days.

2. Complete shipment data
At the beginning of the disposition in the early morning all shipments for the present day are completely available.

3. Homogeneous fleet of vehicles
For the long-distance tours a homogeneous fleet of trucks is available. The trucks are exclusively trucks from subcontractors, which can be booked on the transport market. The prices for the trucks are known. Legal driving and recreation times are only approximately considered by the restriction of a maximum driving time on each tour, which is identical for all tours or vehicles respectively.

4. Planning only long-distance tours
We do not plan the tours of vans and carrier vehicles that pickup and deliver shipments to be transshipped at the depots. For comparability between direct and indirect mode, the costs for indirect pickup and delivery are computed approximately (see Section 4.3).

5. Schedule for line-haul shipments
We will not consider composite shipping paths for line-haul shipments. It is assumed, that line-haul shipments are transported directly between two depots. However, it is possible to consider specific time windows for line-haul shipments.

6. Transportation modes for each shipment
It is sufficient to consider only one kind of shipments, namely partial loads, in the model because of the following reasons:

- As no back and connection loads have to be considered, full loads can be neglected in the model.
- Most of the small shipments are a priori consolidated to line-haul shipments. Line-haul shipments can be dealt as special partial loads between two depots, for which a schedule with fixed starting times could be given.
- Each of the remaining shipments can be picked up and delivered either directly or indirectly (cf. transportation modes for partial loads described in Section 2.). An indirect pickup or delivery is done alternatively by a van or by a carrier vehicle. Shipments which are transported in the transportation modes 1 - 3 are partial loads. If a shipment is picked up and delivered indirectly (transportation mode 4), in practice the shipment could be a piece good shipment or a partial load. As we do not plan the tours of vans and carrier vehicles, the only difference between the broken piece goods transport and the broken partial load transport is the different cost. Therefore, except for the costs we do not distinguish between the transportation mode 4 for partial loads and the piece goods transport.

In spite of the simplifications compared with the problems described in Section 2., the above defined planning situation reflects relatively well the practice of the trunk haulage disposition, as the following discussion of the assumptions shows:

1. One-day planning of trunk haulage
When dealing with partial and full loads, in most cases there are not enough shipment data for the next day available to be able to plan round trips. Even in most cases in practice, the dispatcher is only able to plan the vehicle trips for the current day. If there are some shipments known for the next day, the dispatcher tries to build a tour, which ends close to an origin of connecting loads or back loads. An extension of the model in order to process these connecting and back loads will be made in the future.

2. Complete shipment data
The dynamic incoming shipments can be treated in the model if it is used with a rolling horizon:
The planning process is restarted in regular intervals, e.g. one hour. The planning is based on the shipment data known until this time. From the solution those tours are selected and fixed which correspond to predefined criteria. When restarting the planning process, the shipments in non-fixed tours are planned together with new shipments arrived in the last hour. Criteria for fixing tours could be the use of capacity or of time of the planned tour.

3. Homogeneous fleet of vehicles
In carrier companies which transport piece goods and partial loads usually trucks with two loading bridges with a standard size are used. Therefore, the assumption of identical vehicles with respect to capacities is no further restriction, especially since other vehicle types only slightly differ from the vehicles with loading bridges.
Concerning the legal recreation and driving times, it is realistic that trucks and drivers, respectively, are available on the market which have enough time left to take over the planned tours. Thereby, the restriction of the maximum driving time supports the consideration of legal recreation and driving times.

4. Restriction to the planning of trunk haulage tours
Regarding the carrier vehicles, the driving personal is different for their use by day and by night. Therefore, drivers on trunk haulage tours by night have the complete operating time available. We only neglect, that tours of carrier vehicles should end at depots, from which they can operate during the following day.

5. Composite shipping paths
For carriers which transport only partial loads composite shipping paths on depot-depot-relations do not exist. When also transporting piece goods the shipping paths are medium-term fixed and their consideration complicates the model only in an organizational way.

6. Initial combination of shipments to line-haul shipments
In most of the carrier companies the short-haul disposition and the trunk haulage disposition are organized separately. The trunk haulage dispatcher usually gets only information about the already consolidated line-haul shipments. An initial given consolidation of shipments to line-haul shipments therefore corresponds to the procedure in practice.

All in all the described planning situation contains standard decisions which can be found more or less specialized in almost all carrier companies. Therefore, we can use the model described in the following as basis for a multiplicity of practical disposition problems.

4.2 Basic model

In this Section the model for the above defined simplified planning situation is formulated. The **data** needed are the following specifications about depots, vehicles, shipments, distances and travel times:

Depots $d = 1 \ldots D$, each with

- Adress (postal code and location) which defines the location of d in the street network
- Time window for loading $[A_d^{D,l}, B_d^{D,l}]$ and unloading $[A_d^{D,u}, B_d^{D,u}]$
- Cost functions for evaluating indirect pickup and delivery by vans and carrier vehicles

Homogeneous fleet of trucks

- Capacities C^W (weight), C^P (pallets), C^{LM} (loading meters)
- Maximum allowed driving time C^T
- Cost function for evaluating the trunk haulage tours

Shipments $s = 1 \ldots S$, each with

- Adress (postal code and location) of origin and destination which define the locations O_s and D_s in the street network
- Time windows for loading $[A_s^l, B_s^l]$ and unloading $[A_s^u, B_s^u]$
- Stop times for loading ST_s^l and unloading ST_s^u
- Allowed transportation modes: direct, indirect by carrier vehicle, indirect by van, each for loading and unloading
- Corresponding sending depot SD_s and receiving depot RD_s in case of indirect pickup and/or delivery
- Shipment size, given by q_s^W (weight), q_s^P (pallets) or q_s^{LM} (loading meters)

Street network

- Driving times $t(a, b)$ between two locations a and b; a,b are depot locations, origins or destinations
- Street distances $d(a, b)$ between two locations a and b, corresponding to driving times

According to the main functions of the trunk haulage disposition described in Section 2., the following decisions have to be made and restrictions have to be considered:

Decisions

- Fixing of transportation mode for each shipment
- Combination of shipments to long distance tours. For each long distance tour the sequence of loading and unloading locations has to be determined as well as the temporal course of the tour.
- Assignment of tours to trucks

Constraints

- The weight, pallets and loading meters must not exceed the truck capacities.
- The loading and unloading time windows of all shipments have to be considered.
- The maximum driving time on each tour has to be considered.
- All shipments have to be transported from their origins to their destinations.
- In each tour the first unloading location is visited only after all shipments have been picked up.

Objective

In the described basic model all shipments are given and must be transported. Therefore, the revenues cannot be influenced by the dispatcher. The objective of trunk haulage disposition therefore consists in the minimization of the costs from shipping all given shipments from their origins to their destinations. In the following Section 4.3 the evaluation of the solutions of the trunk haulage problem is discussed in greater detail.

4.3 Evaluation and objectives

The evaluation of the decisions of the dispatcher is done in practice according to different criteria. In addition to qualitative criteria the main objective consists in the maximization of profit. This objective was realized in the models described in Section 3. dependent on the situation by minimization of distances, by maximization of contribution margins or by minimization of costs.

As mentioned above, in the basic model proposed in Section 4.2 the transportation costs are minimized. Thereby the transportation costs for the trunk haulage tours and the approximative costs for the short-haul pickup and delivery tours by vans and carrier vehicles have to be considered. If the model is extended to back and connecting loads, the objective function should get a further component which measures how the trunk haulage tours are acceptable with respect to the use of vehicles on the following day.

Regarding the transportation costs for trunk haulage the following trend can be noted at the German transport market: Until the release of tariffs in the year 1994 the trunk haulage transports are payed according to official tariffs ("Güterfernverkehrstarif GFT" for piece goods and partial loads). Since the repeal of freight price reglementations a multiplicity of individual tariffs and agreements have been established. Further the market prices are dependent on seasons and on weekdays. The application of daily varying real-life costs is an essential precondition for the success of a computer supported optimization of trunk haulage disposition.

Therefore, a system has to process different kinds of cost functions (GFT, weight and or distance dependent tables, ...) which can be daily adapted according to the market situation. More often, the trend on the transport market is that the freight prices lean on the real vehicle costs (see Ebner (1997)). As the vehicle costs usually are composed of fixed and variable cost components, the trunk haulage tours should be evaluated according to these components, if no detailed information is available about the real-life price conditions.

The evaluation of indirect pickups and deliveries is done without exactly planning the vehicle tours. The procedure follows closely the "ring model" of Fleischmann (1997). Because of the different vehicles (vans, carrier vehicles), there is a difference, whether the shipment is carried out by a van or by a carrier vehicle. Therefore, for each pickup or delivery the costs are determined for both possibilities and this vehicle type which causes the lower costs is selected.

5. Solution procedures

As already mentioned in the preceding section, the problem is too complex to use exact mathematic optimization procedures. Therefore we developed a heuristic procedure for solving the basic problem formulated in Section 4.. First, a starting solution is generated. Subsequently, this starting solution is improved successively by different exchange steps. The procedure of the starting heuristic is described in Section 5.1, the different strategies for improving the starting solution are subject of Section 5.2.

5.1 Starting heuristic

The composition of tours in the starting heuristic is similar to the savings procedure for the one depot vehicle routing problem. Starting with single tours, the tours are combined successively according to a preliminary determined sequence.

Generally, a trunk haulage tour T is represented by a sequence of loading locations $(l_1^T, \ldots, l_{n_1}^T, l_{n_1+1}^T, \ldots, l_{n_1+n_2}^T)$. At the first n_1 locations shipments are loaded, at the second n_2 locations shipments are unloaded. A loading location relates only to locations which are visited within the trunk tour T. For shipments which are indirectly picked up and/or delivered, the trunk haulage loading locations are the corresponding sending and receiving depots. Therefore $l_1^T, \ldots, l_{n_1}^T$ are origins of shipments or sending depots and $l_{n_1+1}^T, \ldots, l_{n_1+n_2}^T$ are destinations or receiving depots. At each loading location, especially at depots, more than one shipment can be loaded or unloaded.

To describe a tour, the following informations are necessary:

- weight W^T, number of pallets Pal^T, load meters Lm^T
- number of loading locations AzL^T
- for each loading location $l_i^T, i = 1, \ldots, AzL^T$
  - $[EST_i^T, LST_i^T]$ earliest and latest loading begin
  - Wz_i^T waiting time between arrival time and loading begin
  - Bz_i^T present loading begin
  - Sz_i^T stop time, i. e. time for loading or unloading
  - specification of AzS shipments to be loaded or unloaded

Figure 5.1 gives an overview over the procedure for generating a starting solution.

Initialization

In the first step of the initialization for all shipments with not yet fixed transportation mode it has to be decided whether the shipments are picked up and delivered directly or indirectly. The decision is made according to two weight limits G^P for pickup and G^D for delivery, that is a shipment s is picked up (delivered) directly if $q_s^W \geq G^P$ ($q_s^W \geq G^D$) and indirectly otherwise. The weight limits G^P and G^D are parameters of the starting heuristic.

1. **Initialization**
   - 1.1. Fixing of transportation mode of each shipment
   - 1.2. Formation of single starting tours $T_s, s = 1, \ldots, S$
   - 1.3. Computation of distance measures $A_{i,j}$ for all pairs of shipments $(i, j), i, j = 1, \ldots, S, i < j$
   - 1.4. Sorting of shipment pairs according to increasing distance measures $A_{i,j}$
2. **Iteration**
   - 2.1. Selection of next shipment pair (i, j)
   - 2.2. Check of combination $T = T_i \times T_j$:
     - 2.2.1. If $T_i = T_j$, go to next iteration
     - 2.2.2. If the required capacity of $T_i \times T_j$ is greater than the vehicle capacity, go to next iteration
     - 2.2.3. Construction of combination tour T: $\forall k = 1, \ldots, AzL^{T_j}$:
       - Determine the best feasible insertion position of $l_k^{T_j}$ in T_i
       - If the insertion of $l_k^{T_j}$ in T_i is infeasible, reject the combination of $T_i \times T_j$, go to next iteration
         Else, realize the insertion of $l_k^{T_j}$ in T_i.
3. **Termination** All pairs (i, j) are worked off.
4. **Evaluation of tours**

Fig. 5.1. Starting heuristic

In case of indirect pickup, the pickup can be realized either by a van or a carrier vehicle. If both is possible we choose the possibility which causes the lower approximative pickup cost. In analogy, this is applied to the delivery. In order to initialize the tour composition, subsequently, single starting tours $T_s, s = 1, \ldots, S$, are formed, each containing only one shipment s. (With T_s we describe in the following the tour which contains shipment s). Each tour T_s has two loading locations. $l_1^{T_s}$ is the trunk haulage origin and $l_2^{T_s}$ the trunk haulage destination of shipment s.

The trunk haulage origin is in case of direct pickup the origin of the shipment itself and in case of indirect pickup the corresponding sending depot. The trunk haulage destination is the destination of the shipment (direct delivery) or the receiving depot, respectively (indirect delivery):

$$l_1^{T_s} = \begin{cases} O_s \\ SD_s \end{cases} \quad \text{if } s \text{ is picked up} \quad \begin{cases} \text{directly} \\ \text{indirectly} \end{cases}$$

$$l_2^{T_s} = \begin{cases} D_s \\ RD_s \end{cases} \quad \text{if } s \text{ is delivered} \quad \begin{cases} \text{directly} \\ \text{indirectly} \end{cases}$$

The earliest and latest start time at the first loading location $[EST_1^{T_s}, LST_1^{T_s}]$ is equal to the time window of the origin of shipment s in case of direct pickup and to the time window of the corresponding sending depot in case of indirect pickup. In analogy this also holds for the second loading location.

$$[EST_1^{T_s}, LST_1^{T_s}] = \begin{cases} [A_s^l, B_s^l] \\ \left[A_{SD_s}^{D,l}, B_{SD_s}^{D,l}\right] \end{cases} \quad \text{if } s \text{ is picked up} \quad \begin{cases} \text{directly} \\ \text{indirectly} \end{cases}$$

$$[EST_2^{T_s}, LST_2^{T_s}] = \begin{cases} [A_s^u, B_s^u] \\ \left[A_{RD_s}^{D,u}, B_{RD_s}^{D,u}\right] \end{cases} \quad \text{if } s \text{ is delivered} \quad \begin{cases} \text{directly} \\ \text{indirectly} \end{cases}$$

It is assumed that the present loading begin takes place at the earliest possible time. The loading begin times as well as the waiting times therefore can be computed as follows:

$$\begin{aligned} Bz_1^{T_s} &= EST_1^{T_s} \\ Wz_1^{T_s} &= 0 \\ Bz_2^{T_s} &= Max(EST_2^{T_s}, Bz_1^{T_s} + Sz_1^{T_s} + t(l_1^{T_s}, l_2^{T_s})) \\ Wz_2^{T_s} &= Max(0, EST_2^{T_s} - (Bz_1^{T_s} + Sz_1^{T_s} + t(l_1^{T_s}, l_2^{T_s}))), \end{aligned}$$

Within the initialization, for all shipment pairs (i, j) also the distance measure $A_{i,j}$ is determined. $A_{i,j}$ should tell something about the proximity of two shipments. It is assumed that for two shipments which are close to each other, the joint transport in a trunk tour is desired. Therefore, with the sorting of shipment pairs according to increasing distance measures a sensible sequence for the combination of tours is defined.
For the distance calculation not necessarily the origin and destinations of shipments are relevant, but the trunk haulage origin and destinations. As distance measure $A_{i,j}$ between two shipments i and j we use:

$$A_{i,j} = d(i^l, j^l) + d(i^u, j^u),$$

where i^l and j^l are the trunk haulage origins, i^u and j^u the trunk haulage destinations of shipments i, j. Other definitions of $A_{i,j}$ are possible, e.g. including the similarity of time windows.

Iteration

In an iterative process the tours are combined successively with each other. The sequence of combination steps is determined by the sorting of the $A_{i,j}$. In each iteration the next shipment pair (i, j) is selected. Then it has to be checked whether the tour T_i which contains shipment i can be combined with the tour T_j which contains shipment j. A combination is infeasible,

- if the shipments i and j are already part of the same tour ($T_i = T_j$) or
- if the combination tour exceeds the vehicle capacities.

If these simple conditions for the combination of the two tours are fulfilled, the combination tour $T = T_i \times T_j$ is constructed by successively inserting each loading location $l_k^{T_j} (k = 1, \ldots, AzL^{T_j})$ of tour T_j into tour T_i at the best feasible insertion position. An insertion is feasible, if

- the sequence constraints (all origins before the first destination) are kept,
- the time windows are not violated and
- the driving time does not exceed the maximum allowed driving time.

If the insertion of a loading location is infeasible at all positions of the tour, the combination of tours T_i and T_j is rejected. If more than one feasible insertion position exists, the one with the lowest duration of the new tour is selected. If in two alternatives the duration of the tour is equal, the second criterion for the best insertion position is the detour caused by the insertion. The feasibility of an insertion is checked by a successive forward computation (cf. push forward concept described by Solomon (1987)).

Termination

The iterative combination of tours terminates, if all pairs of shipments are worked off. Alternatively the procedure already terminates, if the distance measure $A_{i,j}$ of the present selected pair (i, j) exceeds a given distance limit. The idea is, that the probability for a successful combination of two tours is small if the loading locations of the shipments are too far remote.

Evaluation of tours and of indirect pickups and deliveries

Like in the traditional savings procedure for the VRP the tours are only evaluated when the combination of tours is completed. However, the combination steps follow the objective of cost minimization. Because of the combination of two tours, in each successful iteration one vehicle is saved. If using a real-life cost function this is equivalent to a cost reduction. Also by encouraging the combination of tours which contain shipments close to each other, the transportation costs are indirectly influenced.

The planned tours can be evaluated by any possible cost function. For the practice tests which are presented in Section 6. the tours are evaluated by a cost function, which contains a fix and a time and distance dependent component.

Regarding the transportation modes the starting heuristic does not support any optimization. The costs of indirect pickup and delivery are fixed by the demarcation of pickup and delivery mode according to the two weight limits G^P and G^D at the beginning of the initialization.
Because of fast computation times, however, it is possible that the starting heuristic is parameterized similarly to the parametric savings procedure for the one-depot VRP (see Paessens (1988)). Different weight limits are used to gererate different solutions. Subsequently the lowest cost solution is used as basis for improvement heuristics described in the following.

5.2 Post-optimization heuristics

The result of the starting heuristic can be improved by a set of post-optimization heuristics based on the shifting of shipments between two tours or within a tour. The selection of shipments to be shifted is random or semi-random. Furthermore, all heuristics allow the increase of the costs during the shift actions up to a given threshold.

During the shift actions also the transportation mode fixed in step 1.1 of the initialization (see Figure 5.1) may be changed. For each feasible mode among the 4 transportation modes (see Section 2.) the effect in the objective value is calculated and the best transportation mode is realized.

In the following three shift heuristics ("S1", "Sn", "Kombi") are described more in detail.

S1 heuristic

The essential steps of the S1 heuristic are summarized in Figure 5.2. The heuristic starts with the present and best known solution, the solution of the starting heuristic z_{pres}. In each iteration one shipment is selected either randomly from all shipments or from a subset of all shipments. The subset contains those shipments for which the shifting into another tour is very promising. Within the scope of the test program only those shipments were selected, which are the only shipment at their loading or unloading location in the tour. Removing these shipments from their tours saves at least one loading location, which has also positive cost effects.

The maximum allowed deterioration of the objective value is computed as a percentage (parameter between 3% and 10%) of the costs of that tour, from which the shipment is removed. In principle also a successive reduction of the threshold percentage during the iterations is possible. Then the reduction should depend from the number of iterations without change in the present solution or without improvement of the optimal solution. So far this traditional threshold accepting strategy has not been tested.

1. **Initialization**
   z_{opt} := objective value of the starting solution, $z_{pres} := z_{opt}$
2. **Iteration**
   2.1. Selection of a shipment i
   2.2. Determination of the maximum allowed increase Δz of the objective value
   2.3. Identification of neighboured tours
   2.4. Best feasible insertion of i in one of the neighboured tours
   2.5. If $z_{new} \leq z_{pres} + \Delta z$:
   $z_{pres} := z_{new}$
   (and save new present solution)
   If $z_{new} < z_{opt} \rightarrow z_{opt} := z_{new}$
   (and save new optimal solution)
3. **Termination**
   after a maximum number of iterations

Fig. 5.2. S1 heuristic

In order to avoid a high computational effort, it is only allowed to shift the selected shipment into one of its neighbouring tours. Tours are considered as neighboured to a shipment, if one of their loading locations is close to the origin and one of their unloading locations is close to the destination of the shipment.

As mentioned above, the transportation mode of a shipment which is inserted into a neighbouring tour can change, if this is feasible and advantageous. If the objective value after shifting the shipment in its best neighbouring tour is less than the threshold, i.e. the sum of the present cost and

the allowed increase in cost, the new solution is accepted as new present solution. If, moreover, the solution is better than the best known solution, e.g. $z_{new} < z_{opt}$, also the best known solution z_{opt} is updated.

The procedure terminates after a maximum number of iterations. It might also be possible to terminate after a maximum number of shift actions without success.

Sn heuristic

The only difference between the Sn heuristic and the S1 heuristic is, that in each iteration simultaneously n ($n \geq 2$ is a parameter which has to be set a priori) shipments are shifted between tours. First all n shipments are removed from their tours. Each single shipment is then inserted into the best of its neighbouring tours.

The selection of shipments is semi-random. A basic shipment is selected randomly. Indeed, the other shipments are also selected randomly, but only from the subset of shipments which are in the neighbourhood of the first shipment. The size of the neighbourhood can be modified by parameter settings.

The Sn heuristic has an higher computational effort, but due to the selective choice of shipments from the neighbourhood of the first shipment the chance to get a new present solution is higher.

Combination heuristic

In the combination heuristic the advantages of the first two heuristics are used, namely

- little computational effort for the S1 heuristic and
- more successful search for improvements in the Sn heuristic.

The S1 and the Sn heuristic are combined, by alternating between steps of both strategies (see Figure 5.3). One switches between the two heuristics always then, when an iteration was not successful, i. e. the present solution has not changed.

However, the Sn-step is simplified insofar, as the selected shipments are only allowed to be shifted into one of the involved tours. Tours are involved if they contain at least one of the shipments to be shifted. Because of the selection rule the shipments are neighboured, so that the shift actions are restricted to a special selection of the neighbouring tours.
Compared to the Sn heuristic, the computational effort for the combination heuristic is lower. Therefore, the combination heuristic can be additionally parameterized. In this case the combination heuristic runs for different values $n = 1, \ldots, n_{max}$ in the Sn-step. The result of one n thereby is used as starting solution for the next pass $n + 1$.

1. **Initialization**
   as in the S1 heuristic
2. **Iteration**
   2.1. S1-step:
   Random selection of a shipment and feasible insertion in the best neighbouring tour
   2.2. If new solution is accepted →
   Begin next iteration with 2.1.
   Else begin next iteration with 2.3.
   2.3. Modified Sn-step:
   - Semi-random selection of n shipments
   - Feasible insertion of each shipment in the best of the **involved** tours.
   2.4. If new solution is accepted →
   Begin next iteration with 2.3.
   Else begin next iteration with 2.1.
3. **Termination**
   after a maximum number of iterations

Fig. 5.3. Combination heuristic

6. Computational Tests

All described solution procedures are implemented in the DELPHI-program ToPFiT. ToPFiT was tested with practical data from three depots of a German carrier company. For each depot data for five days (one week) are available. The number of shipments varies between 50 and 220 per day and depot. In the data the piece goods shipments are already combined to line-haul shipments. For all non-line-haul shipments the choice of the transportation mode is free. Additionally to the vehicle capacity, a maximum number of 6 loading locations will be considered. The maximum driving time is constrained to 10 and 11 hours in two different scenarios. Therefore 30 data sets are available.

In the following we compare the starting heuristic, the S1-heuristic, the Sn-heuristic for $n = 3$ (S3-heuristic) and two variations of the combination-heuristic. In the variation Kombi A the number of shifted shipments n during a Sn-step varies between 2 and 3 ($n_{max} = 3$), in the second variation Kombi B there is a parametrization of the Sn-step for $n = 2, \ldots, 6$ ($n_{max} = 6$). Because of the construction of the combination-heuristic, under the same conditions the second variation Kombi B produces always at least the same result as Kombi A. The S1, S3 and each iteration of Kombi A and Kombi B run for 1000 iterations. The maximum allowed deterioration of objective value is fixed to 3% of those tour costs which contain shipments to be shifted. To initialize the pickup and delivery state we test alternatively 2.5 tons and 5 tons as weight limit.

The evaluation of indirect pickups and deliveries is done by the approximation mentioned in Section 4.3. For evaluating the trunk tours we use costs with a driving time dependent, a distance dependent and a fix component. To favor a higher use of capacity of tours through the objective function a

so-called capacity factor is introduced. With the capacity factor af the cost $K(T)$ of a trunk tour is corrected as follows:

$$K(T) := K(T) + af \cdot (1 - \text{use of capacity(T)}) \cdot K(T) = K(T) \cdot (1 + af \cdot (1 - \text{use of capacity(T)}))$$

If the capacity factor is greater 0 the original cost is supplemented by a penalty cost. These penalty costs are the higher the lower the use of capacity of the tour is. We made tests with 0% and 50%, respectively.

Altogether we considered two different weight limits and two different capacity factors for each of the four improvement strategies . As the starting heuristic is a deterministic procedure with cost-independent construction rules the choice of the capacity factor only has an effect on the objective value but not on the constructed tours. Therefore, during the starting procedure only the two different weight limits were distinguished. Altogether we tested 18 different parameter settings for each of the 30 available data sets.

As a comparison with practical results is not possible because of missing or incomplete information, the procedures are only compared among each other. The use of capacity of vehicles is one objective criterium to evaluate and compare different solutions.

Table 6.1. computation times, capacity utilization and quality of the solution

heuristic	no. of solutions	⌀ computation time	⌀ use of capacity		⌀ GAP
			$af = 0\%$	$af = 50\%$	
SH	60	00:01	62 %	62 %	17,0 %
S1	120	00:16	72 %	74 %	4,7 %
S3	120	01:24	74 %	76 %	3,3 %
Kombi A	120	01:14	73 %	75 %	2,6 %
Kombi B	120	04:20	75 %	78 %	0,7 %

Table 6.1 summarizes the main results of the computational tests. Every line shows average values for the 30 data sets and the two (SH) and 4 (other heuristics) parameter settings.
The average computation time on a Pentium PC varied between a few seconds for generating a starting solution up to over four minutes for the extended combination-heuristic.

The use of capacity after the starting heuristic is very low whereas the values after the postoptimization heuristics increase about 10 %. The effect of two factors can be identified. As expected, the choice of a higher capacity factor favors a higher use of capacity. Furthermore the influence of the capacitated driving time to 10 and 11 hours, respectively, can be recognized. The values in Table 6.1 are average values, also including the two different driving time scenarios. However, when splitting the results, the use of capacity

in the 10 hour-scenario in average is 2 % - 4 % lower than in the tests with a maximum driving time of 11 hours.

The last column in Table 6.1 shows the average relative deviation of the objective values from the objective value of the best found solution. The objective values all are corrected by the penalty cost. As expected the quality of solutions is nearly contrary to the computation times. The starting heuristic produces very low quality solutions. The application of postoptimization strategies therefore is very important. Comparing the improvement heuristics the Kombi B heuristic is clearly the best. The average deviation to the best found solutions is less than one percent.

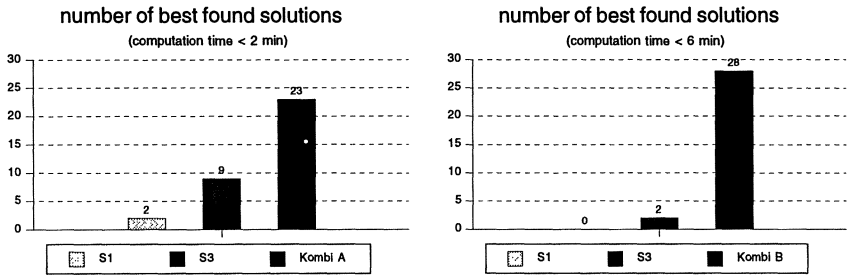


Fig. 6.1. Comparison of postoptimization strategies

If we compare the postoptimization strategies concerning the number of best found solutions, we get the following results (see Figure 6.1):

Comparing only those strategies, which have computation times under 2 minutes (all expect Kombi B), Kombi A found the best known objective value from 23 of the tested 30 data sets. In 9 cases the S3 heuristic and in 2 cases the S1 heuristic was successful. (In some data sets two strategies yielded the best solution at the same time.)

If considering no time limit, the Kombi B heuristic competes with the S1 and S3 heuristic. In this case Kombi B obviously comes out on top with 28 best solutions. The computation times, which are on average slightly under 4,5 minutes, are on average 2,5 minutes concerning the data sets with 50 shipments and 5-6 minutes for the data set with 220 shipments.

For an iterative, dynamic application of the procedures in the praxis a computation time of 5 minutes is an upper limit. When planning several hundred shipments, a relatively extended parametrization like in the Kombi B heuristic seems to last too long. However, the results of the less parameterized Kombi A heuristic with an average deviation of less than 3 % from the best known objective value can be accepted. If realizing the PC-based disposition of trunk haulage in praxis, the automatically planned tours are checked by the dispatcher anyway.

7. Summary and further steps

In this paper the practical problem of routing and scheduling tours and vehicles in trunk haulage was described. Some ideas of literature were presented. For a basic problem we developed a heuristic to construct a starting solution as well as different heuristics for improving the starting solution.

The basic problem restricts to an one-day-planning assuming deterministic shipment data. An embedding in the general case of dynamic arriving orders during the morning can be realized with the concept of rolling horizon in combination with the successive fixing of tours.

As discussed in Section 4. the basic problem reflects the main functions requested by a real-life disposition. The computation times of the presented solution procedures are acceptable within a rolling planning. Further research is needed for the extension of the described model to some of the practical requirements described in Section 2., e.g.

- heterogenous fleet of vehicles with given starting locations and earliest starting times
- consideration of back and connection loads
- composite shipping paths for line-haul shipments
- legal driving and recreation times.

In addition, for better evaluation of the solution quality it is necessary to test the procedures also in practice within a trunk haulage department.

References

AIF-Projekt 9767 "Güterverkehrsnetze" (1995): Literaturrecherche und Algorithmenentwicklung zur Modellierung und Optimierung von Güterverkehrsnetzen. Zwischenbericht.

AIF-Projekt 131/96 "Fuhrpark-Disposition" (1996): Pflichtenheft.

Brown, G.G. / Ronen, D. (1997): Consolidation of Customer Orders into Truckloads at a Large Manufacturer. In: Journal of the Operational Research Society 48, 779 - 785.

Dargel, W. (1983): Ein Problem der Fahrzeug-Einsatzplanung im Getränkevertrieb. Studienarbeit an der Universität Hamburg.

Dejax, P.J. / Crainic, T.G. (1987): A Review of Empty Flows and Fleet Management Models in Freight Transportation. In: Transportation Science, Vol. 21, No. 4, 227 - 247.

Deutsche Gesellschaft für Mittelstandsberatung mbH (Eds.) (1995): Transport- und Speditionswesen: Positionen, Perspektiven, Strategien. Neu-Isenburg.

Dumas, Y. / Desrosiers, J. / Soumis, F. (1991): The pickup and delivery problem with time windows. In: EJOR 54, 7 - 22.

Ebner, G. (1997): Controlling komplexer Logistiknetzwerke. GVB Schriftenreihe 34.

Erdmann, J. (1993): Servicezeitorientierte Distributionslogistik. Logistik und Verkehr 45, Vandenhoeck & Ruprecht.

Feige, D. (1983): Untersuchungen zur rechnergestützten Kfz-Einsatzplanung. Habilitationsschrift Dresden.

Fleischmann, B. (1997): Design of Freight Traffic Networks. In: Stähly et al. (Eds.). Advances in Distribution Logistics. Springer Verlag.

Hemming, H. / Ebner, G. / Kraus, S. / Wlcek, H. (1997): Kosten- und umweltorientierte Optimierung von Güterverkehrsnetzen. Projektbericht.

Ioachim, I. / Desrosiers, J. / Dumas, Y. / Solomon, M. / Villeneuve, D. (1995): A Request Clustering Algorithm for Door-to-Door Handicapped Transportation. Transportation Science, Vol. 29, No. 1, 63 - 78.

Leung, J.M.Y. / Magnanti, T.L / Singhal, V. (1990): Routing in Point-To-Point Delivery Systems: Formulations and Solution Heuristics. In:Transportation Science, Vol. 24, No. 4, 245 - 260.

Paessens, H. (1988): The savings algorithm for the vehicle routing problem. In: EJOR 34, 336 - 344.

Powell, W.B. (1983): The Load Planning Problem of Motor Carriers: Problem Description and a Proposed Solution Approach. In: Transportation Research A, Vol. 17 A, No. 6, 471 - 480.

Powell, W.B. (1986): A Local Improvement Heuristic of Less-than-Truckload Motor Carrier Networks. In: Transportation Science, Vol. 20, No. 4, 256 - 257.

Powell, W.B. (1987): Dynamic Vehicle Allocation Problem with Uncertain Demands. In: Transportation Research, Vol. 21 B, No. 3, 217 - 232.

Powell, W.B. (1996): A Stochastic Formulation of the Dynamic Assignment Problem, with an Application to Truckload Motor Carriers. In: Transportation Science, Vol. 30, Nr. 3, 195 - 219.

Rego, C. / Roucairol, C. (1995): Using Tabu Search for Solving a Dynamic Multi-Terminal Truck Dispatching Problem. In: EJOR 83, 411 - 429.

Savelsbergh / M. W. P., Sol, M. (1995): The General Pickup- and Delivery Problem. Transportation Science, Vol. 29, No. 1, 17 - 29.

Schmidt, J. (1994): Die Fahrzeugeinsatzplanung im gewerblichen Güterfernverkehr. Verlag Peter Lang.

Söllig, M. (1989): LKW-Einsatzplanung für den Fernverkehr in einem Speditionsunternehmen. Diplomarbeit an der Universität Hamburg.

Solomon, M. (1987): Algorithms for the vehicle routing and scheduling problems with time window constraints. Operations Research, Vol. 35, No. 2, 254 - 265.

Wlcek, H. (1997): Local Search Heuristics for the Design of Freight Carrier Networks. In: Stähly et al. (Eds.). Advances in Distribution Logistics. Springer Verlag.

Chapter 4

Operations within the Warehouse

When to apply optimal or heuristic routing of orderpickers

René de Koster[1], Edo van der Poort[2] and Kees Jan Roodbergen[1]

[1] Rotterdam School of Management, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[2] Department of Econometrics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

Abstract. In this paper the differences in performance of heuristic and optimal strategies for routing orderpickers in a warehouse are compared. Specifically, modern warehouses are considered, where orderpicking trucks can pick up and deposit pick pallets at the head of every aisle without returning to a depot. Such environments can be found in many warehouses where paperless picking is performed from pallet locations with pickers having mobile terminals receiving instructions one by one. In order to find orderpicking routes of minimal length in the situation of decentralized depositing, we use an extension of the well-known polynomial time algorithm of Ratliff and Rosenthal (1983) that considered warehouses with a central depot. In practice, the problem is mainly solved by using the so-called S-shape heuristic in which orderpickers move in a S-shape curve along the pick locations. The performance of both routing strategies is compared by using simulation. The optimal algorithm can give substantial reductions in travel time per route. It turns out that the reduction in travel time strongly depends on the lay-out and operation of the warehouse. Simulation is very time consuming, both in creating computer programs and in calculation times. Therefore it is desirable to have statistically based, explicit formulas for the performance of the optimal algorithm and the S-shape heuristic. We adapt formulations of Hall (1993) for the case of decentralized depositing and improve their performance. After adaptation, the formulas perform fairly well, and are a good alternative for simulation to get a rough idea about differences in performance of the S-shape heuristic and the optimal algorithm.

Keywords. warehouse, orderpicking, routing, performance estimation

1 Orderpicking in warehouses

In warehouses and distribution centers, products have to be picked from specified storage locations on the basis of customer orders. In general, the orderpicking process is the most laborious of all warehouse processes. It may consume as much as 60% of all labor activities in the warehouse (cf. Drury (1988)). Another aspect of warehouse activities, i.e. the transport of materials between various functional areas, is discussed in De Koster & Van der Meer (1997).

Especially in distribution environments, the pick process is usually carried out under time constraints. Orders tend, more and more, to arrive late and have to be shipped the same day at pre-fixed departure times per destination (a hub, a group of shops, a geographical area or a customer). This leads to peak loadings and an on-going pressure to carry out the orderpicking process as efficient as possible. Therefore, many warehouses nowadays use paperless orderpicking systems rather than picklists with picking locations, that have to be collected at a central printer. The most commonly used way of paperless orderpicking is via mobile, hand-held or vehicle-mounted, terminals and printers. Paperless orderpicking systems have the clear advantage that orderpickers and storers are connected on-line with the warehouse information system, which results in accurate up-to-date stock information, on-line reaction on exceptional situations, and on-line control of progress. Moreover, the orderpickers can obtain pick and store instructions without leaving the storage area. These aspects lead to pick-error reduction and increased productivity.

The savings may be substantial in view of the picking throughput time per destination, but also in view of the efficient use of expensive special orderpicking equipment like high-bay narrow-aisle orderpicking trucks. The use of mobile terminals offers the possibility of a more decentralized way of operation. For example, in warehouses where orderpicking trucks are used and empty pick pallets (or other carriers) are available at the head of all aisles, orderpicking trucks can drop off full pallets at the head of every aisle. The transportation of the full pallets is taken care of by faster and also cheaper equipment, such as conveyors and forklift trucks. The orderpickers may therefore finish a picking route in any aisle and proceed with the new route in the same aisle. In the sequel of the paper such systems will be called orderpicking systems with *decentralized depositing.*

Another way to achieve savings on orderpickers and equipment is by optimizing orderpicking routes. Given that the orderpicker has to collect a number of products in specified quantities at known locations, in what sequence should the orderpicker visit these locations in order to minimize the distance traveled? The problem of finding shortest orderpicking routes for warehouses with a central depot can be solved in running time linear in the number of aisles and the number of pick locations (see Ratliff & Rosenthal (1983) and Carlier & Villon (1987)). In Van Dal (1992) the algorithm is ex-

tended for different warehouse lay-outs. Gelders & Heeremans (1994) solved the orderpicking problem by applying the branch-and-bound algorithm of Little *et al.* (1963) to a simplified warehouse lay-out for a particular type of warehouse. They report reductions on the total walking distance in the warehouse varying between 9% and 40% (depending on the number of items to be picked). The problem of finding a shortest orderpicking route in the case of decentralized depositing has recently been studied in De Koster and Van der Poort (1997). In practice, the problem of finding orderpicking routes in a warehouse is mainly solved by the so-called S-shape heuristic in which orderpickers move in a S-shape curve along the pick locations skipping the aisles where nothing has to be picked. For warehouse with a central depot, more advanced heuristics are considered in Hall (1993). These heuristics are not considered in this paper, since they have no straightforward analog for decentralized depositing. Besides the routing of orderpickers substantial efficiency gains can also be obtained from proper clustering the orders into routes (see Gibson & Sharp (1992), Rosenwein (1996), and De Koster *et al.* (1997)).

In this paper, we use simulation to investigate the gain in travel time of the optimal algorithm in comparison with the S-shape heuristic for warehouses, where decentralized depositing is applied. We use the polynomial algorithm of De Koster & Van der Poort (1997), that can find shortest orderpicking routes in both warehouses with a central depot and warehouses with decentralized depositing. Based on practice, we consider the following two warehousing situations:

1. Picking with orderpicking trucks in a narrow-aisle high-bay pallet warehouse.

2. Manual picking from shelf racks with decentralized depositing on, for example, a conveyor.

Both warehouse types consist of a number of parallel aisles and use decentralized depositing of picked items. The warehouse with shelf racks has short aisles, only 10 meters, to avoid long walking distances for the orderpickers. The orderpicking in a narrow-aisle high-bay pallet warehouse is performed with orderpicking trucks, that need considerable time to change aisles. Aisles in this type of warehouses are long to decrease the need for aisle changing.

Since simulation is time consuming, it is desirable to have non-simulation techniques that give an indication of possible gains in travel time. To this end we adapt existing explicit statistical formulas (see Hall (1993)) for the case of decentralized depositing. Furthermore, we compare the predictions of the formulas with simulated values and improve the formulas where deviations are encountered.

The paper is organized as follows: Different order picking strategies are presented in Section 2. Section 3 gives numerical results of comparing the optimal algorithm and the S-shape heuristic in the two warehousing situations.

In Section 4, we adapt statistical formulas from Hall (1993) for estimated route length for the situation of decentralized depositing. Practical considerations are mentioned in Section 5 and Section 6 contains concluding remarks.

2 Routing strategies for the orderpicking problem

A warehouse consists of a number of aisles of equal length. The items are stored at both sides of the aisles. In pallet warehouses, usually manned trucks (or cranes) operate in the aisles to pick up items. They can traverse the aisles in both directions, and changing direction is not a problem. Each order consists of a number of items that are usually spread out over a number of aisles. We assume that the items of an order can be picked in a single route. Aisle changes are possible at the front and rear ends of the aisles. Aisle changes with an orderpicking truck or crane to neighboring aisles are very time consuming. In a warehouse with a central depot, the orderpick pallets and pick lists can be picked up and deposited at the depot. In this case the start and finish point of the orderpicking route are known beforehand. In paperless orderpicking systems with decentralized depositing the orderpick pallets and pick information can be picked up and deposited at the head of every aisle. In this case, only the start point of the orderpicking route is known beforehand, because this is the end point of the previous route.

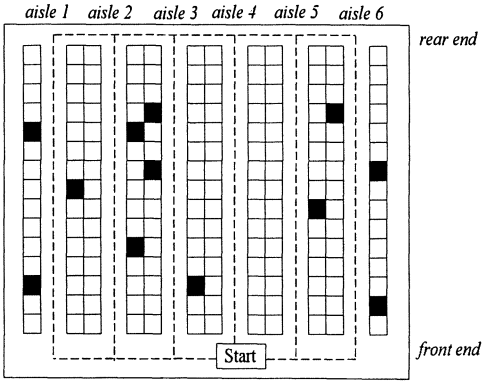


Figure 1: Schematic warehouse lay-out.

In order to determine an orderpicking route of minimum length, the travel time between each pair of adjacent (item) locations in the warehouse needs to be specified. In the specification of the travel time we can take into account the time for entering an aisle and the time for accelerating and decelerating while driving from one location to another. We will only focus on minimizing

the travel time. Other orderpicking activities, like positioning the truck or crane at the pick location, picking items from the pick location and putting them onto a product carrier, have to be performed anyway. Therefore, they do not impact the choice of an orderpicking route. The warehouse lay-out is shown schematically in Figure 1. Closed boxes indicate the section in the rack where items have to be picked. The dotted lines indicate where the order picker may drive.

2.1 The S-shape heuristic

The simplest way to route orderpickers is by using the S-shape heuristic. Any aisle containing at least one item is traversed through the entire length. Aisles where nothing has to be picked are skipped. After picking the last item, the orderpicker returns to the front end of the aisle. In Figure 2 a route is given, that is found by applying the S-shape heuristic to the warehouse lay-out of Figure 1.

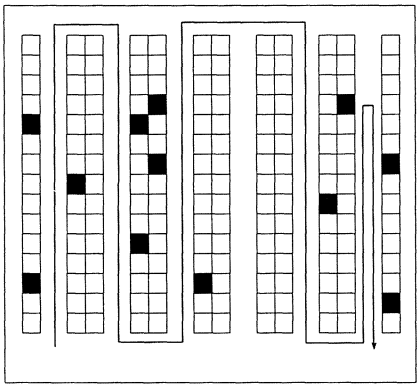


Figure 2: Orderpicking route for decentralized depositing found by applying the S-shape heuristic.

2.2 Optimal routing

If the S-shape heuristic is used, any aisle containing items is traversed entirely (except possibly the last aisle). Many other ways to traverse an aisle are possible. Figure 3 gives the six different ways (transitions) to traverse an aisle that can be identified in order to traverse an aisle optimally (see Ratliff & Rosenthal (1983)). In the figure, the rear end of aisle j is denoted by a_j and the front end by b_j. In transition (5) only the longest double edge is not traversed. Transition (3) and (4) are only possible if there is at least one item in the aisle, transition (5) is only possible if there are two or more items in the aisle and transition (6) is only allowed if the aisle is empty.

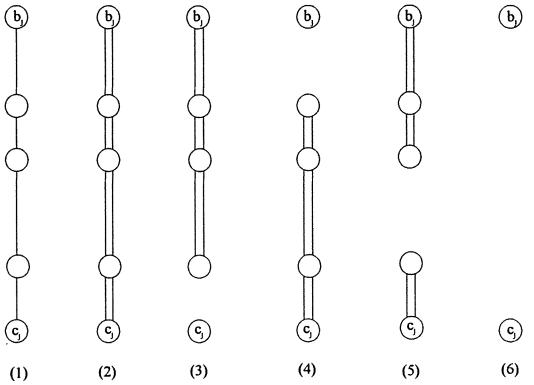


Figure 3: Six ways to traverse the edges in aisle j.

The warehouse with the orderpicking locations can be modeled as a graph with the vertices corresponding to the pick locations and endpoints of the aisles. Any two vertices that correspond to adjacent locations in the warehouse are connected by two parallel edges. No more than two parallel edges are needed, since it was shown by Ratliff & Rosenthal (1983) that a shortest route contains no more than two edges between any pair of vertices. The length of the edges indicates the travel times in the warehouse. Furthermore, a depot vertex is added. For a warehouse with decentralized depositing we introduce for each aisle two edges, with length 0, between the depot and the head of this aisle. The left part of Figure 4 shows the graph for the warehouse and pick locations of Figure 1 in the case of decentralized depositing. In the figure, the vertices v_i for $i = 1, ..., 12$ denote the orderpicking locations, the vertices a_i, b_i, for $i = 1, ..., 6$ the ends of the aisles, and vertex s the depot. Any two vertices that correspond to adjacent locations in the warehouse are connected by edges.

Any orderpicking route will be considered as being a special kind of subgraph of the warehouse graph, and is called a *routesubgraph*. That is, any subgraph of the warehouse graph is called a routesubgraph if its edges form a cycle that includes the depot once and each of the pick locations at least once. The length of a subgraph is defined as the sum of the length of the edges in this subgraph. The right part of Figure 4 shows a routesubgraph for the warehouse lay-out of Figure 1 in the case of decentralized depositing. In Ratliff & Rosenthal (1983) an algorithm is given that constructs an orderpicking route from a given routesubgraph. The problem of finding a shortest orderpicking route can therefore be solved by finding a routesubgraph of minimum length.

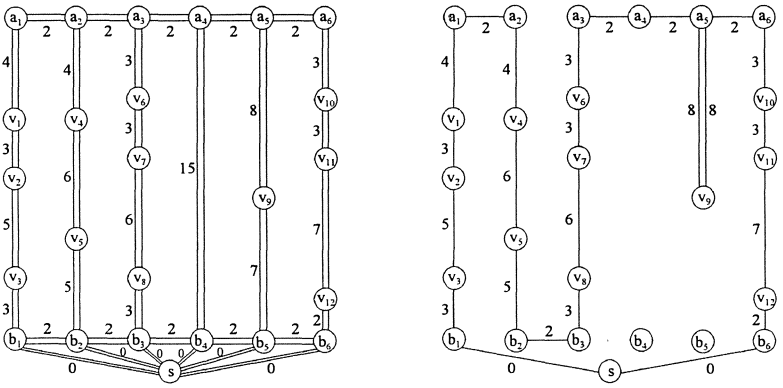


Figure 4: Example of a warehouse graph (left) and a routesubgraph (right).

2.3 Finding a minimum length routesubgraph

In Ratliff & Rosenthal (1983) an algorithm is given that finds a routesubgraph of minimum length for warehouses with a central depot. In De Koster & Van der Poort (1997) this algorithm is extended to find a minimum length routesubgraph in both warehouses with a central depot and warehouses with decentralized depositing. They showed that this extension is necessary since a shortest orderpicking route in a warehouse with decentralized depositing can not be obtained from a shortest orderpicking route in a warehouse with a central depot by leaving out some of the edges. The algorithm determines a shortest routesubgraph by applying dynamic programming (see e.g. Bertsekas (1976)). A short description is given below.

Suppose we have a subgraph of the warehouse graph, consisting of edges and vertices of aisle j (for any $j = 1, ..., n$, where n is the number of aisles in the warehouse) and of edges and vertices of all aisles to the left of this aisle. This subgraph is called an *L_j-partial routesubgraph* if there exists another subgraph, called *completion*, consisting of edges and vertices to the right of aisle j, such that the union of these two subgraphs forms a routesubgraph. Two L_j-partial routesubgraphs are *equivalent* if any completion of one partial routesubgraph is a completion for the other.

The algorithm starts with all L_1-partial routesubgraphs consisting only of vertices and edges of aisle 1. In the next step, L_2-partial routesubgraphs are formed by extending the L_1-partial routesubgraphs with vertices and edges of aisle 2. Continuing this way, we finally get the L_n-partial routesubgraphs, which precisely are the routesubgraphs.

In order to use the concept of dynamic programming, we have to define the potential *states*, the possible *transitions* between states, and the *costs* involved in such a transition. The states correspond to 12 classes of equivalent L_j-partial routesubgraphs. The transitions between states consist of adding

vertices and edges of a new aisle. A transition from aisle $j-1$ to aisle j consists of three steps: first edges between the endpoints of the two aisles are added, in the next step edges from aisle j to and from the depot can be added, and finally edges and vertices within aisle j are added. The cost of each transition is equal to the sum of the lengths of the edges added in the transition.

The algorithm considers all aisles and items, and for each aisle and item a constant number of operations has to be done. Hence, the time-complexity function of the algorithm is linear in the number of aisles and the number of items.

3 A numerical comparison between optimal and heuristic solutions

This section compares the optimal and heuristic solutions in two practical order picking systems; namely a narrow-aisle high-bay pallet warehouse and a shelf area with decentralized depositing of picked items. For each type of warehouse, we consider various combinations of the number of aisles and the number of items to be picked per route. For every such configuration, we generate a number of random orders. The locations of the items in an order are uniformly and independently distributed over the orderpicking area. For each random order, the route length is calculated for both the S-shape heuristic and the optimal algorithm. The difference in average route length is compared by calculating the average percentage improvement in travel time of the optimal algorithm over the S-shape heuristic.

For each simulation experiment, the necessary number of replications needs to be determined such that the estimate for the mean travel time has a relative error smaller than some γ, for $0 < \gamma < 1$. An approximation for the necessary number of replications, such that the relative error is smaller than γ with a probability of $1-\alpha$, is the smallest integer i satisfying

$$i \geq S^2(i)[z_{1-\alpha/2}/\gamma'\overline{X}(i)]^2,$$

where $S_2(i)$ is the sample variance, $z_{1-\alpha/2}$ the $1-\alpha/2$ percentile of the normal distribution, $\overline{X}(i)$ the sample mean and $\gamma' = \gamma/(1+\gamma)$. To obtain a relative error γ smaller than 2% with a probability of 95% for all situations considered in this paper, a replication size of 10,000 is sufficient. Hence, for each warehouse type and every combination of number of aisles and number of items, we generate 10,000 random orders.

We assume that all orders are processed in order of arrival. That is, the route for any new order starts in the aisle where the previous route ended. This is the most straightforward way. If a set of orders has to be picked on a order by order basis, there is an additional opportunity for savings on travel time by finding the best sequence of orders. This can reduce the travel time

needed to go from the aisle where the last route ended to the aisle where the next route starts. In our earlier experiments, it has appeared that the potential reduction in travel time of such sequencing is usually less than 2%. Therefore, the option of sequencing orders is not considered further here.

3.1 Narrow-aisle high-bay pallet warehouse

This kind of order picking systems is often used for the picking of fairly large, not fast moving, items that are stored in pallet areas. The following assumptions are made. The aisles are 50 meters long. The average travel speed within the aisles is 1.5 meters/second and outside the aisles 1 meter/second. The distance between two neighboring aisles is 4.3 meter and the time needed to leave or enter an aisle is 15 seconds. The simulations were carried out for 2 to 10 aisles. Each curve in Figure 5 gives the relationship between the number of items per route and the percentage improvement in travel time of the optimal algorithm over the S-shape heuristic. Each curve in the figure corresponds to a fixed number of aisles (values indicated next to the curves). Solid lines correspond to an even number of aisles, dotted lines to an odd number of aisles.

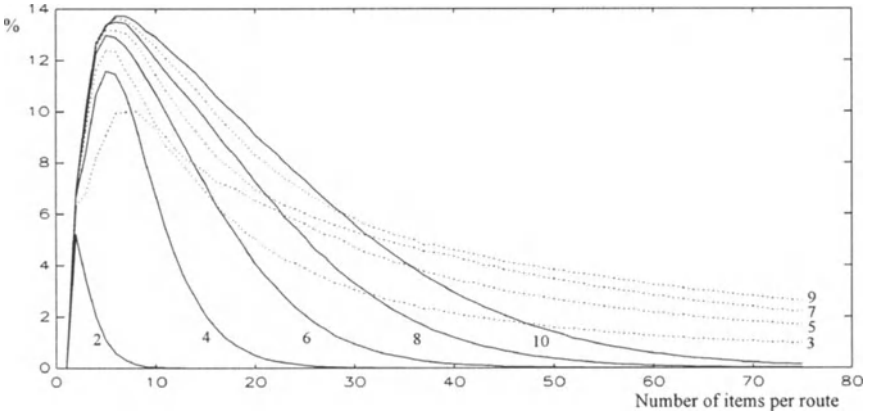


Figure 5: Percentage improvement in travel time of the optimal algorithm over the S-shape heuristic versus the number of pick locations per route for a narrow-aisle high-bay pallet warehouse.

There are three striking features in Figure 5. Firstly, every curve has a single peak. This can be explained as follows. In the case of a single item, the route for the optimal and heuristic strategy are the same. If the number of items increases, more possibilities arise for the optimal algorithm to find efficient routes, that differ significantly from the routes generated by the S-shape heuristic. But if the number of items increases beyond a certain point,

this effect is outweighed by an increasing probability that the optimal way to pick items from an aisle is to traverse the entire aisle. This will make an optimal route look more like a route of the S-shape heuristic, thus making the difference between the two smaller. Secondly, we can see that if the number of aisles increases, then the percentage improvement also increases, though with a decreasing rate. Increasing the number of aisles increases the number of possibilities for routing, thus increasing the probability that the optimal algorithm can find a shorter route than the S-shape heuristic. The effect of the number of aisles is further analyzed in Section 3.3.

Finally, there is a remarkable difference between curves with an odd number of aisles and curves with an even number of aisles: the curves for an odd number of aisles have thicker tails. This is due to the fact that the S-shape heuristic is especially fit for situations with an even number of aisles. If all aisles have to be visited, the heuristic traverses every aisle exactly once in case there is an even number of aisles. If the heuristic has to traverse an odd number of aisles, then the route enters the last aisle from the front end, visits all picking locations and has to return through the same aisle, thus traveling it twice (see e.g. Figure 2). If only one item has to be picked in the last aisle the expected travel time for this aisle is equal to traversing the aisle once. If the total number of items increases, the density of pick locations in the last aisle will rise. Having to travel from the front end to the item farthest away and back to the front end will therefore give an expected travel distance for the last aisle approximating twice the aisle length. Since it only concerns travel in one aisle, the difference in the thickness of the tails tends to disappear if the number of aisles increases.

3.2 Picking in a shelf area with decentralized depositing

We now consider a shelf store area where order pickers pick items in batch with a small pick cart. They receive pick instructions via a mobile terminal, with possibly a label printer. At the front end of the aisles, there is a conveyor where the picked items can be dropped off. After dropping off the picked items, the pickers receive information over the items to be picked in the next route. The following assumptions are made. The aisles have a length of 10 meters. The average walking speed within and outside the aisles is 0.6 meters per second. The distance between two neighboring aisles is 2.4 meters and no additional time is needed for aisle changing. The simulations were carried out for 2 to 10 aisles.

Each curve in Figure 6 gives the relationship between the number of items per route and the percentage improvement in travel time of the optimal algorithm over the S-shape heuristic. Each curve in the figure corresponds to a fixed number of aisles (values indicated next to the curves). Solid lines correspond to an even number of aisles, dotted lines to an odd number of aisles.

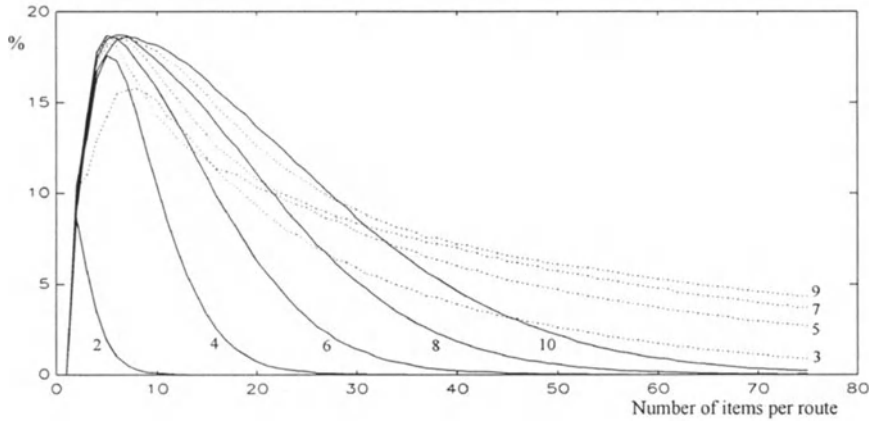


Figure 6: Percentage improvement in travel time of the optimal algorithm over the S-shape heuristic versus the number of pick locations per route for a shelf area with decentralized depositing of picked items.

The shape of the curves in Figures 5 and 6 are very similar. Only the peaks are about 5% higher in Figure 6. The main difference between the two situations is the time needed in the narrow-aisle high-bay pallet warehouse to change aisles. Therefore, we evaluate in Section 3.4 the effects of aisle change times more closely.

3.3 Effects of the number of aisles

The same warehouse lay-out is considered as in Section 3.2. Only situations with an even number of aisles are used in order to keep the figure clear. Similar results can be obtained for an odd number of aisles. Simulations are carried out for 6, 10, 14, ... , 30 aisles. Each curve in the figure corresponds to a fixed number of aisles (values indicated next to the curves).

From Figure 6 it is known that the maximum gain of the optimal algorithm over the heuristic increases if the number of aisles increases from 2 to 5. Figure 7 shows that the maximum gain stabilizes for more than 5 aisles. The locations of the peaks in Figure 7 show a trend. For 6 aisles the peak is located at 5 items, for 10 aisles at 7 items, for 14 aisles at 11 items etc. That is, the location of the peak shifts to a higher number of items if the number of aisles increases. To explain this, consider the two following opposite effects. First, if the number of items to be picked increases, then the expected number of aisles containing at least one pick location rises. This gives more possibilities for the optimal algorithm to improve on the S-shape heuristic. On the other hand, an increase in the number of items leads to a higher probability that an aisle has to be traversed entirely. Therefore, the

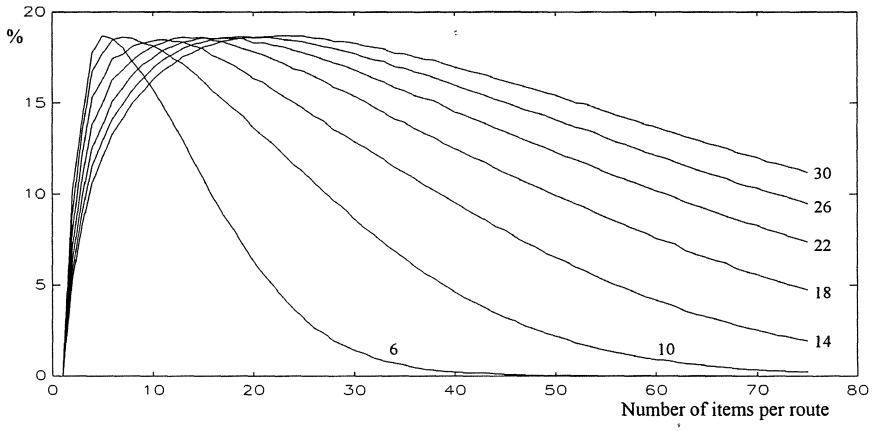


Figure 7: Percentage improvement in travel time of the optimal algorithm over the S-shape heuristic versus the number of pick locations per route for a shelf area with decentralized depositing.

highest potential to improve on the S-shape heuristic is in situations where the number of items to be picked is close to, but strictly smaller than the number of aisles.

3.4 Varying aisle change times

Here we consider the same warehouse lay-out as in Section 3.1. The picking area consists of 6 aisles. The time needed to leave or enter an aisle is varied between 0 and 25 sec. Each curve in Figure 8 gives the relationship between the number of items per route and the percentage improvement in travel time of the optimal algorithm over the S-shape heuristic for a fixed aisle change time. Each curve corresponds to a fixed time to enter or leave an aisle (values in seconds indicated next to the curves). Similar results can be obtained for a different number of aisles.

From Figure 8 the impact can be seen of the aisle change time. If the time needed to change aisles increases, the gains of the optimal algorithm decrease drastically. However, the location of the peak is not influenced.

3.5 Concluding remarks

The numerical results suggest that the savings in travel time may be substantial when using the optimal algorithm instead of the S-shape heuristic. However, the algorithm is more complex than the S-shape heuristic. In view of the extra expenses and risks of implementing the optimal algorithm, we consider situations in which it is really worthwhile to implement the optimal algorithm. That is, situations for which the improvement of total route time

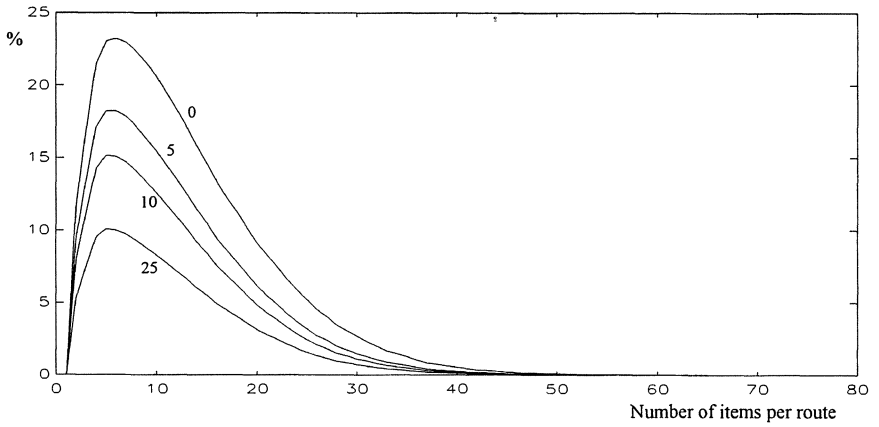


Figure 8: Percentage improvement in travel time of the optimal algorithm over the S-shape heuristic versus the number of pick locations per route for a narrow-aisle high-bay pallet warehouse with 6 aisles.

is substantial, i.e. more than 5%. The travel time usually amounts substantially to total route time, in Tompkins *et al.* (1996) this is estimated on 50% of total route time. Therefore, the gains in travel time of the optimal algorithm over the S-shape heuristic should for practical purposes be over 10%.

It can be concluded from the numerical results that the proportional gain in average travel time strongly depends on the average number of items per aisle. Situations in which there are many orders containing just one or two items will profit little from introducing the optimal algorithm. For a narrow-aisle high-bay pallet warehouse gains in travel time will fall below 10% if the number of pick locations per route is more than twice the number of aisles. In case of a shelf area with decentralized depositing this point is roughly if the number of pick locations is three times the number of aisles.

There are also other factors that influence the proportional gain in travel and total time, for instance the time needed to leave or enter an aisle. If this time increases, then it will become increasingly unattractive to enter aisles more than once, and consequently the difference in travel time between the optimal and heuristic solutions will decrease. The time needed to enter aisles used in the simulation experiments of Section 3.1 is chosen realistically for the case of order picking trucks. If aisle-changing cranes are used instead of trucks then the time for entering aisles will be substantially higher, thus drastically diminishing the need to use an optimal algorithm in stead of the S-shape heuristic.

4 Explicit statistical formulas for estimating route length

In Section 3 we have seen that there are many situations in which the optimal algorithm will perform significantly better than the S-shape heuristic. But there are also situations, especially those with very few or many items, in which the optimal algorithm is less useful. In principle, it is possible to perform simulations for all possible configurations, that may be encountered in a specific practical case. However, simulations are very time consuming, both in creating computer programs and calculation times. It would therefore be desirable to have explicit formulas available to get a quick first impression of possible gains.

In Hall (1993) statistical formulations are given for various heuristics and the optimal algorithm. These formulas are designed for the case of a central depot and random distribution of pick locations over aisles and locations. We adapt them for the case of decentralized depositing. Note that statistical formulas are often mainly fit for large numbers. Consequently, these formulas are likely to perform better for warehouse situations with a larger number of items and aisles. In practice, a picking zone will consist of only a few aisles. Therefore, we analyze the performance of the formulas for 2 up to 10 aisles.

4.1 An adaptation for decentralized depositing

In this section we adapt existing formulas for estimating route length from Hall (1993) for the the case of decentralized depositing. We define the following variables:

m = number of items to be picked,
n = number of aisles in the warehouse,
x = distance between the heads of aisle 1 and N,
y = length of an aisle,
s_x = travel speed outside the aisles,
s_y = travel speed within aisles,
t_a = time needed to leave or enter an aisle.

Any formulation for a distance D given in Hall (1993) is denoted by $\tilde{D}$. Distances concerning the optimal algorithm are denoted with an asterisk. Distances traveled outside the aisles are indexed with an x, distances traveled inside the aisles are indexed with an y. For example, $\tilde{D}_y^*$ denotes the distance traveled inside the aisles for the optimal algorithm, according to a formulation from Hall (1993).

The expected values of the minimum, and maximum, of m continuous uniformly distributed [0,1] variables $\{X_i\}$ are well known, and equal:

$$E[\min\{X_1, X_2, ..., X_m\}] = 1/(m+1) \tag{1}$$

$$E[\max\{X_1, X_2, ..., X_m\}] = m/(m+1) \tag{2}$$

Under the assumption that items are distributed uniformly over aisles and locations, we know from Hall (1993) that for the S-shape heuristic the number of aisles containing items, denoted by A, has an expected value of:

$$E[A] = n \cdot \left(1 - \left(\frac{n-1}{n}\right)^m\right),$$

which is n times the probability that an aisle contains at least one item.

The expected value for the distance traveled inside the aisles is given by:

$$E[\tilde{D}_y] = y \cdot E[A] + y/2 \tag{3}$$

In Hall (1993) it is assumed that an order picker always traverses an aisle entirely, that is even if the last aisle is entered from the front end, it is traversed entirely to the rear end and thereafter back through the same aisle to the front end. Thus, the last aisle is traversed twice if the number of aisles containing items is odd. In Equation (3), the distance traveled inside the aisles is adjusted for this excess travel in the last aisle by a correction term of $y/2$. This is based on the assumption that the number of aisles containing at least one item is odd or even with equal probability. We now make a different assumption for travel in the last aisle: after picking the last item the picker returns directly to the front end, without first going to the rear end. The correction term has to be adjusted accordingly. Let $E[I] = m/E[A]$ be the expected number of items in an aisle given the fact that the aisle contains at least one item. With Equation (2) we can determine the expected location of the item farthest from the front end of the aisle to be approximately $E[I]/(E[I]+1)$. Thus, the expected travel time in the last aisle if the number of aisles is odd, is given by:

$$E[T] = 2y\frac{E[I]}{E[I]+1}.$$

Formula (3) already accounts for one aisle traversal of length y. The expected excess travel in the last aisle if the number of aisles to visit is odd, is therefore:

$$E[T] - y \tag{4}$$

Given an equal probability that the number of aisles is odd or even, gives the correction term:

$$\tfrac{1}{2}(E[T] - y) \tag{5}$$

Thus, the expected travel inside the aisles is:

$$E[D_y] = y \cdot E[A] + \tfrac{1}{2}(E[T] - y) \tag{6}$$

Using Equations (1) and (2), the expected distance traveled outside the aisles in a warehouse with central depositing can be found to equal:

$$E[\tilde{D}_x] \approx 2x \cdot \frac{m-1}{m+1},$$

which is twice the expected distance between the left and right most aisle containing items. For the case of decentral depositing, a straightforward approximation for the distance traveled outside the aisles is:

$$E[D_x] = \tfrac{1}{2}E[\tilde{D}_x],$$

which is the distance traveled outside the aisles between the expected location of the left and right most aisles containing items.

For small values of m, an adjustment factor is given in Hall (1993) to account for the expected distance from the central depot to the nearest aisle containing a pick location. For the case of decentral depositing we need an adjustment factor to account for the expected travel from the end aisle of the previous route to the first aisle of the next route. We have $E[D_x]$ as an estimate for the distance between the left and right most aisle containing items. The total distance between the heads of the aisles 1 and n is equal to x. Therefore, the endpoints of a route can vary over a distance of $x - E[D_x]$. The start point of the new route and end point of the previous route are uniformly distributed over this distance. Treating aisle locations as continuous random variables and using Equations (1) and (2), an approximation for the expected distance between the location of the two points is:

$$\tfrac{1}{3}(x - E[D_x]).$$

The time needed to leave and enter an aisle is $2t_a$ and the expected number of aisles containing at least one item is $E[A]$. Thus, the total time needed for entering and leaving aisles for the route is approximated by:

$$2t_a E[A].$$

Total travel time for the S-shape heuristic in case of decentral depositing can now be estimated by:

$$\begin{aligned} E[T_{S-shape}] &= \frac{E[D_y]}{s_y} + \frac{E[D_x]}{s_x} + 2t_a \cdot E[A] + \frac{\frac{1}{3}(x - E[D_x])}{s_x} \\ &= \frac{E[D_y]}{s_y} + \frac{2E[D_x]}{3s_x} + 2t_a \cdot E[A] + \frac{x}{3s_x} \end{aligned} \tag{7}$$

For the optimal algorithm Hall (1993) gives as lower bound for the distance traveled inside the aisles:

$$E[\tilde{D}_y^*] = ny \cdot \sum_{i=1}^{m} \binom{m}{i} \left(\frac{1}{n}\right)^i \left(\frac{n-1}{n}\right)^{m-i} \left[1 - \left(\frac{1}{2}\right)^i\right] \tag{8}$$

This formulation is based on the assumption that the distance traveled is minimized for each aisle individually, without considering any consequences for the total route. The term between square brackets is the expected fraction of the aisle length traveled in an aisle containing i items. The remaining part in the summation gives the probability that an aisle contains i items. Thus, the summation gives the expected fraction of the aisle traveled.

For travel outside the aisles the same approximation is used as for the S-shape heuristic, that is:

$$E[D_x^*] = E[D_x],$$

To compensate for travel from the end aisle of the previous route to the first aisle of the following route we use the same correction term as for the S-shape heuristic. Since the number of times an aisle has to be entered is minimal for S-shape, we have that the time needed to leave and enter aisles is at least as large for the optimal algorithm as for the S-shape heuristic. Therefore, a conservative estimate for the optimal algorithm is to set the time needed for entering and leaving aisles equal to that of the S-shape heuristic. Total travel time for the optimal algorithm can thus be estimated by:

$$E[T_{optimal}] = \frac{E[\tilde{D}_y^*]}{s_y} + \frac{2E[D_x^*]}{3s_x} + 2t_a E[A] + \frac{x}{3s_x}. \qquad (9)$$

4.2 A comparison between simulation and formulas

We use the warehouse situation described in Section 3.1 to compare the results from simulations with the results from the explicit formulas. The number of aisles is varied from 2 to 10. The percentage difference between the result of simulation and the answer of Formula (9) for the optimal algorithm is calculated as $(E[T_{optimal}] - L_{simulation})/L_{simulation}$, where $L_{simulation}$ gives the average travel time of routes generated by the optimal algorithm in the simulation. Each curve in Figure 9 gives the relationship between the number of items per route and the percentage difference for a fixed number of aisles. The number of aisles varies from 2 to 10 (values 3, 5 and 7 indicated next to the curves).

Figure 9 reveals two undesirable features of the explicit formula. First, for few (5 or less) items the formula gives answers of 10% to 25% below the simulated values. The explanation for this is simple. For example, if one item has to be picked, the formula assumes a distance to be traveled within the aisle equal to half the aisle length. That is, it is possible to enter the aisle from both the front and the rear end, dependent on which is the shortest. However, a real route will have to enter the aisle from the front end, go to the pick location and return to the front end. If the item location is uniformly distributed over the aisle this gives an expected distance equal to the aisle length. Similar effects occur for other low values of the number of items. In

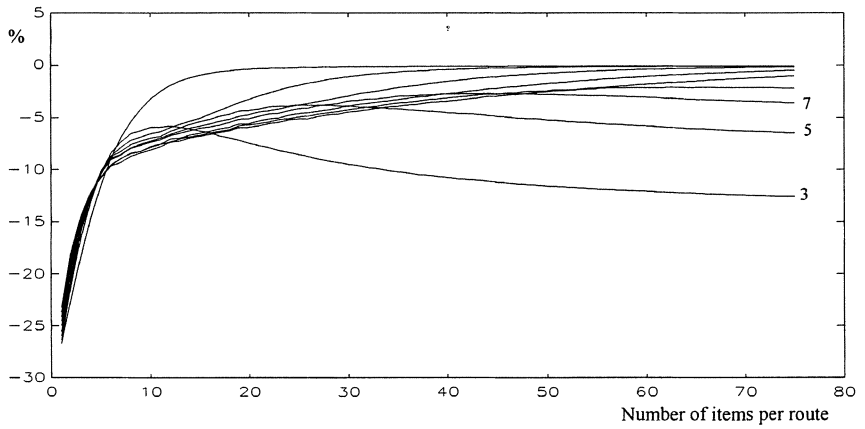


Figure 9: Percentage difference between the results of simulation and the explicit Formula (9) versus the number of pick locations per route for the optimal algorithm in a narrow-aisle high-bay pallet warehouse.

case the number of aisles is odd, the formula underestimates the travel time if the number of items rises. This can be explained as follows. The formula assumes that for each aisle individually the distance is minimized. In case of a large number of items this will typically lead to traversing each aisle entirely. A route traversing an odd number of aisles ends up at the rear end of the warehouse instead of at the front end. The optimal algorithm does return to the front end and consequently gives a longer distance. The magnitude of the effect will diminish if the number of aisles increases, because the contribution to total route length is smaller.

In Figure 10 the relationship between the number of items per route and the percentage difference between Formula (7) and the simulation for the S-shape heuristic is given. Each curve corresponds to a fixed number of aisles varying from 2 to 10 (values indicated next to the curves).

The formula tends to underestimate the average length if the number of aisles is odd and to overestimate it if the number of aisles is even. For example, consider a warehouse with 3 aisles. If the number of items to be picked is high, then an orderpicking route will typically traverse the first and second aisle and go up and down the third aisle. If the number of items increases, the total distance traveled within the aisles will approach 4 times the aisle length. If the warehouse has two aisles and there are many items to be picked, a route will traverse the two aisles exactly once. The formula accounts for three, respectively two, aisle traversals plus the correction term from Formula (5) for excess travel in the last aisle. This correction adds the expected excess distance traveled in the last aisle multiplied by a probability of $\frac{1}{2}$. In fact this correction should not be added for two aisles, whereas it

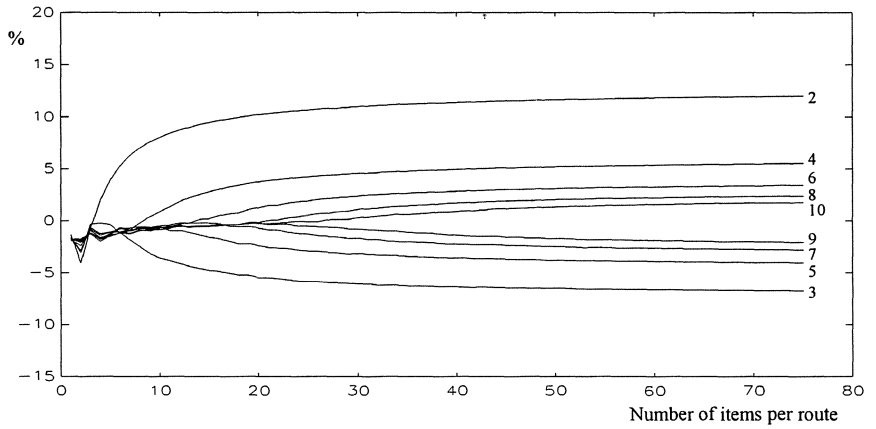


Figure 10: Percentage difference between the results of simulation and the explicit Formula (7) versus the number of pick locations per route for the S-shape heuristic in a narrow-aisle high-bay pallet warehouse.

should be added with probability 1 for the case of 3 aisles with a large number of items. This type of correction will only work accurately if the number of aisles is high (at least more than 10). The small negative peak located at 2 items is due to the same effect. That is, there is a high probability that both items are in one aisle. Having a high probability that the number of aisles is odd, we know that the formula will underestimate the route length.

4.3 Improvement for S-shape heuristic

To cure the problems we encountered in the previous section we change the correction term. Suppose k aisles have to be traversed, where k is odd. Furthermore, suppose that the total number of items to be picked is m. The average number of items in an aisle is then m/k. Following the same line of argument we used to obtain Formula (4), we find that the distance traveled within the aisles should be adjusted by:

$$C_k = 2y \cdot \frac{\frac{m}{k}}{\frac{m}{k}+1} - y,$$

which is the average excess distance traveled in the last aisle.

If there are n aisles, then the probability that all items are in exactly k aisles, is given by:

$$P_k = \binom{n}{k}\left(\frac{k}{n}\right)^m \cdot \left[1 - \sum_{i=1}^{k-1}(-1)^{i+1}\binom{k}{k-i}\left(\frac{k-i}{k}\right)^m\right].$$

In the formula above, the left part of P_k is the probability that the m items are in at most k aisles. The term between square brackets is 1 minus the probability that all items are in $k-1$ or less aisles out of a given set of k aisles. To obtain this, use is made of the inclusion-exclusion rule. The new correction term can now be formulated as:

$$C^+ = \sum_{k \in S} P_k \cdot C_k, \text{ where } S = \{i \leq n \mid i \text{ is odd }\}. \tag{10}$$

The estimate for travel within the aisles for the S-shape heuristic can now be restated as:

$$E[D_y^+] = y \cdot E[A] + C^+. \tag{11}$$

Using Formula (11) instead of Formula (6) gives significantly better results as can be seen from Figure 11. In the figure, each curve corresponds to a fixed number of aisles varying from 2 to 10.

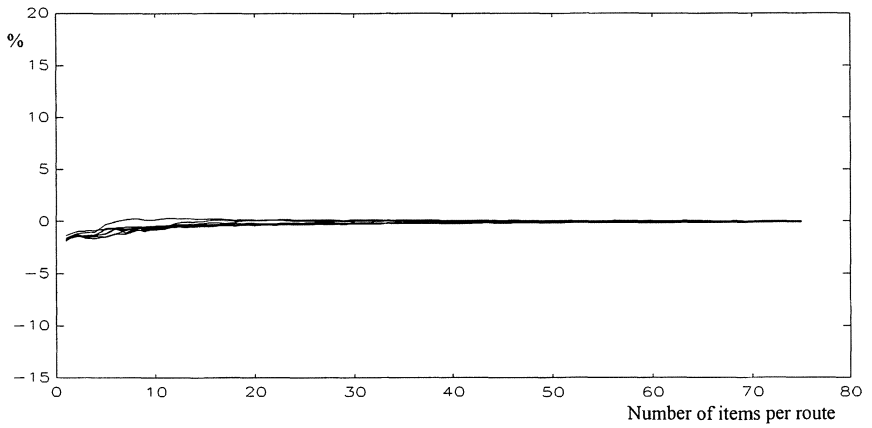


Figure 11: Percentage difference between the results of simulation and the explicit Formula (7) (with $E[D_y]$ replaced by $E[D_y^+]$) versus the number of pick locations per route for the S-shape heuristic in a narrow-aisle high-bay pallet warehouse.

4.4 Improvement for the optimal algorithm

Clearly, it would be also desirable to improve the performance of Formula (9) for the optimal algorithm. Since the optimal algorithm has a structure that is far more complex than the S-shape heuristic, this is not just as simple. Firstly, consider the following. From Section 4.1 we know that $[1 - (\frac{1}{2})^i]$ is the expected fraction of the aisle length traveled in an aisle containing i items, if the aisle is traveled optimally. The expected number of items in an

aisle is given by $E[I] = m/E[A]$. Thus, the expected fraction of the aisle traveled can be estimated by:

$$1 - \left(\tfrac{1}{2}\right)^{E[I]}$$

and consequently the total distance traveled inside all aisles by:

$$E[D_y^*] = E[A] \cdot y \cdot \left[1 - \left(\tfrac{1}{2}\right)^{E[I]}\right] \tag{12}$$

Formula (12) is clearly easier to use than Formula (8). From comparison with simulated values it appears that Formula (12) gives slightly better results than Formula (8).

We propose two improvement factors. First, we intend to decrease the large deviation for a small number of items. Consider any orderpicking route visiting two or more aisles. In Formula (12) it is assumed that for every aisle individually the optimal traversal strategy is chosen. But, the route has to start and finish at the front end of the warehouse. Therefore, in two aisles the transitions (3) and (5) of Figure 3 are not possible. We first determine the expected distance traveled in an aisle where only transitions (1) and (4) are possible (Hall (1993) already excludes transition (2) and transition (6) is not necessary since the aisle is considered to be non-empty). Suppose there are i items in this aisle. If all items are in the front half of the aisle, then transition (4) is used. This gives an expected length of

$$y \cdot \frac{i}{i+1}.$$

This event has a probability of $\left(\tfrac{1}{2}\right)^i$. With probability $1 - \left(\tfrac{1}{2}\right)^i$ there is at least one item in the rear half of the aisle. In this case the shortest way to pick the items is to traverse the entire aisle, with length y. The expected number of items in an aisle is $E[I]$. Thus, the expected distance traveled in an aisle, where only transitions (1) and (4) are possible, can be estimated by:

$$E[D_y^{(1),(4)}] = y \cdot \left(\frac{1}{2}\right)^{E[I]} \cdot \frac{E[I]}{E[I]+1} + y \cdot \left[1 - \left(\frac{1}{2}\right)^{E[I]}\right]$$

Formula (12) already accounts for a distance of:

$$y \cdot \left[1 - \left(\frac{1}{2}\right)^{E[I]}\right].$$

Therefore, the adjustment needed in the two aisles, where only transitions (1) and (4) are possible, is for each of these two aisle equal to:

$$E[D_y^{(1),(4)}] - y \cdot \left[1 - \left(\frac{1}{2}\right)^{E[I]}\right] = y \cdot \left(\frac{1}{2}\right)^{E[I]} \cdot \frac{E[I]}{E[I]+1}.$$

All items are in one aisle with probability:

$$P^{(4)} = n \cdot \left(\frac{1}{n}\right)^m .$$

In this case, the whole route consists of one time transition (4), with expected length:

$$E[D_y^{(4)}] = 2y \cdot \frac{m}{m+1}.$$

The expected excess distance is:

$$E[D_y^{(4)}] - E[D_y^*].$$

Combining the previous formulas gives the correction term:

$$C_1^* = P^{(4)}\left(E[D_y^{(4)}] - E[D_y^*]\right) + 2\left(1 - P^{(4)}\right)\left(y\left(\frac{1}{2}\right)^{E[I]} \cdot \frac{E[I]}{E[I]+1}\right). \quad (13)$$

Next we consider the underestimation of Formula (9) for an odd number of aisles and a large number of items. The effect appears to be similar to what we encountered with the S-shape heuristic. However, using Formula (10) as a correction proves to be nearly as bad as not correcting at all. Therefore, we seek for a different solution. The correction is needed for the cases, with an odd number of aisles and a high probability that all aisles are traversed.

The probability that an aisle is traversed entirely is given by:

$$P[traversal] = \sum_{i=1}^{m} \binom{m}{i} \left(\frac{1}{n}\right)^i \left(\frac{n-1}{n}\right)^{m-i} \cdot \left[1 - \prod_{k=0}^{i-1}\left(\frac{1}{2} + \frac{1}{2k+2}\right)\right]$$

For a proof see Appendix A.

As a correction term we state:

$$C_2^* = (P[traversal])^n \left[2y \cdot \frac{\frac{m}{n}}{\frac{m}{n}+1} - y\right],$$

where $(P[traversal])^n$ is an estimate for the probability that it is optimal for every aisle to traverse it entirely and the term between square brackets an estimate for excess travel in the last aisle (based on the S-shape heuristic).

We can now state the new estimate for travel inside the aisles for the optimal algorithm as:

$$E[D_y^{**}] = E[D_y^*] + C_1^* + C_2^*. \quad (14)$$

To avoid interference of the two correction terms, we could multiply C_1^* by $1-(P[traversal])^n$. However, since the influence on the results has appeared to be negligible, this is left out.

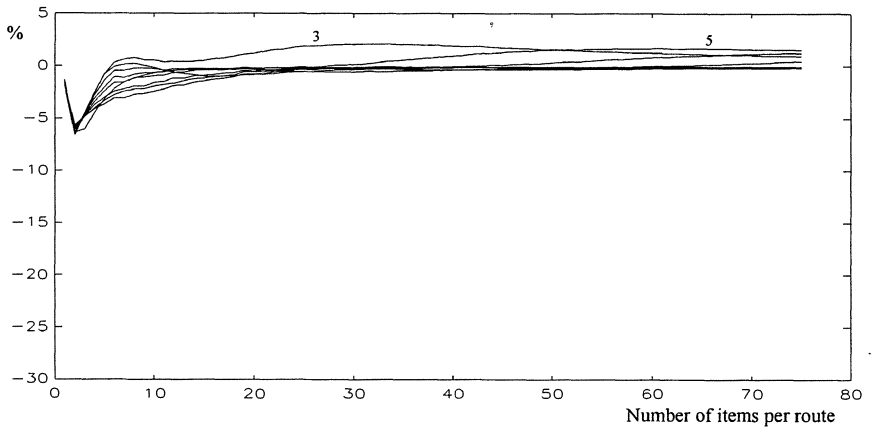


Figure 12: Percentage difference between the results of simulation and the explicit Formula (9) (with $E[\tilde{D}_y^*]$ replaced by $E[D_y^{**}]$) versus the number of pick locations per route for the optimal algorithm in a narrow-aisle high-bay pallet warehouse.

Figure 12 gives the relationship between the number of items per route and the percentage difference between Formula (9) (with $E[\tilde{D}_y^*]$ replaced by $E[D_y^{**}]$) and the simulation for the optimal algorithm. In the figure, each curve corresponds to a fixed number of aisles varying from 2 to 10.

If we compare Figure 12 with Figure 9 the effects of the corrections can be seen. The high deviation for a small number of items has decreased and the performance for an odd number of aisles has improved (even though it is slightly positive now, whereas it was negative).

5 A practical comparison between optimal and heuristic solutions

In practice, the optimal routing algorithms are not frequently used. This is partly caused by the fact that the algorithm is not widely known among engineers that design warehouses and warehouse information systems. Also, any application to lay-outs different than with parallel aisles and a central depot requires considerable non-trivial alterations of the algorithm (see e.g. De Koster & Van der Poort (1997)).

In order to use the algorithm, it has to be incorporated in the existing warehouse management system software, which means a change into the heart of such a system. Usually, it will be difficult to add a new routing algorithm to an existing system. Also, there are extra expenses and risks of implementing the optimal algorithm, because the algorithm is complex (based on dynamic

programming) and not transparent.

Since the optimal algorithm is complex, calculation time will increase compared to a heuristic. The S-shape heuristic considers for each aisle two possibilities: traverse it or do not traverse it. The optimal algorithm updates 12 equivalence classes for every subsequent aisle, choosing one of the transitions from Figure 3. Therefore, the optimal algorithm can usually be expected to require at least 12 times more calculation time than the S-shape heuristic. This is confirmed by the simulations described in this paper. However, solving one instance with the optimal algorithm generally requires a few milliseconds (on a P166 computer). This is not likely to cause any problems if calculations in a warehouse have to be performed on-line.

A heuristic like the S-shape is applicable in all situations and has predictable results. Savings of using optimal algorithms in practice are difficult to forecast. Furthermore, if heuristics and optimal algorithms are compared for practical problems we have to look for differences in total route time. Total route time includes travel time, but also the time needed for other order picking activities. These activities include positioning the truck or crane at the pick location, identifying the location and the product on the location, picking the proper quantity from the location, confirming the pick on a pick list or mobile terminal and putting the picked items on a product carrier. In practical situations the travel time is often about 50% of the total route time (see e.g. Tompkins *et al.* (1996)). Since the optimal algorithm only minimizes travel times, the savings on total route time may be considerably smaller than the savings on travel time reported in Section 3, depending on the exact amount the travel time contributes to total travel time.

From Section 3 and other studies (e.g. Gelders & Heeremans (1994)) it is known that substantial gains can be obtained with the optimal algorithm compared to S-shape. However, the gains depend strongly on the warehouse type and the materials handling equipment used. For example, aisle-change times have a significant effect on the possible improvement by the optimal algorithm (cf. Section 3.4).

6 Conclusion

From the simulation experiments we found rough guidelines when to apply the optimal algorithm and when to apply the S-shape heuristic. If the number of items per route is on average very low (1 or 2), then optimal routing will not give a significant gain in total route time. If the number of items is larger than two or three times the number of aisle, then the gain of the optimal algorithm will also be too small to be of practical use. In between of these two extremes, optimal routing may be of use. In those cases the usage of the optimal algorithm can lead to substantial savings in total route time and hence in the number of pickers and order picking devices needed.

To obtain a first indication of possible gains use can be made of sta-

tistically based, explicit formulas. The existing explicit formulas (see Hall (1993)) give for the case of a central depot a good indication for the actual performance of the optimal algorithm and the S-shape heuristic as long as the number of aisles and the number of pick locations is large (both greater than 10). We adapted these formulas for the case of decentralized depositing and corrected some of the problems that were encountered in case the number of aisles or the number of items was less than 10. These adapted formulas can prove to be useful in the process, in which there has to be decided between optimal or heuristic routing.

Appendix A

Probability that it is optimal to traverse an aisle entirely

Consider one aisle. Without loss of generality we assume that the aisle has length 1. Define the largest gap Y_i of this aisle as the largest distance in the aisle not containing items, given that the aisle contains i items. If $Y_i < \frac{1}{2}$, then traversing the entire aisle is the optimal way to pick the items in the aisle, otherwise either transition (3), (4) or (5) of Figure 3 is optimal. From Equation (A1) in the appendix of Hall (1993) we can derive the following probability:

$$P(Y_{i+1} > \tfrac{1}{2} \mid \text{item } i+1 \text{ falls within largest gap and } Y_i = \tilde{y} > \tfrac{1}{2}) = 2\tilde{y} - 1. \quad \text{(i)}$$

Furthermore it can be derived that:

$$P(Y_{i+1} > \tfrac{1}{2} \mid \text{item } i{+}1 \text{ falls outside largest gap and } Y_i = \tilde{y} > \tfrac{1}{2}) = 1{-}\tilde{y}. \quad \text{(ii)}$$

Therefore:

$$P(Y_{i+1} > \tfrac{1}{2} \mid Y_i = \tilde{y} > \tfrac{1}{2}) = \tilde{y}. \quad \text{(iii)}$$

Given that the new gap is still larger than $\frac{1}{2}$, we have that the expected size of the new gap is:

$$E(Y_{i+1} \mid Y_i = \tilde{y},\ \text{item } i+1 \text{ falls within the largest gap}, Y_{i+1} > \tfrac{1}{2}) = \tilde{y} - \tfrac{1}{2}(\tilde{y} - \tfrac{1}{2}) = \tfrac{1}{2}\tilde{y} + \left(\tfrac{1}{4}\right). \quad \text{(iv)}$$

$$E(Y_{i+1} \mid Y_i = \tilde{y},\ \text{item } i+1 \text{ falls outside the largest gap}, Y_{i+1} > \tfrac{1}{2}) = \tilde{y}. \quad \text{(v)}$$

Using Equations (i) - (v) we obtain:

$$E[Y_{i+1} \mid Y_i = \tilde{y}, Y_{i+1} > \tfrac{1}{2}] = \frac{2\tilde{y}-1}{\tilde{y}} \cdot (\frac{1}{2}\tilde{y} + \frac{1}{4}) + \frac{1-\tilde{y}}{\tilde{y}} \cdot \tilde{y} = 1 - \frac{1}{4\tilde{y}}.$$

It is easy to calculate that $E[Y_1] = \frac{3}{4}$. Now it is a straightforward exercise to prove with induction that:

$$E[Y_i | Y_i > \tfrac{1}{2}] = \frac{1}{2} + \frac{1}{2i+2}. \tag{vi}$$

Standard probability theory yields:

$$\begin{aligned} P(Y_{i+1} > \tfrac{1}{2}) = P(Y_{i+1} > \tfrac{1}{2} \text{ and } Y_i > \tfrac{1}{2}) = \\ P(Y_i > \tfrac{1}{2}) \cdot P(Y_{i+1} > \tfrac{1}{2} | Y_i = \tilde{y} > \tfrac{1}{2}). \end{aligned} \tag{vii}$$

Using Equation (iii) we obtain:

$$P(Y_{i+1} > \tfrac{1}{2} | Y_i = \tilde{y} > \tfrac{1}{2}) = \tilde{y} = E[Y_i | Y_i > \tfrac{1}{2}]. \tag{viii}$$

It is easy to calculate that $P(Y_1 > \frac{1}{2}) = 1$.
With Equations (vi) - (viii) and using induction, it can be derived that:

$$P(Y_i > \tfrac{1}{2}) = \prod_{k=0}^{i-1} \left(\frac{1}{2} + \frac{1}{2k+2} \right).$$

The probability that traversing the entire aisle is optimal, given that there are i items in the aisle, is:

$$P(Y_i \leq \tfrac{1}{2}) = 1 - P(Y_i > \tfrac{1}{2}) = 1 - \prod_{k=0}^{i-1} \left(\frac{1}{2} + \frac{1}{2k+2} \right). \tag{ix}$$

The probability that an aisle is traversed entirely can now be stated as:

$$P[traversal] = \sum_{i=1}^{m} \binom{m}{i} \left(\frac{1}{n} \right)^i \left(\frac{n-1}{n} \right)^{m-i} P(Y_i \leq \tfrac{1}{2}),$$

where the first part of the summation is the probability that the aisle contains i items and the second part is given by Equation (ix).

References

Bertsekas, D.P. (1976): Dynamic programming and stochastic control, Academic Press.

Carlier, J. / Villon, P. (1987): A well-solved case of the traveling salesman problem, Technical Report, University of Compiegne.

De Koster, R. / Van der Meer, J.R. (1997): Centralized versus decentralized control of internal transport, a case study, Advances in Distribution Logistics, (Springer) Berlin.

De Koster, R. / Van der Poort, E. / Wolters, M. (1997): Efficient order-batching methods in warehouses, Working paper, Rotterdam School of Management.

De Koster, R. / Van der Poort, E. (1997): Routing orderpickers in a warehouse: A comparison between optimal and heuristic solutions, Working paper, Rotterdam School of Management.

Drury, J. (1988): Towards more efficient orderpicking, IMM Monographs 1, Institute of Materials Management, Cranfield, UK.

Gelders, L. / Heeremans, D. (1994): Het traveling salesman probleem toegepast op orderpicking ('The traveling salesman problem applied to orderpicking'), Tijdschrift voor Economie en Management 19, 381-388.

Gibson, D.R. / Sharp, G.P. (1992): Order batching procedures, European Journal of Operations Research 58, 57-67.

Hall, R.W.H. (1993): Distance approximations for routing manual pickers in a warehouse, IIE Transactions 4, Vol. 25, 76-87.

Little, J.D.C. / Murty, K.G. / Sweeney, D.W. / Karel, C. (1963): An algorithm for the traveling salesman problem, Operations Research 11, 972-989.

Ratliff, H.D. / Rosenthal, A.S. (1983): Orderpicking in a rectangular warehouse: A solvable case of the traveling salesman problem, Operations Research 31, 507-521.

Rosenwein, M.B. (1996): A comparison of heuristics for the problem of batching orders for warehouse selection, International Journal of Production Research 34(3), 657-664.

Tompkins, J.A. / White, J.A. / Bozer, Y.A. / Frazelle, E.D. / Tanchoco, J.M.A. / Trevino, J. (1996): Chapter 9 in Facilities planning, (John Wiley & sons, inc.) New York.

Van Dal, I.C. (1992): Special cases of the traveling salesman problem, Ph.D.-thesis University of Groningen.

Centralized versus Decentralized Control of Internal Transport, a Case Study

René de Koster[1] and J. Robert van der Meer[1]

[1] Rotterdam School of Management, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Abstract. Since the introduction of wireless truck terminals and the translation of this technology in material handling control systems, a new control area has emerged in the control of such trucks. In this paper special attention is devoted to Fork Lift Trucks (FLTs) with wireless truck terminals which are controlled by a central Warehouse Management System (WMS). In order to justify investments in wireless truck terminals, it is necessary to specify the reduction in the number of vehicles needed and to indicate the impact on response times and throughput times. This is investigated for the case of a distribution center, where a wireless truck terminal system has been introduced. Two situations have been compared via simulation: decentralized conventional control (without mobile terminals) and centralized control with a WMS using so-called work lists. It is shown that, a centralized control system outperforms the conventional control systems. Such a system leads to a 29 % reduction of the number of FLTs needed, and a simultaneous reduction in pallet response times. Furthermore, warehouse performance is almost insensitive to the structure of the work lists when centralized control systems with work lists are used.

Keywords. Vehicle control, forklift truck, warehousing, work lists

1 Introduction

In warehouses and manufacturing plants, use is made of internal transportation equipment to transport material between locations. Such locations may be an unloading dock door and a storage location, or a storage location and a workcenter. The most commonly used internal transport equipment are conveyors and vehicles. Conveyors are often used when large volumes of conveyable goods have to be transported over rather short distances. Conveyors are static and have little flexibility. Vehicles such as Fork Lift Trucks (FLTs) or Automated Guided Vehicles (AGVs) have higher flexibility in routing and in the material that can be transported. Since the introduction of wireless truck terminals and the translation of this technology in material handling control systems, it has become apparent that there are large similarities between the control of AGVs, and the control of

FLTs equipped with vehicle-mounted terminals. In the design stage of the warehouse or plant, a number of questions has to be answered concerning such vehicle systems, such as: the number of vehicles needed, the exact routings of the vehicles, the way to control the vehicles and the products the vehicle should transport.

It is in general difficult to answer such questions. However, in view of savings that can be made by reducing the number of expensive vehicles and drivers, it is worth investigating the problem. If there are too many FLTs, than there are many (expensive) drivers necessary and there is also a greater probability of congestion. The latter will lead to high waiting times for loads to be picked up. On the other hand, if there are not enough FLTs, drivers will be overworked, load waiting times will be too high and due times will not be met.

In this paper we look at how the waiting times of loads (pallet response times), and the number of FLTs needed, are influenced by different control systems for the FLTs at the European distribution center of a multinational wholesaler in computer hardware and software. We do this by implementing and modeling all relevant data, such as the throughput volume, material flow, load generation times, the number of FLTs present, position of the FLT track layout, in the Automod simulation software program. Each FLT can transport one pallet and follows a specific path. Vehicles can pass each other, if necessary.

Most existing literature on Automated Guided Vehicle Systems involve studies based on operational control problems. Basically the relevant literature can be divided into two major categories: decentralized control and centralized control. Decentralized control can be subdivided into control with loops, control with multiple non-overlapping loops, and control with multiple partially overlapping loops. These control systems can be analyzed using analytical models or simulation models. In the next paragraph a selection of the relevant literature is presented and discussed to give an indication of the differences between those studies, and the study presented in this paper.

Srinivasan, Bozer and Cho (1994) present a general-purpose analytical model to compute the approximate throughput capacity of a material handling system used in a manufacturing setting. For given flow data, the model can be used to rapidly determine the throughput capacity of a wide range of handling and layout alternatives. The analytical model, as the authors say, is no substitute for simulation since it is an approximate model and simulation is still required to 'fine tune' the system.

An approximate analytical model to estimate the expected waiting times for move requests that occur in single-device trip-based handling systems is presented in a paper by Bozer, Cho and Srinivasan (1994). They assume that the empty device is dispatched according to the Modified First-Come-First-Served rule (see also Bartholdi and Platzman (1989)), which is comparable in performance to the Shortest-Travel-Time-First rule, which they also introduce in their paper. They evaluate the performance of the analytical model through simulation.

Johnson and Brandeau (1993) considered the problem of designing a multi-vehicle AGV system as an addition to an existing non-automated material handling

system. Raw components are delivered by a pool of vehicles from a central storage area to workcenters throughout the factory floor. The pool of vehicles is modeled as an *M/G/c* queuing system and the design model is formulated as a binary program. They illustrate their model with an example of an actual design problem, and present computational experience for other example design problems.

Traditionally, AGV systems have been implemented and analyzed assuming that every vehicle is allowed to visit any pick up/deliver location in the system. Bozer and Srinivasan (1991) introduce a conceptually simple and intuitive approach where the system is decomposed into non-overlapping, single-vehicle loops operating in tandem. They also develop an analytical model to study the throughput performance of a single vehicle loop. The model can also be used to measure the impact of using a bi-directional vehicle, reconfiguring the guidepath, adding new stations and changing the flow values.

An evaluation of tandem configurations is done by Bischak and Stevens (1991). They compare the performance of a tandem AGV system with that of conventional AGV track systems. In the tandem system the track is divided into non-overlapping, single-vehicle closed loops. Using simulation they show that, because of trips requiring delivery across loops, the tandem system has a higher expected travel time per load and thus a greater average time in the system than with the conventional control system. Measurements of the performance of an AGV system include AGV utilization, throughput, and a load's average time in the system.

In the previously mentioned papers, the AGVs are not controlled centrally by a central information system. In knowledge-based centralized control systems, a Warehouse Management System (WMS) keeps track of all movements within the warehouse. Such a WMS, uses a database with information of where loads are to be picked up and/or delivered, and selects an AGV to do so, according to specified logistic rules. Kodali (1997) describes a prototype knowledge-based system for selecting an AGV and a workcenter from a set of workcenters, simultaneously requesting the service for transport of a part in on-line scheduling of flexible manufacturing systems.

In this paper we extend the research based on evaluations between conventional control, however this time with overlapping loops, and control using a centralized control system.

The paper is organized as follows. Section 2 gives a brief outline of different vehicle control systems. Section 3 describes the problem at the wholesaler. In section 4 the model is described in detail. In section 5 an introduction is given of different control systems for FLTs which are analyzed. Section 6 gives an overview of the results obtained with the various AGV control systems, and in section 7 conclusions are drawn.

2 Control of Internal Transport

The transportation control system of vehicles is either *centralized* or *decentralized.* A centralized control system implies that all transportation tasks are considered simultaneously by a central computer, or the Warehouse Management System

(WMS). A control system where vehicles drive in loops and perform the first transportation task they encounter, (First-Encountered First-Served rule), is a typical decentralized way of control, (see Mantel and Landeweerd (1997)).

With centralized control, vehicles are not restricted to drive in loops, but are free to move anywhere. Furthermore, with the use of a computer, it is possible for the load to claim a vehicle, instead of the vehicle having the initiative of claiming a load. Thus there are two system types: load driven, and vehicle driven. When a load is offered to the system, it can claim a vehicle for transport on basis of several logic rules, such as: claim the closest vehicle, the vehicle with shortest travel time, the first vehicle which becomes idle or the vehicle which has been idle the longest. When the vehicle takes the initiative it can claim loads by similar rules, such as: the closest load (in travel time or distance), the load with longest waiting time, the location with the smallest queue size, or the load in the longest queue. In addition, it is possible to give priorities to certain locations where loads are to be picked up. This can be done by entering work lists in the central computer. This will be illustrated with an example. Suppose we are investigating a warehouse with 3 locations (or groups of locations). When a vehicle becomes idle after dropping off a load at location 1, the central computer (WMS) will search the work list of location 1 (see figure 1). First location 1 is checked for work by the central computer, because that location is on top of the work list. If a load is waiting there to be picked up, then the WMS instructs a vehicle to retrieve the load. If there is no work at location 1, then location 2 is checked, etc.. When a vehicle becomes idle at location 2, location 2 is checked first for work (see figure 1), then location 3, and lastly location 1. At location 3, only location 3 is checked with more priority next *all* other locations in a random or a particular (for example FIFO) order.

Location 1	Location 2	Location 3
Location 1	Location 2	Location 3
Location 2	Location 3	ALL
Location 3	Location 1	

Figure 1 Example of work lists for centralized control

The following section gives a detailed description of the case study, followed by the description of the model of the case, and the vehicle control systems used.

3 Case Description

The case concerns the transportation of pallet loads at the European distribution center of a computer hardware and software wholesaler. This wholesaler distributes computer products to different retail stores in Europe and anticipates how much to purchase and store to be able to comply to the demand of the retailers. Because computer products change quickly over time, it is necessary to keep inventory levels low and the storage times as short as possible. A large part

of the incoming products are packed in cartons. These cartons are stacked per product on pallets. The pallets are transported by FLTs. A computer, the Warehouse Management System (WMS) keeps track of inventory, and the position of stored products. The FLTs are manned and can move a single pallet at a time.

The distribution center (DC) can be divided into several areas (see figure 2). Each day, trucks arrive at the *Receiving Lanes* of the DC where the pallets (loads) are unloaded. In total there are 5 Receiving Lanes. If the cartons on the pallets contain returned or broken products they are manually transported to one of the 5 *Return stations*. If it is unclear what the content is of the cartons, the pallets are manually transported to the *Check-in Area*. There are 12 Check-in stations in total (see figure 2). At each of the previously mentioned stations, the pallets are labeled with a so called license plate number (also bar code). This license plate number contains information about the content of the cartons and the location the pallet should be brought to. At the moment the license plate is placed on the pallet, the pallet is entered into the WMS.

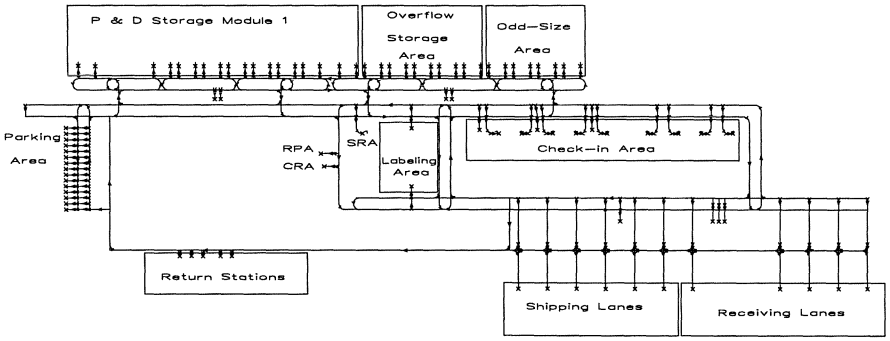


Figure 2 FLT path layout connecting all pick up and delivery locations, all main transport tracks are uni-directional

If the cartons on the pallet are odd-shaped, or if the pallet is one of many with the same product, it will be transported to the *Odd-Size* or *Overflow Storage Area* respectively. The Odd-Size Storage Area and the Overflow Storage Area have 10 and 8 stations respectively. Otherwise the pallets go to one of the 18 Pick & Drop (P&D) locations of *P&D Storage Module 1*. Within the storage modules, pallets are stored and orders are picked, see De Koster *et al.* (1997) for a detailed discussion on the routing of orderpickers. From Storage Module 1 pallets can be transported to the Repalletization Area (*RPA*), the Shelf Replenishment Area (*SRA*), the Central Return Area (*CRA*), the *Shipping Lanes* or to the *Labeling Area* (see the material flow diagram in figure 3). The Labeling Area has one delivery station and one pick up station. RPA, CRA and SRA have one station each, and there are 6 shipping lanes in total (see figure 2).

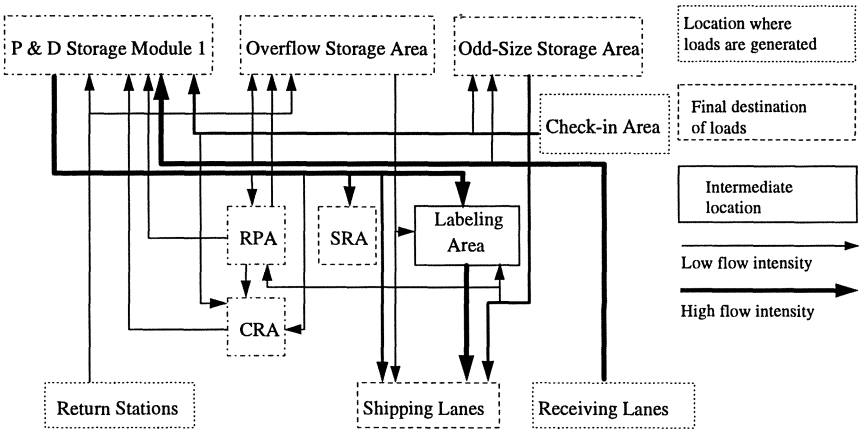


Figure 3 Material flow between all locations

From RPA, pallets move to Storage Module 1 or to CRA. At SRA the cartons of the pallets are placed on a conveyor belt, and will be transported to the shelf area where products are hand picked.

Pallets at CRA always move to Storage Module 1. At the Labeling Area, pallets receive customer stickers and packing lists.

The final stations are at the Shipping Lanes. At the end of the Shipping Lanes trucks arrive at dock doors to transport products to retail stores.

The aim is to find a control system for the FLTs such that loads are transported on the FLT track layout (see figure 2) to and from the correct location with a given flow intensity (see figure 3), while keeping the number of FLTs (and drivers) as low as possible and keeping load waiting times as short as possible. Short load waiting times, or response times are important to realize due times for the trucks waiting at the Shipping Lanes also to and keep queues at stations as short as possible. To find the control system capable of this task, the case has been implemented in the model described in the next section.

4 The Model

To calculate the performance of each control system, the design of the warehouse (figure 2) and other relevant specifications of the warehouse and FLTs have been modeled in the Automod simulation software package.

All the parameters are kept the same for each control scenario. These parameters include: the material flow (see figure 3, and table 2), the number and locations of loads generated in the system, load generation instants, the speed of the vehicles, vehicle capacity, the paths via which the vehicles may travel (see figure 2), the load pick up and set down time, the number of simulated days and the number of working hours per day. The length of the simulation is 10 days

(which means about 5.800 loads to be transported), where the first 2 days are used as a warm up period. Table 1 below gives a summary of the values of the data.

Table 1 The parameters used for each scenario

FLT speed	2 m/s
Pick up time of a load	15 s
Set down time of a load	15 s
Vehicle capacity	1 load (pallet)
Simulation period	10 days
Number of working hours per day	7.5 hours

Vehicles always use the path with the shortest travel time when traveling to a pick up or delivery location. In table 2 the throughput per day from one location to another is given.

Table 2 Total throughput in pallets per day

From / To		1	2	3	4	5	6	7	8	9	10	11	**Total**
1	Labeling Area	0	0	159	0	0	0	0	0	0	0	0	159
2	Check-in Area	0	0	0	0	0	0	22	0	0	0	0	22
3	Shipping Lanes	0	0	0	0	0	0	0	0	0	0	0	0
4	Receiving Lanes	0	0	0	0	0	0	109	2	2	0	0	113
5	SRA	0	0	0	0	0	0	0	0	0	0	0	0
6	RPA	0	0	0	0	0	0	9	0	0	0	1	10
7	P&D Storage Module 1	144	0	31	0	17	5	2	0	0	0	0	199
8	Overflow Storage Area	4	0	12	0	0	0	0	0	0	0	0	16
9	Odd-Size Area	11	0	40	0	0	1	0	0	0	0	0	52
10	Return Stations	0	0	0	0	0	0	6	0	0	0	0	6
11	CRA	0	0	0	0	0	0	4	0	0	0	0	4
Total		159	0	242	0	17	6	152	2	2	0	1	**581**

The loads are independently exponentially generated for a total of 10 days. Thus the interarrival times of loads follow a Poisson process. Each day is in turn divided into 4 periods. Period 1: from the start of the day until the coffee break, period 2: from the coffee break until lunch, period 3: from lunch until the tea break, and period 4: from the tea break until the end of the working day. These periods are introduced in order to realistically represent the variation in the interarrival rates over the day. For example, in period 4 there are more loads transported to the shipping lanes than in period 1.

Loads are generated at the Return Stations, RPA (where pallets are split or restacked), CRA (where damaged pallets or pallets with unknown contents are checked), Storage Areas, Check-in Area and Receiving Lanes (see figure 3). All other locations are end locations or intermediate locations.

The loads are transported by FLTs which can be controlled in many ways (see section 2). The following section describes the vehicle control systems used in this case study.

5 Different Ways to Control FLTs

In this paper we look at 4 different vehicle control systems by which the FLTs are controlled. The first is conventional control where the FLTs drive in fixed loops. The last three are control systems where the FLTs have vehicle-mounted terminals which are linked to a central computer (WMS). The central computer keeps track of the pallet movements and instructs the FLTs accordingly. The loop configuration discussed below has partially overlapping loops, and will be compared with a configuration where vehicles are free to go anywhere.

The following paragraphs describe the control systems individually in more detail.

5.1 Conventional control system with loops

This is the current situation of the warehouse of the computer hardware and software wholesaler. All the pick-up and delivery locations in the warehouse are divided in 2 main uni-directional loops. Loop 1 contains the Return Stations, RPA, CRA and Labeling Area (see figure 2). Loop 2 consists of all stations except the Return Stations. Thus there are two partially overlapping loops for the FLTs. Each FLT is assigned to a fixed loop, with 1 FLT in loop 1, and 6 FLTs in loop 2. The driver is responsible for selecting the proper load and transporting it. The FLTs of one loop are not allowed to pick up pallets in the other, pallets can only be *delivered* at locations of the same or the other loop. An FLT which delivers a pallet in the other loop immediately returns to the nearest point of its own loop. The vehicles are in this case always in motion, driving in their own loop checking for work at the stations they pass. If there is no work at a particular station, the FLT travels to the next station in the loop. If there is work at that location, than the FLT picks it up and brings it to its destination (which could be in the other loop).

Although this is a commonly used way to control FLTs (without truck terminals) in a warehouse, it is inefficient. If more stations are added, or the FLT path layout is changed, then the design of the loops changes also. This means that a new vehicle assignment has to be made to balance the performance in the new loops. To make more efficient use of the FLTs, we introduce a vehicle control system which uses a centralized computer.

5.2 Control system with work lists

In this control system, the FLTs are equipped with vehicle-mounted terminals that are in constant communication with a central computer. The central computer, the Warehouse Management System (WMS), keeps track of the pallets and the so called work lists (see for example figure 4). Each delivery or drop-off location has a work list. If an entry on the work list contains multiple locations, then these are selected in a first-come first-served order. If there are no more locations to check on the list, and still no work has been found, the FLT is instructed to park at the nearest parking place, and waits until it is called for again.

In this scenario there are many work lists, a unique one for every drop-off location. For example, at the labeling area, the first search location on the work list is *labeling area* then *P&D module 1* then *return* etc., at the end of the list *all* remaining stations are checked for possible work (see figure 4).

Labeling Area	Shipping Lanes	SRA	RPA
Labeling Area	Odd-Size Area	RPA	RPA
P&D Module 1	Overflow Area	CRA	CRA
Return Stations	Check-in stations	P&D Module 1	Return Stations
Receiving Lanes	Receiving Lanes	Return Stations	Receiving Lanes
ALL	P&D Module 1	Receiving Lanes	P&D Module 1
	Labeling Area	Labeling Area	Labeling Area
	ALL	ALL	ALL

P&D Storage Module 1	Overflow Area	Odd-Size Area	CRA
P&D Module 1	Overflow Area	Odd-Size Area	CRA
RPA	Odd-Size Area	Overflow Area	RPA
CRA	P&D Module 1	P&D Module 1	P&D Module 1
Return Stations	RPA	RPA	Return Stations
Overflow Area	CRA	CRA	Receiving Lanes
Receiving Lanes	Return Stations	Return Stations	Labeling Area
Labeling Area	ALL	ALL	ALL
ALL			

Figure 4 Work lists for all delivery locations for the control system with work lists

The work lists are constructed in such a way, that in most cases, the locations around the current position of the idle FLT are checked first for work. Furthermore, the route the idle FLT should follow next is consistent (in most cases) with the uni-directional flow of the paths. This reduces the probability of circulating around without a load, to pick up a load which has been made available just 'behind' the current location of the idle FLT. These are also the lists the company has investigated and are going to implement.

Although these work lists are constructed in such a way that the WMS searches for work in neighboring locations, they might not give the best results. Because the Return Stations, RPA and CRA do not appear in every work list, or appear on the top of the work lists, it is expected that the load waiting times (or pallet response times) will be rather high for these areas. To decrease these pallet response times, the work lists are updated as described next.

5.3 Control system with updated work lists

This control system is the same as the *Control System with Work Lists* described above. The difference is that more priority is given to stations where relatively little happens. Due to the structure of the current work lists, the pallet response times may be high at the CRA, RPA and Return Stations.

Labeling Area	Shipping Lanes	SRA	RPA
RPA	RPA	RPA	RPA
CRA	CRA	CRA	CRA
Return Stations	Return Stations	Return Stations	Return Stations
Labeling Area	Odd-Size Area	P&D Module 1	Receiving Lanes
P&D Module 1	Overflow Area	Receiving Lanes	P&D Module 1
Receiving Lanes	Check-in stations	Labeling Area	Labeling Area
ALL	Receiving Lanes	ALL	ALL
	P&D Module 1		
	Labeling Area		
	ALL		

P&D Storage Module 1	Overflow Storage Area	Odd-Size Area	CRA
RPA	RPA	RPA	RPA
CRA	CRA	CRA	CRA
Return Stations	Return Stations	Return Stations	Return Stations
P&D Module 1	Overflow Area	Odd-Size Area	P&D Module 1
Overflow Area	Odd-Size Area	Overflow Area	Receiving Lanes
Receiving Lanes	P&D Module 1	P&D Module 1	Labeling Area
Labeling Area	ALL	ALL	ALL
ALL			

Figure 5 The work lists for all delivery locations for the control system with updated work lists

The idea is, that relatively little happens there (see table 2), so the priority is low. The result is that busy areas are always checked first for work. Because many pallets need to be moved there, the FLTs are instructed to leave immediately, and instructions to go to the Return Areas are therefore rare. This will result in high pallet response times at the Return Areas, and so in this scenario (see figure 5), they are placed on top of every work list (i.e. they have the most priority).

To update and maintain all the work lists is time consuming in practice. Especially if stations are added and the number of work lists increases. To ease the workload of maintaining all the work lists, and to keep the advantage of a centralized computer, a new control system is introduced with only one work list. This control system with a single work list is described in the next paragraph.

5.4 Control system with a single flow intensity based work list

The difference between this control system and the previously described control system is, that there is only one work list altogether. However, use is still made of a centralized computer for the vehicle control system. Every drop-off location has the same work list (see figure 6). The location with the least daily outgoing flow intensity is placed on top of the work list. The locations are added to the list, in ascending order of flow intensity. The idea is that locations with little work, will now not be neglected so soon. Because relatively little happens at these locations,

they are often skipped quickly and the central computer ends up checking the busy areas anyway. This is a very simple control system, since there is only one list, which is based only on the flow intensity per location or area. It is also rather easy to implement and easily maintained.

All Stations
CRA
Return Stations
RPA
Overflow Area
Check-in Stations
Odd-Size Area
Receiving Lanes
Labeling Area
P&D Module 1
ALL

Figure 6 The work list for all delivery locations for the control system with a single flow intensity based work list

6 Results

The 4 different scenarios have been evaluated in the warehouse, with the same FLT track layout (see figure 2), material flow generation, FLT speed, etc. as described in section 4 (see table 1). The only difference is the logic of the vehicle control systems, described in section 5. In this section we present and evaluate the results for each of the 4 scenarios. The following performance criteria are compared:

- number of vehicles needed
- pallet response times
- number of pallets waiting for transport
- vehicle idle time/percentage of utilization

In the case of conventional control with loops, the FLTs are continuously driving around, even if they have no load. In other words, they never park and are therefore never idle. So the vehicle idle time is a performance measure which is only useful for the last 3 scenarios where the WMS can instruct the FLTs to park, after the work lists have been searched and no work has been found.

The first 2 days of the 10 day simulation period are used as a warm up period. So the results mentioned in the next paragraphs are averages for the last 8 days of the simulation run.

6.1 Results conventional control system with loops

This is the current situation of the hardware and software wholesaler. Experience has shown that a total of 7 FLTs are necessary to transport all loads. Loop 1 with

CRA, RPA, labeling and the return stations (see figure 2) has 1 FLT. The other loop contains all stations except the Return Stations and has 6 FLTs. This division has been verified with the simulation model. Figure 7 gives an overview of the number of pallets which are waiting to be moved for the last 8 days (i.e. hours 15 until 75). This statistic is updated every minute, the figure also includes a line representing the moving average for the number of pallets waiting to be moved.

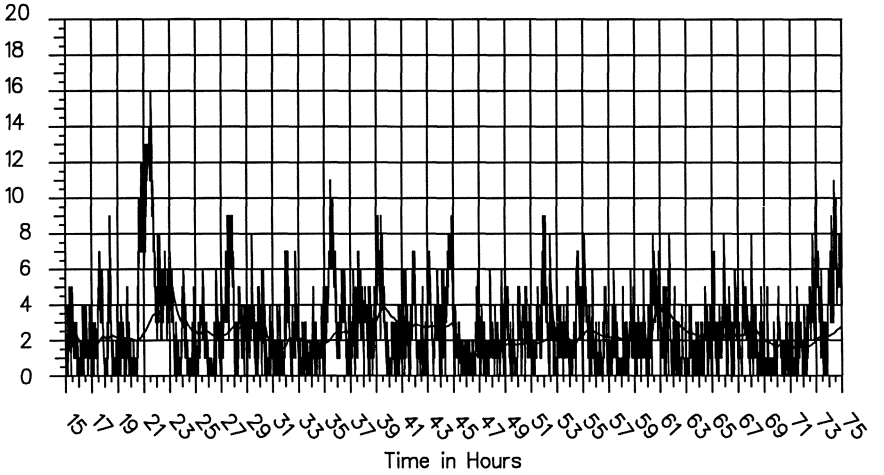


Figure 7 The total number of waiting pallets as a function of time for the case of conventional control

The wave in the moving average per day, can be explained by the different intensities in which loads are generated over the day.

The waiting times of the pallets (figure 7) are presented in table 3. The pallet response times for the Return Stations, RPA and CRA are grouped together as 'Return Areas'. The row named 'Outbound' summarizes the response times for all loads with the Shipping Lanes as final destination. The 'Inbound' row summarizes the response times for all other loads. The last row ('Total') gives an overview of the response times of all waiting pallets.

The FLT in loop 1, mainly responsible for the Return Areas, is able to handle the pallet movements satisfactorily when compared to the other stations. When only 5 instead of 6 FLTs are used in the second loop, the total average response time increases with more than 100 seconds and the standard deviation more than doubles.

Table 3 Pallet response time per station in seconds

Location	Minimum	Mean	Maximum	Std. Dev.
Labeling Area	27	119	464	82
Return Areas	16	134	613	116
P & D Module 1	25	221	4513	339
Receiving Lanes	33	119	875	83
Check-in Stations	28	172	1335	177

Inbound	16	140	1719	138
Outbound	25	183	4513	282
Total	16	170	4513	248

Although an average response time of less than 3 minutes (170 seconds) is not bad, substantial improvements can be obtained by using a Warehouse Management System. This is shown in the next paragraph.

6.2 Results control system with work lists

In this case, only 5 FLTs were necessary in total. The system with a WMS using work lists, therefore out-performs the conventional system with loops. Figure 8 gives an overview of the number of pallets which are waiting to be moved. Close observation points out that the moving average of the number of pallets waiting to be picked up is in pattern almost the same as that of figure 7.

The moving average generally stays in a bandwidth of 1 till 3 pallets, whereas in figure 7 this was between 2 till 4. The minute to minute peaks show a different pattern, with a maximum of 18 pallets waiting to be picked up at a certain point in time.

The response times of the pallets (figure 8) are presented in table 4. As can be seen in the last row, the total average waiting time of the pallets is comparable to (even less then) that of the conventional control system (see table 3).

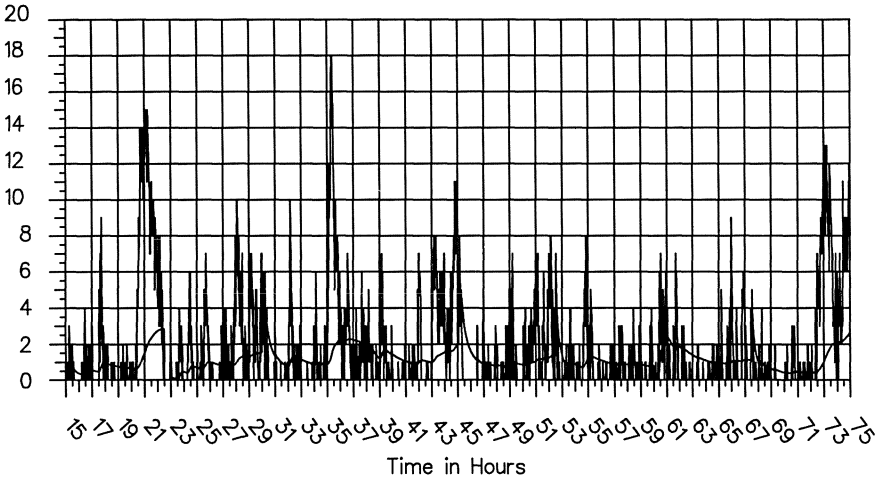


Figure 8 The total number of waiting pallets as a function of time for the case of control with work lists

It should be noted however, that with conventional control, 7 FLTs were necessary instead of the 5 used now. This is an improvement of 28.6 % for the number of FLTs needed.

Table 4 Pallet response time per station in seconds

Station	Minimum	Mean	Maximum	Std. Dev.
Labeling Area	51	113	507	51
Return Areas	60	210	1699	282
P & D Module 1	33	127	1417	75
Receiving Lanes	42	328	4072	500
Check-in Stations	47	127	503	71
Inbound	37	266	4072	425
Outbound	33	122	1417	72
Total	33	165	4072	249

Figure 9 can be used to verify whether the number of logged in FLT is sufficient to handle the material flow. In this figure a minute to minute representation is made of the number of FLTs that are idle. It can be seen that, at a certain points in time, all (5) FLTs are idle and also that all FLTs are busy at some times. FLTs spend about 20.3 % of their logged in time in the idle state.

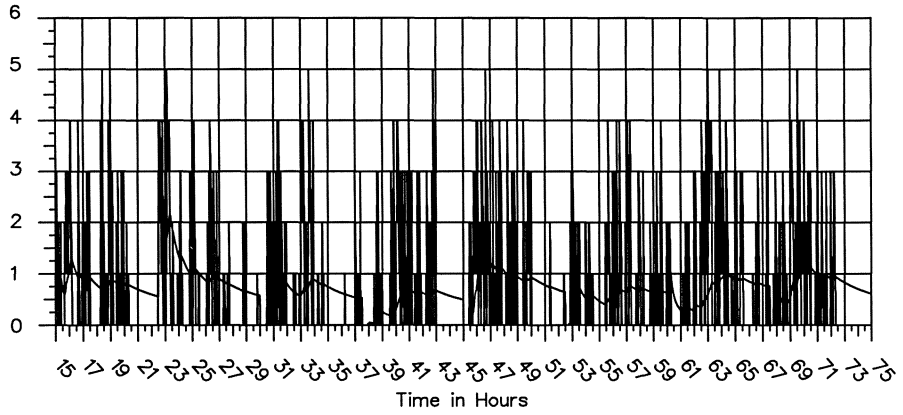


Figure 9 The total number of idle FLTs as a function of time for the case of control with work lists

6.3 Results control system with updated work lists

Although the total average pallet response time for the WMS control system is less than the response time of the conventional system, the average pallet response time at the Return Areas has increased. This was expected, as mentioned earlier, because the positions of the Return Areas are not very high on the work lists, if they are on the lists at all (see figure 4). The result is that other locations are given more priority and the locations in the Return Areas are often neglected.

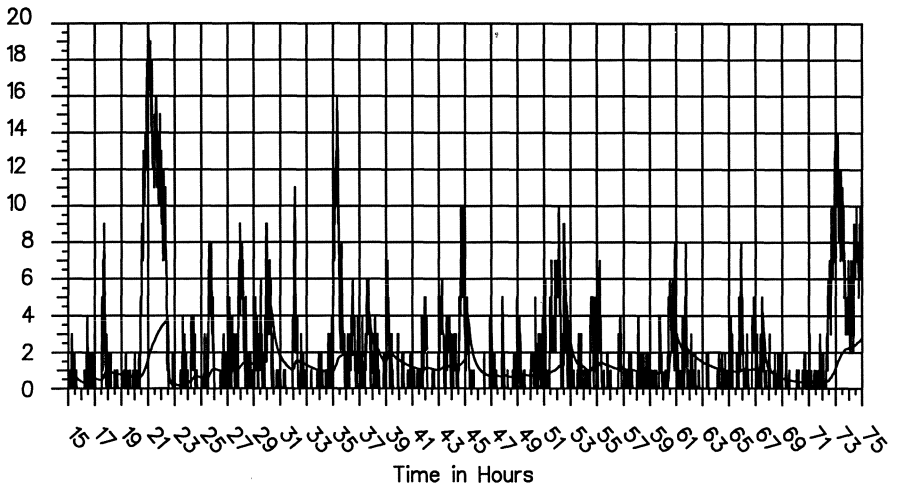


Figure 10 The total number of waiting pallets as a function of time for the case of control with updated work lists

The work lists have therefore been updated with the Return Areas always on top (see figure 5). In total there are still only 5 FLTs needed. Figure 10 gives an overview of the number of pallets which are waiting to be moved for the last 8 days. The moving average of the number of pallets waiting to be picked up is in pattern almost the same as that in figure 8. The moving average also stays in a bandwidth of 1 to 3 pallets. The peaks are higher, with a maximum of 20 pallets waiting to be picked up at a certain point in time.

The waiting times of the pallets are presented in table 5. As can be seen in the last row, the average waiting time of all pallets is comparable with the previous control system (see table 4). As expected however, the response time at the Return Areas has decreased considerably, from 210 seconds to 102 seconds. The vehicle utilization in this case is 80 %, as well.

Table 5 Pallet response time per station in seconds

Station	Minimum	Mean	Maximum	Std. Dev.
Labeling Area	48	115	457	65
Return Areas	55	102	215	34
P & D Module 1	33	128	1690	81
Receiving Lanes	42	340	5558	556
Check-in Stations	50	117	322	55
Inbound	37	265	5558	467
Outbound	33	123	1690	79
Total	33	166	5558	271

The result of giving the Return Areas more priority, is a slight decrease in performance for the pallet response time at the Receiving Lanes. However, this is only about 12 seconds. The other response times generally stayed the same. It may

therefore be said that this control system, is not better, but not worse either, than the previous control system.

6.4 Results control system with a single flow intensity based work list

This control system makes use of only one work list (see figure 6), to which the WMS refers every time a FLT becomes idle (i.e. after delivering a load).

In total there are still only 5 FLTs necessary. Figure 11 gives an overview of the number of pallets which are waiting to be moved. The peaks are the lowest of all control systems with a maximum of 12 pallets waiting to be picked up at a certain point in time.

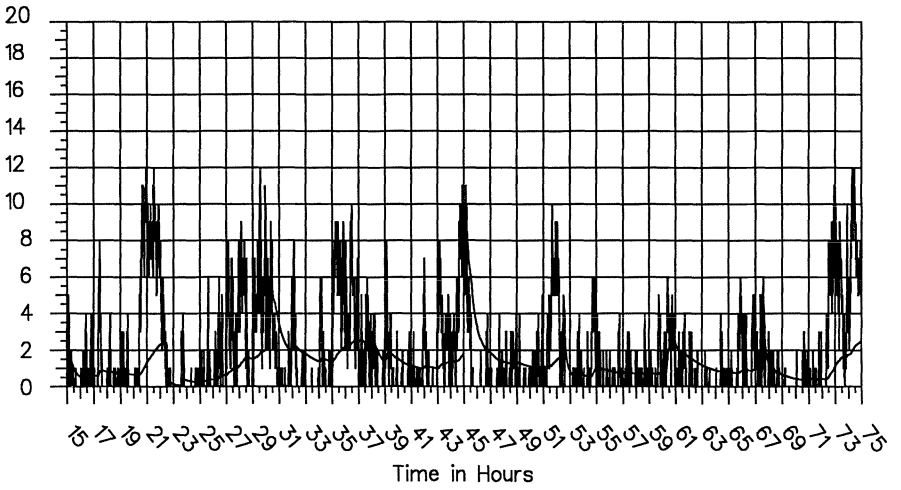


Figure 11 A minute to minute representation of the number of pallets waiting with modern control and a single flow intensity based work list

The waiting times of the pallets are presented in table 6. As can be seen in the last row, the average waiting time of all the pallets is, except for the response time at the P&D module 1, comparable to, or less than, that of the other control systems (see tables 3, 4 and 5). However, the response time at P&D Module 1 has increased with almost half a minute. This is no surprise because these stations are given the least priority. This is due to the fact that most of the material flow takes place here, and so they are at the bottom of the flow intensity based work list. This results in a slight increase of the total average response time of about 10 seconds. The average vehicle utilization in this case is still only 80.5 %.

Table 6 Pallet response time per station in seconds

Station	Minimum	Mean	Maximum	Std. Dev.
Labeling Area	51	106	746	56

Return Areas	52	107	213	33
P & D Module 1	32	256	5996	441
Receiving Lanes	42	122	363	49
Check-in Stations	45	101	232	31
Inbound	38	133	2916	142
Outbound	32	201	5996	363
Total	32	180	5996	315

Due to the simplicity of this control rule, and the reasonable results, it may be said that this control system scores well. Especially because still only 5 FLTs are needed.

7 Conclusion

In this paper we looked at 4 different vehicle control systems, one decentralized conventional system with two partially overlapping loops, and three centralized control systems. The object was to see whether centralized knowledge-based systems can improve warehouse performance. The performance can be measured in several dimensions: the number of FLTs needed, pallet response time, vehicle utilization, model simplicity and robustness. The idea is that conventional control is too expensive because it makes inefficient use of possible information.

The results are summarized in table 7. They show that less FLTs are needed in order to get generally the same (even slightly lower) response times when a centralized computer is used. In total, the conventional decentralized control system needs 40 % more FLTs.

When conventional control is used, the FLTs are constantly in motion, and are therefore more liable to break down. Also, the constant driving without work to be done can cause driver irritation. So, in view of this, the number of vehicles needed and the higher pallet response time, it is clear that in this case control with work lists outperforms the conventional loop control systems.

Table 7 Summarized results of centralized control versus decentralized control

Control system	Number of FLTs needed	Total average response time	FLT utilization	Max. pallets waiting
Decentralized: Conventional loops	7	170 s	100 %	16
Centralized: Work lists	5	165 s	79.7 %	18
Centralized: Updated work lists	5	166 s	80 %	20
Centralized: Flow intensity based work list	5	180 s	80.5 %	12

The performance of centralized control systems can be improved as well. That is, the response time at some locations can be reduced. However, this will usually

increase the response time at other locations. This was done in the case of the control system with updated work lists. Pallet response times at the Return Areas decreased with almost 50% while other statistics basically remained the same.

Even though the centralized systems win in the number of vehicles needed and lower pallet response times, they lose in simplicity from the conventional systems. Implementing, maintaining and fine tuning work lists is not an easy task. To make this easier, a centralized system with only one 'flow intensity based' work list was introduced. This is a fairly simple control system and easy to construct. Results show (see table 7), that the flow intensity based control system performs comparable to the other centralized system. The average total pallet response time increased only 10 seconds.

We believe that decentralized conventional control systems with loops will in general be outperformed by centralized control systems with work lists, because the latter makes better use of available information. Furthermore, when using a central computer with work lists to control the vehicles, the performance is almost insensitive to the structure of the lists.

References

Bartholdi III, J.J. / Platzman, L.K. (1989): Decentralized Control of Automated Guided Vehicles on a Simple Loop, in: IIE Transactions, Vol. 21, No. 1, 76-81

Bischak, D.P. / Stevens, K.B., Jr (1995): An evaluation of the tandem configuration automated guided vehicle system, in: Production Planning & Control, Vol. 6, No. 5, 438-444

Bozer, Y A. / Cho, M. / Srinivasan, M M. (1994): Expected waiting times in single-device trip-based material handling systems, in: European Journal of Operational Research, Vol. 75, 200-216

Bozer, Y.A. / Srinivasan, M.M. (1991): Tandem configurations for automated guided vehicle systems and the analysis of single vehicle loops, in: IIE Transactions, Vol. 23, No. 1, 72-82

De Koster, R. / Van der Poort, E. / Roodbergen, K.J. (1997): When to apply optimal or heuristic routing of orderpickers, in: Advances in Distribution Logistics, (Springer) Berlin

Johnson, M.E. / Brandeau, M.L. (1993): An analytical model for design of a multivehicle automated guided vehicle system, in: Management Science, Vol. 39, No. 12, 1477-1489

Kodali, R. (1997): Knowledge-based systems for selection of an AGV and a workcentre for transport of a part in on-line scheduling of FMS, in: Production Planning & Control, Vol. 8, No. 2, 114-122

Mantel, R.J. / Landeweerd, R.A. (1995): Design and operational control of an AGV system, in: Graves, R.J., McGinnis, L.F., Medeiros, D.J., Ward, R.E., and Wilhelm, M.R. (eds.), *Progress in Material Handling Research*: (1994), pp.309-323, ISSN: 1080-711X.

Srinivasan, M.M. / Bozer, Y.A. / Cho, M. (1994): Trip-based material handling systems: throughput capacity analysis, in: IIE Transactions, Vol. 26, No. 1, 70-89

Chapter 5

Inventory Control and Forecasting

Transshipments in a divergent 2-echelon system

E.B. Diks[1] and A.G. de Kok[1,2]

[1] Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[2] Department of Technology Management, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Summary. Consider a two-echelon inventory system consisting of a central depot (CD) supplying a number of retailers. Only the retailers face customer demand. Both the CD and every retailer adopt a periodic review echelon order-up-to policy, i.e., periodically an order is placed to raise the echelon inventory position to the order-up-to-level. At the CD incoming stock is allocated by using the *Balanced Stock* (BS) rationing policy. When the orders arrive at the retailers, an instantaneous rebalancing of the total echelon stock of the retailers takes place. This rebalancing is realized by lateral transshipments, assuming zero transshipment lead times. The used rebalancing policy is also a BS rationing policy.

The objective of this paper is to determine all the control parameters, such that the target customer service levels are attained at minimal expected total costs. Exact expressions are developed to determine these parameters. The order-up-to-levels are computed by a heuristic, since the exact expressions are intractable. A numerical study reveals that the performance of this heuristic is good. Furthermore, this model is compared with the same model without lateral transshipments, which yields insight into the conditions under which transshipment could be useful.

1. Introduction

In this paper we consider a single-item distribution system consisting of a central depot (CD) supplying a number of retailers. This system is also referred to as a divergent two-echelon system. Periodically the CD places replenishment orders at an external supplier, which has sufficient capacity to guarantee a fixed lead time. Both the CD and every retailer control their stock by a so-called *echelon stock* policy. This policy was introduced by Clark & Scarf (1960). The echelon stock of a stockpoint equals all stock at this stockpoint plus in transit to or on hand at any downstream stockpoints minus the backorders at its downstream stockpoints. The echelon inventory position of a stockpoint is defined as the sum of its echelon stock and the material in the pipeline towards this stockpoint. The inventory in the system is controlled by a periodic review mechanism. That is, every review period the CD issues a replenishment order that raises the echelon inventory position to its order-up-to-level. Upon arrival it is decided how to allocate this order to the CD and the retailers. For the purpose of allocating stock to retailers we need a rationing policy.

A lot of research has been done to determine the control parameters (e.g. the order-up-to-level and the allocation policy of the CD) such

that the expected total costs (holding plus penalty costs) are minimized. Diks and de Kok (1998b) proved that an echelon stock policy is optimal given the balance assumption, which assumes that the allocation policy allocates nonnegative quantities. Under some additional assumptions Diks and De Kok (1998a) derive an approximately optimal control policy. For a survey on cost-optimal policies we refer to Federgruen (1993) and Van Houtum et al. (1996).

A major disadvantage of the approach of minimizing the expected total costs is that in practice penalty costs are often unknown. Usually operational surrogates (e.g. customer service levels) are required to determine 'reasonable' penalty costs. Hence the cost-optimal approach yields the policy which minimizes the holding costs *plus* the penalty costs. While in our opinion the objective of the analysis should be to determine the policy which minimizes the expected holding costs given some customer service level constraints (see also Tüshaus and Wahl (1997)). De Kok (1990) determines a control policy which satisfies customer service levels in a divergent two-echelon system with a stockless depot (i.e. the CD serves merely as a coordinator). In De Kok et al. (1994) the model is extended by allowing the CD to hold stock. Furthermore, they introduce the Consistent Appropriate Share (CAS) rationing policy which is a generalization of the well-known Fair Share (FS) rationing policy of Eppen & Schrage (1981). Van der Heijden (1997) introduces the related Balanced Stock (BS) rationing policy. For an overview on these rationing policies we refer to the survey paper of Diks et al. (1996).

A possible way to decrease the stock needed to operate the system, but still guarantee the target customer service levels, is by allowing transshipments between the retailers. In the situation where often some end-stockpoints have excess inventory while others faces shortages, lateral transshipments has gained in popularity as the appropriate recourse action to avoid shortages. However, by allowing these lateral transshipments extra transportation (and handling) costs are involved. So the appropriateness of using lateral transshipments depends on the trade-off between the extra costs involved with the transshipments and the decrease in holding costs.

In this paper we allow lateral transshipments between retailers. These transshipments take place upon arrival of the replenishment orders of the retailers. Like Karmarkar and Patel (1977), Hoadley and Heyman (1977), Cohen et al. (1986), Tagaras (1989), Tagaras and Cohen (1992) and Diks and De Kok (1996) we assume that these transshipments are instantaneous, i.e., all transshipment lead times have a zero duration. In practice this may correspond to shipping stock overnight. In this way the inventory system experiences zero transshipment lead times, while in reality it takes some time to ship the stock. Furthermore, the model with non-zero lead times are extremely hard to deal with analytically.

Diks and De Kok (1996) analyzed the divergent two-echelon system with lateral transshipments. They adopted the CAS rationing policy both to allocate the stock at the CD and to rebalance the stock at the retailers. An

extensive numerical study of Van der Heijden et al. (1997) indicates that the BS rationing policy outperforms the CAS rationing policy in divergent multi-echelon systems (without lateral transshipments). Hence, in this paper we consider the same model as Diks and de Kok (1996) except we adopt the BS rationing policy at the CD and at the retailers.

A similar system with lateral transshipments has been analyzed by Jönsson and Silver (1987). They considered a divergent 2-echelon system with a stockless CD, but with positive transshipment lead times. A major difference between their model and ours is the duration of the review period compared to the depot lead time plus the retailer lead time. They considered a large review period, whereas we consider a small review period. Therefore in our model it is possible that more than one replenishment order is outstanding at the same time. As already noted by Diks and de Kok (1996), this complicates the analysis considerably.

Also Tagaras (1989) analyzes a two-echelon distribution system with lateral transshipments. In our opinion the usefulness of their model is doubtful due to the restrictive assumptions (only two retailers are considered, and the CD has an infinite capacity with a zero replenishment lead time). His model is characterized by complete pooling in that if there is an economic incentive to transship one item, then the maximum amount will be sent.

The paper is organized as follows. In Sect. 2. we describe the system under consideration. In Sect. 3. we present the allocation policy adopted at the CD and the rebalancing policy adopted at the retailers. For both policies we use the BS rationing policy. In Sect. 4. all the control parameters of the system are determined, when every review period the cumulative echelon stock of the retailers is rebalanced. In Sect. 5. we extend the results to divergent N-echelon systems. In Sect. 6. we present some numerical results of our model, and compare them with the model of Van der Heijden (1997) (which uses BS rationing and transshipments are *not* allowed) and the model of Diks and de Kok (1996) (which uses CAS rationing and transshipments are allowed). In Sect. 7. we extend the model by not rebalancing every review period, but only when the cumulative echelon stock of the retailers drops below a certain level. Finally, we give a few concluding remarks in Sect. 8..

2. Model description

Consider a divergent two-echelon system consisting of a central depot (CD) supplying a number of retailers (see Fig. 2.1). The CD can place orders at an external supplier, and retailers can place orders at the CD. Thus we do not allow for direct shipments from the external supplier to a retailer. Such a kind of model is addressed by Klein and Dekker (1997).

The inventory system is controlled by a periodic review mechanism. That is, every period the CD issues a replenishment order that raises the echelon inventory position to its order-up-to-level S_0. Like many papers (cf. Houtum

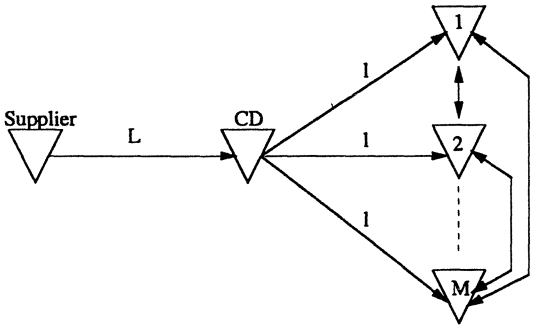


Fig. 2.1. Schematic representation of the inventory distribution system.

and Zijm (1991) and Van der Heijden (1997)) we assume that the supplier has sufficient capacity to deliver the complete order arrives after a fixed lead time L. Upon arrival it is decided how to allocate the order to the CD and the retailers. When allocating stock there are two possibilities:

1. The physical stock at the CD is sufficient to raise the echelon inventory position of each retailer to its order-up-to-level S_i. Then the required amounts are sent to the successors and excess stock is kept at the CD to be allocated at the next occasion.
2. The physical stock is not sufficient to reach the levels S_i. Then a fraction p_i of the difference is subtracted from the amount that is sent to retailer i with $\sum_n p_n = 1$.

It takes a fixed lead time of l periods to ship products from the CD to each retailer. The assumption that all retailer lead times are identical is essential for the analysis, since otherwise dependencies in time between the rationing decisions and the rebalancing decisions at the retailers becomes rather complicated. However, usually the retailer lead times can be regarded as identical, since the retailers are positioned around the CD such that the transportation time is approximately the same. Immediately after the arrival of shipments at the retailers, it is decided whether the cumulative echelon stock of all retailers should be rebalanced or not. In case it is rebalanced, some products are shipped from retailers with excess inventory to those which are low on inventory. In this paper the rebalancing policy corresponds to the Balanced Stock (BS) rationing policy of Van der Heijden (1997).

The order in which the rationing policy at the CD and the rebalancing policy at the retailers are applied influences the material flow in the system. We use the following order:

1. Arrival of shipments at CD and retailers.
2. Decide on whether to rebalance the cumulative echelon stock of all retailers or not. In case of rebalancing the BS rationing policy is used.
3. Retailers place their orders at the CD.

4. Decide on how to allocate the echelon stock of the CD by using the BS rationing policy.
5. CD places order at the external supplier.

The objective of this paper is to determine all the control parameters in the system such that every retailer attains a predetermined target customer service level. We use the fill rate as service measure, which is defined as the the fraction of demand satisfied immediately from the stock on hand. This service measure is widely used in practice (cf. Silver and Peterson (1985), De Kok (1990) and Lagodimos (1992)). Besides attaining the target fill rates we also like to minimize the holding costs and lateral transshipments costs.

3. System dynamics

In this section we explain how the material flows through the system in case all control parameters are *given*. For sake of clarity, we first explain the system dynamics when every period it is decided to rebalance. This enables us to determine the behavior of the stock levels at both the CD and the retailers. From that we derive: (1) the attained fill rates at the retailers, and (2) the expected amount transshipped per period. Next, in Sect. 4. we determine how to set the control parameters such that every retailer attains its target fill rate at minimal expected costs. In Sect. 7. the results are extended to the case in which not every period rebalancing takes place.
For the mathematical analysis we use the notation as listed below:

Stockpoint status

For all order arrivals $t \geq 0$ we define

I_t^i	Echelon inventory position of retailer i at time t just **before** rationing.
$\hat{I}_t^i$	Echelon inventory position of retailer i at time t just **after** rationing.
J_t^i	Echelon stock of retailer i at time t just **before** rebalancing.
$\hat{J}_t^i$	Echelon stock of retailer i at time t just **after** rebalancing.

Lead times and demand

L	Lead time between the external supplier and the CD.
l	Lead time between the CD and a retailer.
μ_i	Mean of the demand of retailer i per review period.
σ_i	Standard deviation of the demand at retailer i per review period.
D_{t_1,t_2}	Demand at all retailers during $(t_1, t_2]$.
$D^i_{t_1,t_2}$	Demand at retailer i during $(t_1, t_2]$.
D_t	Demand at all retailers during t periods.
D_t^i	Demand at retailer i during t periods.

Control parameters

S_0	Order-up-to-level of the CD (with respect to the echelon inventory position).
S_i	Order-up-to-level of retailer i with respect to echelon inventory position.
S_i'	Order-up-to-level of retailer i with respect to echelon stock.
p_i	Allocation fraction of retailer i.
Δ_0	Maximum physical stock at the CD, $\Delta_0 = S_0 - \sum_{n=1}^{M} S_n$.
Δ_i	Variable denoting $S_i - S_i'$.

Performance measures

T	Total expected stock transshipped between all retailers per period,
T_i	Expected stock transshipped by retailer i per period,
β_i	target fill rate of retailer i.

3.1 Rationing policy at the CD

Consider the system at the beginning of an arbitrary period, say $t - L$. At this time the CD places an order at the external supplier to raise the echelon inventory position to S_0. Since the lead time equals L, this order arrives at the beginning of period t. So the echelon stock of the CD just after the arrival of this order equals

$$S_0 - D_{t-L,t}. \tag{3.1}$$

If this amount exceeds the sum of the order-up-to-levels of the retailers, every retailer is able to raise its echelon inventory position to its order-up-to-level S_i. Thus,

$$D_{t-L,t} \leq \Delta_0 \quad \Longrightarrow \quad \hat{I}_t^i = S_i, \qquad i = 1, 2, \ldots, M. \tag{3.2}$$

However, if (3.1) is less than $\sum_{n=1}^{M} S_n$, the *complete* echelon stock of the CD is allocated to the retailers by using an appropriate rationing policy. We suggest to apply the BS rationing policy suggested by Van Donselaar (cf. Van der Heijden (1997)), since it is easy implementable and the performance is outstanding (cf. Van der Heijden et al. (1997)). This means that

$$D_{t-L,t} > \Delta_0 \quad \Longrightarrow \quad \hat{I}_t^i = S_i - p_i(D_{t-L,t} - \Delta_0), \qquad i = 1, 2, \ldots, M. \tag{3.3}$$

For the allocation fractions p_i holds $\sum_{n=1}^{M} p_n = 1$. From (3.2) and (3.3) it follows

$$\hat{I}_t^i = S_i - p_i(D_{t-L,t} - \Delta_0)^+, \qquad i = 1, 2, \ldots, M, \tag{3.4}$$

where $X^+ = \max(0, X)$ for any expression of X.

The amount of products which are sent to retailer i at time t equals

$$q_t^i = \hat{I}_t^i - I_t^i.$$

Note that $q_t^i < 0$ may occur for some i and t, which means that a *negative* amount of the physical stock at the CD is allocated to retailer i at time t. This can be interpreted as taking q_t^i products from the pipeline of retailer i, and allocating them to other retailers. In practice, however, this is usually impossible. Therefore the occurrence of $q_t^i < 0$ is referred to as imbalance. Typically, this phenomenon applies to echelon stock policies. To deal with this phenomenon most papers on echelon stock policies assume the balance assumption (or a similar kind of assumption): allocation rule (3.4) only yields nonnegative allocation quantities q_t^i. Empirical evidence shows that even rather strong violation of this assumption has only a limited affect on the quality of the analysis, unless the variation of the lead time demand is very high (cf. Van Donselaar and Wijngaard (1987)). In this model the balance assumption is even less restrictive since the echelon stock at the retailers is balanced every period.

In Van der Heijden (1997) the allocation fractions $\{p_i\}$ are determined so as to minimize an approximate expression for the imbalance. This yields

$$p_i = \frac{\sigma_i^2}{2\sum_{n=1}^{M} \sigma_n^2} + \frac{1}{2M}, \qquad i = 1, 2, \ldots, M. \tag{3.5}$$

Note that p_i is independent of the order-up-to-levels $\{S_i\}$. Despite the fact that these allocation fractions are determined to minimize the imbalance in a divergent two-echelon system *without* lateral transshipments we propose to use the allocation fractions of equation (3.5). There are several reasons for this. First, it simplifies the analysis considerably. Second, the allocation fractions are easy to compute and therefore very suited for practical purposes. Third, we expect that these allocation fractions does not play a very important role due to the rebalancing every period.

3.2 Rebalancing policy at the retailers

After the allocation at the CD at time t, it takes l periods to ship the orders from the CD to retailers. During these periods: (1) retailer i faces a customer demand of $D_{t,t+l}^i$, and (2) the cumulative echelon stock of all retailers is rebalanced at time $t+1$, $t+2$, $\ldots$, $t+l-1$. Therefore the echelon stock of retailer i at time $t+l$ (just after the arrival of the order, but before rebalancing) equals

$$J_{t+l}^i = \hat{I}_t^i - D_{t,t+l}^i + \sum_{s=t+1}^{t+l-1} \tau_s^i, \qquad i = 1, 2, \ldots, M, \tag{3.6}$$

where τ_s^i denotes the amount of products which arrives at retailer i at time s (if $\tau_s^i \geq 0$), or the amount of products which are shipped from retailer i to other retailers at time s (if $\tau_s^i < 0$). By definition,

$$\tau_t^i = \hat{J}_t^i - J_t^i. \tag{3.7}$$

Since rebalancing only reallocates the available stock we know that $\sum_{n=1}^{M} r_t^n = 0$ for every t. From (3.4), (3.6) and the aforementioned property it can be shown that the cumulative echelon stock of all retailers equals

$$\sum_{n=1}^{M} S_n - (D_{t-L,t} - \Delta_0)^+ - D_{t,t+l}. \tag{3.8}$$

Note that the rebalancing of the cumulative echelon stock can be interpreted as the allocation of the echelon stock of stockpoint 0′ (see Fig. 3.1) to the retailers. Therefore we suggest to use the BS rationing policy to reallocate

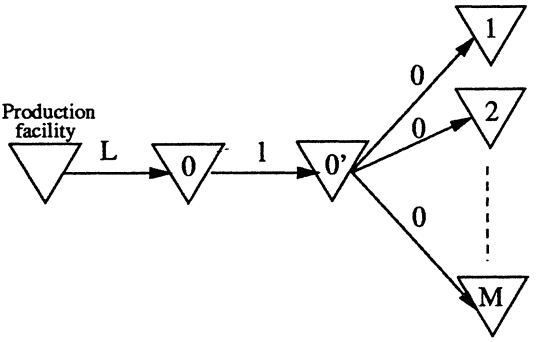


Fig. 3.1. A divergent multi-echelon system for which the rationing policy in stockpoint 0′ coincides with the rebalancing process.

these products over the retailers. For that purpose we introduce the control parameter S_i' for retailer i. This S_i' can be interpreted as the order-up-to-level of retailer i with respect to its echelon *stock*, in contrast with S_i which represents the order-up-to-level of retailer i with respect to its echelon *inventory position*. An important distinction between the BS rationing policy used as the allocation policy at the CD and the rebalancing policy at the retailers is the following. The allocation policy at the CD allows for retaining stock at the CD (therefore we distinguished between $D_{t-L,t} \leq \Delta_0$ and $D_{t-L,t} > \Delta_0$). The rebalancing policy at the retailers, however, cannot retain any stock since all the available products need to be allocated. In order to accomplish this we assume

$$\sum_{n=1}^{M} S_n = \sum_{n=1}^{M} S_n'. \tag{3.9}$$

Assumption (3.9) ensures that the cumulative echelon stock given by (3.8) is always less than $\sum_{n=1}^{M} S_n'$, which means that rationing is always required (we do not have to distinct between two cases anymore). So one might think of S_i' as a control parameter in order to ration the stock appropriately, instead of the number of products retailer i really requires to have after rebalancing.

The amount of products short to raise the echelon stock of every retailer to the level S_i' equals

$$\sum_{n=1}^{M} S_i' - \left(\sum_{n=1}^{M} S_n - (D_{t-L,t} - \Delta_0)^+ - D_{t,t+l} \right) \overset{(3.9)}{=} (D_{t-L,t} - \Delta_0)^+ + D_{t,t+l}.$$

The BS rationing policy suggests to divide this shortage over the retailers by the allocation fractions p_i given by (3.5). Hence,

$$\hat{J}_{t+l}^i = S_i' - p_i((D_{t-L,t} - \Delta_0)^+ + D_{t,t+l}), \qquad i = 1, 2, \ldots, M. \tag{3.10}$$

Using (3.10) we are able to give the service equation which has to hold for every retailer (cf. Hadley and Whitin (1963))

$$\beta_i = 1 - \frac{\mathrm{E}[D_{t+l,t+l+1}^i - \hat{J}_{t+l}^i]^+ - \mathrm{E}[-\hat{J}_{t+l}^i]^+}{\mu_i}, \qquad i = 1, 2, \ldots, M. \tag{3.11}$$

4. Determination of the control parameters

In this section we determine all the control parameters. This is done by decomposing the determination of these parameters into subproblems. First, in Sect. 4.1 we determine $\Delta_i = S_i - S_i'$ for every retailer i by minimizing the expected total stock transshipped per period. Second, assuming Δ_0 is given, we determine $\{S_i'\}$ such that the customer service level constraints are satisfied (see Sect. 4.2). Finally, in Sect. 4.3 we determine Δ_0 such that the total expected holding costs are minimized.

4.1 Determination of the control parameters $\{\Delta_i\}_{i=1}^M$

The order-up-to-levels $\{S_i\}_{i=1}^M$ *only* interact with $\{S_i'\}_{i=1}^M$ through eq. (3.9). In Sect. 4.2 we determine $\{S_i'\}$ such that the service level constraints are satisfied. Thus, we can chose $\{S_i\}_{i=1}^M$ independently of $\{S_i'\}_{i=1}^M$, provided that $\sum_n S_n = \sum_n S_n'$. These $(M-1)$ degrees of freedom are used to determine $\Delta_i = S_i - S_i'$ such that the expected total stock transshipped per period, denoted by T, is minimized. Note that this corresponds to minimizing the expected transshipment costs per period in case the costs of shipping stock from one retailer to another are equal. The problem we have to solve can be formulated as a non-linear optimization problem with M variables

$$\min T(\{\Delta_n\}) \quad \text{s.t.} \quad \sum_{n=1}^{M} \Delta_n = 0. \tag{4.1}$$

The objective function T can be computed as follows

$$T(\{\Delta_n\}) = \sum_{n=1}^{M} T_n(\Delta_n) \quad \text{with} \quad T_i(\Delta_i) = \mathrm{E}[\tau_{t+l}^i]^+. \tag{4.2}$$

In order to compute $T_i(\Delta_i)$ we need to have a tractable expression for τ_{t+l}^i. Substitution of (3.4) in (3.6), and substituting both the result and (3.10) in (3.7) yields

$$\tau_{t+l}^i = D_{t,t+l}^i - p_i D_{t,t+l} - \Delta_i - \sum_{s=t+1}^{t+l-1} \tau_s^i, \quad i = 1, 2, \ldots, M. \tag{4.3}$$

Define $\tilde{\tau}_t^i := D_{t-1,t}^i - p_i D_{t-1,t} - \tau_t^i$, then from (4.3) it follows that

$$\sum_{s=t+1}^{t+l} \tilde{\tau}_s^i = \Delta_i, \quad i = 1, 2, \ldots, M. \tag{4.4}$$

If $\tilde{\tau}_s^i$ is known for $l-1$ subsequent time periods, say $s = t+1, t+2, \ldots t+l-1$, then $\tilde{\tau}_{t+l}^i$ immediately results from (4.4). Specifically, $\tilde{\tau}_t^i = \tilde{\tau}_{t+kl}^i$ for any integer value k, $\tilde{\tau}_t^i$ is known for every t. Assuming τ_t^i is a stationair process, i.e. $\mathrm{E}[\tau_t^i] = \mathrm{E}[\tau_{t+1}^i]$ for any t, then from (4.4) and the definition of $\tilde{\tau}_t^i$ it follows that $\tilde{\tau}_t^i = \Delta_i / l$. In Appendix A. we discuss how to choose the initial state of the system such that indeed τ_t^i is a stationair process. Hence,

$$\tau_t^i = D_{t-1,t}^i - p_i D_{t-1,t} - \frac{\Delta_i}{l}, \quad i = 1, 2, \ldots, M. \tag{4.5}$$

To calculate the mean amount transshipped per period, $T_i(\Delta_i)$, we use a two-moment approximation for τ_t^i. Because τ_t^i has a probability distribution function on the entire interval $(-\infty, \infty)$ we use an approximation by a normal distribution. Hence, we can approximate $T_i(\Delta_i) = \mathrm{E}[\tau_t^i]^+$ by

$$T_i(\Delta_i) = \sigma_{\tau_t^i} \phi\left(\frac{\mu_{\tau_t^i}}{\sigma_{\tau_t^i}}\right) + \mu_{\tau_t^i} \Phi\left(\frac{\mu_{\tau_t^i}}{\sigma_{\tau_t^i}}\right), \tag{4.6}$$

where

$$\mu_{\tau_t^i} = \mu_i - p_i \sum_{n=1}^{M} \mu_n - \frac{\Delta_i}{l} \quad \text{and} \quad \sigma_{\tau_t^i} = \sqrt{(1-2p_i)\sigma_i^2 + p_i^2 \sum_{n=1}^{M} \sigma_n^2}. \tag{4.7}$$

Applying the Lagrange-multiplier technique on (4.1) yields

$$\frac{dT_i(\Delta_i)}{d\Delta_i} = \lambda, \quad i = 1, 2, \ldots, M. \tag{4.8}$$

By differentiating (4.6) to Δ_i, and substituting the result in (4.8) yields

$$-\frac{1}{l} \Phi\left(\frac{\mu_{\tau_t^i}}{\sigma_{\tau_t^i}}\right) = \lambda, \quad i = 1, 2, \ldots, M. \tag{4.9}$$

Substituting (4.7) in (4.9), and rewriting the result yields

$$\Delta_i = l(\mu_i - p_i \sum_{n=1}^{M} \mu_n - \Phi^{-1}(-\lambda l)\sigma_{\tau_t^i}), \quad i = 1, 2, \ldots, M. \tag{4.10}$$

From $\sum_{n=1}^{M} \Delta_n = 0$ it follows that $\Phi^{-1}(-\lambda l) = 0$. Hence,

$$\Delta_i = S_i - S_i' = l(\mu_i - p_i \sum_{n=1}^{M} \mu_n), \quad i = 1, 2, \ldots, M. \tag{4.11}$$

Substitution of (4.11) in (4.7) yields $\mu_{T^i} = 0$. This means that every retailer i is in balance, since the mean amount of products which are shipped *out of* retailer i equals the amount of products which are shipped *to* retailer i. This is intuitively clear, since if μ_{T^i} would be less than 0, then the CD is systematically allocating to much products to retailer i. Therefore upon arrival some of these products are immediately shipped to other retailers.

Substitution of $\mu_{T^i} = 0$ in equation (4.6) yields a simple approximation of the expected number of products which are transshipped to stockpoint i per period,

$$T_i = \sqrt{\frac{1-2p_i}{2\pi}\sigma_i^2 + \frac{p_i^2}{2\pi}\sum_{n=1}^{M}\sigma_n^2}, \quad i = 1, 2, \ldots, M. \tag{4.12}$$

4.2 Determination of the control parameters of the rebalancing policy

In this section we elaborate on how the order-up-to-level S_0 and the control parameters $\{S_i'\}$ can be determined *given* Δ_0. The control parameter S_i' is implicitly defined by service equation (3.11). This can be showed by substituting (3.10) in (3.11), which yields

$$\beta_i = 1 - \frac{\mathrm{E}[\Psi_{1i} - S_i']^+ - \mathrm{E}[p_i\Psi_2 - S_i']^+}{\mu_i}, \qquad i = 1, 2, \ldots M. \tag{4.13}$$

with $\Psi_{1i} := p_i((D_L - \Delta_0)^+ + D_l) + D_1^i$ and $\Psi_2 := (D_L - \Delta_0)^+ + D_l$.
This expression is valid for general demand distributions. To determine S_i' we suggest to solve equation (4.13) numerically. We use a similar approach as in Seidel and de Kok (1990), De Kok et al. (1994) and Van der Heijden (1997). It is convenient to use some simple two-moment approximation for the random variables Ψ_{1i} and Ψ_2. A good candidate is a mixture of two Erlang distributions, since it is closely related to the gamma distribution. Therefore all advantages using the gamma distribution for lead time demand (Burgin (1975)) remain valid, but the computations are greatly simplified.

This means that the random variable, say X, follows an E_{k_1,λ_1} with probability θ_1, and an E_{k_2,λ_2} with probability $\theta_2 := 1 - \theta_1$. Hence, the probability density function f is defined by

$$f(x) = \theta_1\lambda_1^{k_1}\frac{x^{k_1-1}}{(k_1-1)!}e^{-\lambda_1 x} + \theta_2\lambda_2^{k_2}\frac{x^{k_2-1}}{(k_2-1)!}e^{-\lambda_2 x}.$$

By using the fitting procedure of Tijms (1994) we can easily obtain the parameters λ_1, λ_2, k_1, k_2, θ_1 and θ_2 by matching the first two moments. If the

squared coefficient of variation of X, denoted by $c^2[X]$, is less than 1, then we approximate the distribution of X by the mixed Erlang distribution with parameters

$$k_1 := \lfloor 1/c^2[X] \rfloor, \qquad k_2 := k_1 + 1,$$
$$\lambda_1 := \frac{k_2 - \theta_1}{\mu_X}, \qquad \lambda_2 := \lambda_1,$$
$$\theta_1 := \frac{1}{1 + c^2[X]} \left(k_2 c^2[X] - \sqrt{k_2(1 + c^2[X]) - k_2^2 c^2[X]} \right).$$

Otherwise, we approximate the distribution of X by the so-called Coxian-2 distribution with the gamma-normalization. Then the parameters of the mixed Erlang distribution are

$$k_1 := 1, \qquad k_2 := 1,$$
$$\lambda_1 := \frac{2}{E[X]} \left(1 + \sqrt{\frac{c^2[X] - \frac{1}{2}}{c^2[X] + 1}} \right), \qquad \lambda_2 := \frac{4}{E[X]} - \lambda_1,$$
$$\theta_1 := \frac{\lambda_1(\lambda_2 E[X] - 1)}{\lambda_2 - \lambda_1}.$$

For more details regarding the fitting procedure we refer to Tijms (1994).

Because the random variables D_L, D_l and D_1^i of Ψ_1 are independent, we know that the mean and variance of Ψ_{1i} are

$$\mathrm{E}[\Psi_{1i}] = p_i \left(\mathrm{E}[D_L - \Delta_0]^+ + l \sum_{n=1}^{M} \mu_n \right) + \mu_i,$$
$$\mathrm{var}[\Psi_{1i}] = p_i^2 \left(\mathrm{var}[D_L - \Delta_0]^+ + l \sum_{n=1}^{M} \sigma_n^2 \right) + \sigma_i^2.$$

The mean and variance of $(D_L - \Delta_0)^+$ can be determined by fitting a mixed Erlang distribution on the lead time demand D_L, and next applying the formulas derived in Appendix B.. Analogously, the mean and variance of Ψ_2 can be determined. Next, we approximate the distributions of Ψ_{1i} and Ψ_2 by fitting mixed Erlang distributions on its first two moments. Again, by applying the results derived in Appendix B. we compute $\mathrm{E}[\Psi_{1i} - S_i']^+$ and $\mathrm{E}[p_i\Psi_2 - S_i']^+$. Now equation (4.13) can be solved for S_i' by using bisection.

Finally, we determine the order-up-to-level at the CD by

$$S_0 = \Delta_0 + \sum_{n=1}^{M} S_n \overset{(3.9)}{=} \Delta_0 + \sum_{n=1}^{M} S_n'. \tag{4.14}$$

Since S_i' is known for every retailer i, S_0 immediately results from (4.14).

4.3 Determination of Δ_0

In Sect. 4.1 we determined $\{\Delta_i\}$ such that the expected total amount transshipped per period is minimized. From (4.11) it follows that every Δ_i is independent of Δ_0. Next, in Sect. 4.2 we determined $\{S_i'\}$ such that every retailer

attains its target customer fill rate. From (4.13) it is clear that these control parameters depend on Δ_0. In this section we determine how to set Δ_0 so as to minimize the total expected holding costs at the *end* of every period. Following Langenhoff and Zijm (1990),Van Houtum and Zijm (1991), and Diks and De Kok (1998b) holding costs h_0 are incurred for every product at the CD or in one of the pipelines towards a retailer, and holding costs $h_0 + h_i$ are incurred for a product at retailer i. Note that h_i can be interpreted as the added value in stockpoint i (where $i = 0$ denotes the CD, and $1 \leq i \leq M$ denotes a retailer). For this cost structure the expected total holding costs at the end of a period equals

$$\begin{aligned} HC(\Delta_0) := \quad & h_0 \mathrm{E}[S_0 - D_{L+1}] \\ & + \sum_{i=1}^{M} h_i \mathrm{E}[S_i' - p_i \left((D_L - \Delta_0)^+ + D_l\right) - D_1^i] \\ & + \sum_{i=1}^{M} (h_0 + h_i) \mathrm{E}[p_i \left((D_L - \Delta_0)^+ + D_l\right) + D_1^i - S_i']^+. \end{aligned}$$

Empirically we found that for most instances the function HC is a convex function with respect to Δ_0. However, for some instances HC appeared to be concave for small values of Δ_0 and convex for larger values of Δ_0. By performing some one-dimensional search over Δ_0 we obtain the optimal value Δ_0^* which minimizes HC. Since in distribution systems usually no (or little) value is added to the product at retailer level ($h_i \approx 0$), empirically we find that Δ_0^* usually lies between 0 and $0.95L \sum_{n=1}^{M} \mu_n$. This means that no stock or very little stock is retained at the CD. This coincides with the results of other studies (cf. De Kok et al. (1994)).

5. Extension to divergent N-echelon system

In this section we illustrate how to extend the results to divergent N-echelon systems. This is done considering the example in Fig. 5.1.

Fig. 5.1 depicts a 3-echelon distribution system of a single product. First, there is a production facility which supplies a product to a national distribution center (NDC). Second, this national depot center supplies two regional depots (RDC1 and RDC2). Finally, both regional depots supplies two retailers. The retailers which are supplied by the same regional depot is referred to as a *pooling group* (cf. Lee (1987)). We make the assumption that the lead time of every retailer within a pooling group are identical. Furthermore, every stockpoint in the network uses the periodic review mechanism as described in Sect. 2..

Before presenting the analysis of this example we slightly change and introduce some additional notation:

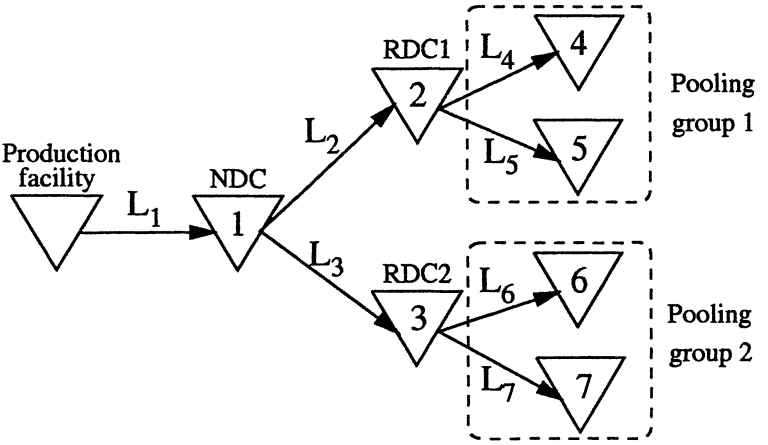


Fig. 5.1. Example of a 3-echelon distribution system.

Network structure

The notation below is clarified by an example in the brackets referring to the situation of Fig. 5.1.

- $ech(i)$ Set of stockpoints that constitute the echelon of stockpoint i (e.g. $ech(2) = \{2, 4, 5\}$).
- $pre(i)$ Preceding stockpoint of stockpoint i (e.g. $pre(3) = 1$).
- V_i All stockpoints which are supplied by i (e.g. $V_1 = \{2, 3\}$).

Lead times and demand

- L_i Lead time between stockpoint $pre(i)$ and stockpoint i,
- D_t^i The demand at the end-stockpoints in $ech(i)$ during t periods.
- D_{t_1,t_2}^i The demand at the end-stockpoints in $ech(i)$ during $(t_1, t_2]$.

Control parameters

- S_i Order-up-to-level of stockpoint i (with respect to echelon inventory position),
- p_i Allocation-fraction from stockpoint $pre(i)$ to stockpoint i,
- Δ_i Maximum physical stock at stockpoint i, $\Delta_i = S_i - \sum_{j \in V_i} S_j$.

The material flow in the upstream part of the network has already been analyzed by Van der Heijden et al. (1997). We explain the behavior of this material flow by the example depicted in Fig. 5.1. At the beginning of a period, say $t - L_1$, the NDC raises the echelon inventory position to S_1. Since the lead time equals L_1, this order arrives at the beginning of period t. So the echelon stock of the NDC just after the arrival of this order equals

$$S_1 - D_{t-L_1,t}^1. \tag{5.1}$$

If this amount (5.1) exceeds the sum of the order-up-to-level of its successors, i.e., $S_2 + S_3$, then both regional depots are able to raise their echelon inventory position to its order-up-to-level. Thus,

$$S_1 - D^1_{t-L_1,t} \geq S_2 + S_3 \quad \Longrightarrow \quad \hat{I}^j_t = S_j, \quad j \in \{2,3\}. \tag{5.2}$$

Otherwise the *complete* echelon stock is allocated to both regional depots by using the rationing policy as introduced in Sect. 2.. Thus,

$$\begin{aligned} &S_1 - D^1_{t-L_1,t} < S_2 + S_3 \quad \Longrightarrow \\ &\hat{I}^j_t = S_j - p_j \left(\sum_{n \in V_1} S_n - (S_1 - D^1_{t-L_1,t}) \right), \quad j \in \{2,3\}, \end{aligned} \tag{5.3}$$

with $\sum_{n \in V_1} p_n = 1$. From (5.2) and (5.3) it follows

$$\hat{I}^j_t = S_j - p_j \left(D^1_{t-L_1,t} - \Delta_1 \right)^+, \quad j \in \{2,3\}. \tag{5.4}$$

Next, we consider one of the regional depots, say RDC1. At the beginning of time t this regional depot places an order at the NCD to raise its echelon inventory position to S_2. However, since the regional depot is supplied by a stockpoint with a finite capacity, it is possible that this order can only be satisfied partially. This (partial) order arrives at the beginning of period $t + L_2$. Hence, at the beginning of period $t + L_2$ the echelon stock of the RDC1 equals

$$\hat{I}^2_t - D^2_{t,t+L_2}. \tag{5.5}$$

If this amount (5.5) exceeds the sum of the order-up-to-level of its successors, i.e., $S_4 + S_5$, then both retailers in pooling group 1 are able to raise the echelon inventory position to its order-up-to-level. Thus,

$$\hat{I}^2_t - D^2_{t,t+L_2} \geq S_4 + S_5 \quad \Longrightarrow \quad \hat{I}^j_{t+L_2} = S_j, \quad j \in \{4,5\}. \tag{5.6}$$

However, if (5.5) is less than $S_4 + S_5$, then the echelon stock of RDC1 is rationed over the retailers in pooling group 1 as follows

$$\begin{aligned} &\hat{I}^2_t - D^2_{t,t+L_2} < S_4 + S_5 \quad \Longrightarrow \\ &\hat{I}^j_{t+L_2} = S_j - p_j \left(\sum_{n \in V_2} S_n - (\hat{I}^2_t - D^2_{t,t+L_2}) \right), \quad j \in \{4,5\}. \end{aligned} \tag{5.7}$$

Substitution of (5.4) in (5.7) yields

$$\hat{I}^j_{t+L_2} = S_j - p_j \left(D^2_{t,t+L_2} - \Delta_2 + p_2 \left(D^1_{t-L_1,t} - \Delta_1 \right)^+ \right)^+, \quad j \in \{4,5\}. \tag{5.8}$$

From equation (5.8) it follows that the cumulative echelon stock of all retailer in pooling group 1 at time $t + L_2 + L_4$ (recall $L_4 = L_5$) equals

$$\sum_{j \in V_2} \hat{I}^j_{t+L_2} - D^2_{t+L_2,t+L_2+L_4}.$$

Substitution of (5.8) in the expression above yields that the cumulative echelon stock of the retailers at $t + L_2 + L_4$ equals

$$\sum_{j \in V_2} S_j - \left(D^2_{t,t+L_2} - \Delta_2 + p_2 \left(D^1_{t-L_1,t} - \Delta_1 \right)^+ \right)^+ - D^2_{t+L_2,t+L_2+L_4}.$$

Like in Sect. 3.2, this stock is reallocated to the retailers by a BS rationing policy. Since $\sum_{j\in V_1} S'_j = \sum_{j\in V_1} S_j$, we obtain

$$\hat{J}^k_{t+L_2+L_4} = S'_k - p_k\left(\left(D^2_{t,t+L_2} - \Delta_2 + p_2\left(D^1_{t-L_1,t} - \Delta_1\right)^+\right)^+ + D^2_{t+L_2,t+L_2+L_4}\right).$$

Analogue to the analysis in Sect. 4.1 it can be shown that

$$\Delta_k = S_k - S'_k = L_k(\mu_k - p_k \sum_{n\in V_j} \mu_n), \quad j \in V_1, \quad k \in V_j.$$

Furthermore, the rebalancing levels $\{S'_k\}$ can be computed similarly as in Sect. 4.2 by solving the service level equation for every retailer k:

$$\beta_k = 1 - \frac{\mathrm{E}[\Psi_{1jk} - S'_k]^+ - \mathrm{E}[p_k\Psi_{2jk} - S'_k]^+}{\mu_k}, \qquad j \in V_1 \quad \text{and} \quad k \in V_j,$$

where

$$\begin{aligned}\Psi_{1jk} &:= p_k\left(\left(D^j_{L_j} - \Delta_j + p_j\left(D^1_{L_1} - \Delta_1\right)^+\right)^+ + D^j_{L_k}\right) + D^k_1,\\ \Psi_{2jk} &:= \left(D^j_{L_j} - \Delta_j + p_j\left(D^1_{L_1} - \Delta_1\right)^+\right)^+ + D^j_{L_k}.\end{aligned}$$

6. Numerical results

In this section we compare the results of our model with the very related models of Van der Heijden (1997) and Diks and de Kok (1996). Van der Heijden analyzes a two-echelon system without transshipments, where the CD adopts the BS rationing policy. In this section we refer to this model as the BS-model. Diks and de Kok analyzes a two-echelon system with lateral transshipments. Both the allocation policy at the CD and the rebalancing policy at the retailers are CAS rationing policies. In this section we refer to this model as TCAS-model. The model as discussed in this paper is referred to as the TBS-model.

Consider the example of Diks and de Kok (1996) where one CD supplies 5 non-identical retailers (see Table 6.1). The depot lead time and the retailer lead time equals 4 and 1, respectively. Every retailer needs to attain the same target fill rate β.

Table 6.1. The demand characteristics of the 5 retailers in the system.

Retailer i	1	2	3	4	5
μ_i	5	10	10	15	20
σ_i^2	25	60	60	225	560

Table 6.2 depicts the safety stock needed to satisfy the fill rate constraints, where the safety stock is defined by $S_0-(L+l+1)\sum_{n=1}^{M}\mu_n$. We computed this safety stock for three different fill rates $\beta = 0.7$, 0.85 and 0.99. Furthermore, we varied the mean stock retained at the CD by considering $\gamma = 0$, 0.9 and 1.2, where $\gamma := \Delta_0/(L\sum_{n=1}^{M}\mu_n)$. From Table 6.2 we conclude that by allowing transshipments the safety stock needed to operate the system decreases considerably (between 8% and 22%). When transshipments are allowed the TCAS-model appears to be slightly better than the TBS-model for $\beta^* = 0.7$, whereas the TBS-model appears to be better than the TCAS-model for $\beta^* = 0.99$. It is hard to draw a well-founded conclusion about which model is the best model. The TBS-model attains fill rates (see Table 6.3) which exceeds the target fill rate for $\beta^* = 0.7$ (and therefore requires more safety stock), whereas the TBS-model attains lower fill rates (see Table 6.3) than the TCAS-model for $\beta^* = 0.99$ (and therefore requires less safety stock).

Table 6.2. Safety stock in a divergent two-echelon system with 5 non-identical retailers ($L = 4$ and $l = 1$).

γ	0	0	0	0.9	0.9	0.9	1.2	1.2	1.2
β	BS	TCAS	TBS	BS	TCAS	TBS	BS	TCAS	TBS
0.70	45	35	38	47	41	43	85	78	81
0.85	103	89	87	110	92	93	141	125	123
0.99	275	259	230	318	266	253	334	288	264

Table 6.3 depicts the customer fill rates obtained by simulation. From this table we conclude that the performance of the heuristic to numerically solve the service level equation (4.13) (see Sect. 4.2) is excellent. All the fill rates attained only slightly differ from the target service levels (at most 0.012), except for the fill rate attained at retailer 1 in the BS-model. This can be explained by the imbalance experienced in the BS-model.

Table 6.4 depicts the percentage that the allocation policy of the CD allocates a negative quantity to at least one of the retailers. From this table we conclude that, indeed, in the BS-model frequently imbalance occurs (especially for $\Delta = 0$). By allowing transshipments imbalance occurs less frequent (especially for the TBS-model).

Table 6.5 depicts the mean amount transshipped per period. For $\beta = 0.7$ the expected total amount transshipped per period in the TBS-model is less than in the TCAS-model (see Table 6.5) (for $\beta = 0.99$ the TCAS-model is slightly better). However, in our opinion the reduction of safety stock is more important than the reduction in transportation costs. The reason for this is that when the model rebalances every review period, then it is not so important anymore how much to transship. The frequency of transshipping plays a more important role in the total costs involved with transshipping

Table 6.3. Attained customer fill rates in a divergent two-echelon system with 5 non-identical retailers ($L = 4$ and $l = 1$).

	γ	0	0	0	0.9	0.9	0.9	1.2	1.2	1.2
β	Retailer	BS	TCAS	TBS	BS	TCAS	TBS	BS	TCAS	TBS
	1	0.748	0.702	0.706	0.724	0.697	0.703	0.709	0.708	0.711
	2	0.703	0.701	0.705	0.694	0.697	0.701	0.697	0.702	0.706
0.70	3	0.703	0.701	0.704	0.694	0.697	0.700	0.697	0.703	0.705
	4	0.706	0.698	0.704	0.696	0.696	0.702	0.698	0.698	0.706
	5	0.708	0.698	0.703	0.695	0.699	0.702	0.695	0.699	0.708
	1	0.878	0.846	0.847	0.870	0.842	0.848	0.858	0.845	0.848
	2	0.848	0.846	0.847	0.849	0.844	0.848	0.849	0.845	0.849
0.85	3	0.846	0.846	0.846	0.848	0.844	0.847	0.848	0.845	0.848
	4	0.842	0.847	0.839	0.845	0.845	0.844	0.847	0.847	0.843
	5	0.840	0.846	0.838	0.847	0.845	0.844	0.841	0.848	0.840
	1	0.991	0.989	0.986	0.994	0.989	0.990	0.992	0.988	0.986
	2	0.987	0.989	0.986	0.991	0.990	0.989	0.990	0.989	0.987
0.99	3	0.987	0.989	0.986	0.991	0.989	0.990	0.990	0.989	0.987
	4	0.981	0.989	0.980	0.989	0.989	0.986	0.989	0.989	0.983
	5	0.979	0.988	0.980	0.990	0.989	0.986	0.989	0.989	0.982

Table 6.4. Percentage that the allocation policy of the CD allocates a negative quantity to at least one of the 5 retailers ($L = 4$ and $l = 1$).

γ	0	0	0	0.9	0.9	0.9	1.2	1.2	1.2
β	BS	TCAS	TBS	BS	TCAS	TBS	BS	TCAS	TBS
0.70	0.43	0.16	0.01	0.28	0.10	0.00	0.10	0.08	0.00
0.85	0.43	0.02	0.01	0.28	0.02	0.00	0.10	0.02	0.00
0.99	0.43	0.01	0.01	0.28	0.01	0.00	0.10	0.01	0.00

than the amount transshipped, since usually these amounts are small. Also note that the expected total amount transshipped per period in the TBS model does not depend on β or γ (this coincides with our approximation (4.12)), whereas in the TCAS-model it both depend on β and γ.

Table 6.5. Mean amount transshipped per period in a divergent two-echelon system with 5 non-identical retailers ($L = 4$ and $l = 1$).

γ	0	0	0.9	0.9	1.2	1.2
β	TCAS	TBS	TCAS	TBS	TCAS	TBS
0.70	19.72	18.36	19.52	18.36	19.39	18.37
0.85	18.62	18.36	18.56	18.36	18.36	18.37
0.99	18.04	18.36	18.03	18.36	18.03	18.37

The impact of the number of retailers on the reduction of the safety stock is analyzed in Table 6.6. It turns out that in the case $\beta = 0.7$ and $\beta = 0.85$ the results of the TCAS-model and the TBS-model are similar, but for $\beta = 0.99$ the TBS-model dominates the TCAS-model. Especially, when the number of retailers is large. Then the reduction of the safety stock can even become 24% (of the safety stock needed in the BS-model).

Table 6.6. Safety stock in a divergent two-echelon system with M identical retailers with $\mu_i = 10$ and $\sigma_i^2 = 80$ ($L = 4$ and $l = 1$).

	γ	0	0	0	0.9	0.9	0.9	1.2	1.2	1.2
β	M	BS	TCAS	TBS	BS	TCAS	TBS	BS	TCAS	TBS
0.70	2	14	13	13	15	15	15	28	27	27
0.70	3	18	16	16	20	18	18	39	37	37
0.70	4	23	19	19	24	21	21	50	46	46
0.70	5	27	21	21	28	24	24	61	55	56
0.85	2	34	32	32	36	33	33	46	43	43
0.85	3	46	41	40	48	42	42	64	57	57
0.85	4	57	49	47	60	50	50	83	72	71
0.85	5	69	57	54	72	59	58	102	85	85
0.99	2	95	89	87	103	92	92	105	92	93
0.99	3	127	116	109	141	119	119	150	125	122
0.99	4	158	143	130	178	146	143	194	158	152
0.99	5	189	172	150	216	174	168	239	195	181

7. Extension of the model

A disadvantage of the model presented so far is that *every* period the echelon stock of the retailers is rebalanced. In principle the retailers should only rebalance, when it is really necessary. E.g., when some retailers face large demands resulting in low inventories, whereas others have excess inventory. In this section we address a possible way of extending the model as described in Sect. 2. such that not every review period rebalancing takes place, but only when it is necessary. How to determine the control parameters for this extended model is a topic for further research.

Suppose that the echelon stock of the retailers is only rebalanced when the cumulative echelon stock drops below a certain level Q. Let the binary random variable X_t we indicate whether the stock is rebalanced ($X_t = 1$) or not ($X_t = 0$). Then from (3.8) it follows that

$$\delta_{t+l} := Pr(X_{t+l} = 1) = Pr\Big(\sum_{n=1}^{M} S_n - \big(D_{t-L_1,t} - \Delta_1\big)^+ - D_{t,t+l} < Q\Big). \quad (7.1)$$

Note that this δ_{t+l} is independent of t since we assumed stationair customer demand. Therefore we define δ as the probability of rebalancing the cumulative echelon stock at the retailers at the end of an arbitrary period. Think of δ as a managerial parameter. Management probably determines the value of δ based on the trade off between the fixed costs of rebalancing and the decrease in costs due to making better use of the available stock in the system. Given this value of δ it is immediately clear how to choose Q such that the frequency of transshipping matches the target set by management. From (7.1) we obtain

$$Q = \sum_{n=1}^{M} S_n - F_Z^{-1}(1-\delta) \quad \text{with} \quad Z := \big(D_{L_1} - \Delta_1\big)^+ + D_l.$$

In our opinion, this an appropriate way to extend the model such that not every review period the echelon stock of the retailers is rebalanced. There are several reasons for this. First, when the cumulative echelon stock of the retailers is small it is important that this stock is appropriately distributed among the retailers. The smaller this cumulative echelon stock becomes the larger the probability becomes of having one or more retailers facing backorders while others have excess stock. Second, this way of incorporating that the retailers do not rebalance their echelon stock every period is still analytically tractable. Finally, management can simply influence the frequency of transshipping by setting only one parameter.

8. Conclusions and further research

In this paper we considered a two-echelon distribution system consisting of a central depot supplying a number of retailers. Every review period an in-

stantaneous rebalancing of the cumulative echelon stock of the retailers takes place, by transshipping stock from one retailer to another. The rebalancing policy used is the BS rationing policy of Van der Heijden (1997). Also the CD adopts this BS rationing policy to allocate incoming stock to the retailers. In this paper we determine all the control parameters of the inventory system such that the target service levels are satisfied. Furthermore, we minimize the total expected stock transshipped per period.

The policy derived in this paper (referred to as TBS) is easy to implement and very robust. The allocation fractions of the rationing policy at the CD and of the rebalancing policy at the retailers are given by (3.5). The parameters Δ_i is determined by (4.10). Only for the computations of the order-up-to-levels we used a heuristic. Numerical results indicate that the performance of this heuristic is excellent. The TBS-policy can easily be extended to divergent N-echelon systems (see Sect. 5.). Also we compared the results of the TBS-policy with:

1. the BS-policy of Van der Heijden (1997), who analyzes a divergent two-echelon system without lateral transshipments, where the CD adopts a BS rationing policy.
2. the TCAS-policy of Diks and de Kok (1998b), who analyzes the same model as analyzed in this paper, except they adopt the CAS rationing policy instead of the BS policy.

By allowing lateral transshipments the TCAS model needs far less safety stock than the BS-policy. For the examples considered in Table 6.3 the mean reduction of safety stock is 13%. The TBS-policy even accomplishes a larger reduction, since the mean reduction is 15%. Furthermore, the amount of imbalance for the TBS-policy is negligible (less than the TCAS-policy, and considerably less than the BS-policy; see Table 6.4).

Finally, in Sect. 7. we made a first step to extend the model such that rebalancing does not takes place every review period, but only when the echelon stock drops below a critical level. More research should be done on this model, to analyze the trade off between the decrease of the rebalancing set-up costs and the increase in safety stock needed to satisfy the customer service levels.

References

Burgin, T. (1975): The gamma distribution and inventory control. *Operational Research Quarterly* **26**, 507–525.

Clark, A.J., Scarf, H. (1960): Optimal policies for a multi-echelon inventory problem. *Management Science* **6**, 475–490.

Cohen, M.A., P.R. Kleindorfer, P.R., Lee, H.L. (1986): Optimal stocking policies for low usage items in multi-echelon inventory systems. *Naval Research Logistics Quarterly* **33**, 17–38.

Diks, E.B., Kok, A.G. de (1996): Controlling a divergent two-echelon network with transshipments using the consistent appropriate share rationing policy. *International Journal of Production Economics* **45**, 369–379.

Diks, E.B., Kok, A.G. (1998a): Computational results for the control of a divergent N-echelon inventory system. *International Journal of Production Economics*, In press.

Diks, E.B., Kok, A.G. de (1998b): Optimal control of a divergent N-echelon inventory system. *European Journal of Operational Research*, To appear.

Diks, E.B., Kok, A.G. de, Lagodimos, A.G. (1996): Multi-echelon systems: A service measure perspective. *European Journal of Operational Research* **95**, 241–263.

Donselaar, K. van, Wijngaard, J. (1987): Commonality and safety stocks. *Engineering Costs and Production Economics* **12**, 197–204.

Eppen, G., Schrage, L. (1981): Centralized ordering policies in a multi-warehouse system with lead times and random demand. in: L.B. Schwarz (ed.), *Multi-level Production/Inventory Control Systems: Theory and Practice*, North-Holland Publishing Company, 51–67.

Federgruen, A. (1993): Centralized planning models for multi-echelon inventory systems under uncertainty. in: Graves, S.C., Rinnooy Kan, A.H.G., Zipkin, P.H. (eds.), *Logistics of production and inventory*, Handbooks in Operations Research and Management Science 4, Elsevier Science Publishers B.V., Amsterdam, North-Holland, Chapter 3, 133–173.

Hadley, G., Whitin, T.M. (1963): *Analysis of Inventory Systems*. Prentice-Hall Inc., Englewood Cliffs, New Jersey.

Heijden, M.C. van der (1997): Supply rationing in multi-echelon divergent systems. *European Journal of Operational Research*, To appear.

Heijden, M.C. van der, Diks, E.B., Kok, A.G. de and A.G. de Kok (1997): Stock allocation in general multi-echelon distribution systems with (R, S) order-up-to-policies. *International Journal of Production Economics* **49**, 157–174.

Hoadley, H., Heyman, D.P. (1977): A two-echelon inventory model with purchases dispositions shipments, returns and transshipments. *Naval Research Logistics Quarterly* **24**, 1–19.

Houtum, G.J. van, Inderfurth, K., Zijm, W.H.M. (1996): Materials coordination in stochastic multi-echelon systems. *European Journal of Operational Research* **95**, 1–23.

Houtum, G.J. van, Zijm, W.H.M. (1991): Computational procedures for stochastic multi-echelon production systems. *International Journal of Production Economics* **23**, 223–237.

Jönsson, H., Silver, E.A. (1987): Analysis of a two-echelon inventory control system with complete redistribution. *Management Science* **33**, 215–227.

Karmarkar, U.S., Patel, N.R. (1977): The one-period, N-location distribution problem. *Naval Research Logistics Quarterly* **24**, 559–575.

Kok, A.G. de (1990): Hierarchical production planning for consumer goods. *European Journal of Operational Research* **45**, 55–69.

Kok, A.G. de, Lagodimos, A.G., Seidel, H.P. (1994): *Stock allocation in a two-echelon distribution network under service-constraints*. EUT 94-03, Department of Industrial Engineering and Management Science, Eindhoven University of Technology.

Klein, M., Dekker, R. (1997): Using break quantities for tactical optimisation in multistage distribution systems. in: Fleischmann, B., Nunen, J. van, Speranza, G., Stähly, P. (eds.), *Advances in Distribution Logistics*, series Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Chapter 17.

Lagodimos, A.G. (1992): Multi-echelon service models for inventory systems under different rationing policies. *International Journal of Production Research* **30**, 939–958.

Langenhoff, L.J.G., Zijm, W.H.M. (1990): An analytical theory of multi-echelon production/ distribution systems, *Statistica Neerlandica* **44**, 149–174.

Lee, H.L. (1987): A multi-echelon inventory model for repairable items with emergency lateral transshipments, *Management Science* **33**, 1302–1316.

Seidel, H.P., Kok, A.G. de (1990): *Analysis of stock allocation in a 2-echelon distribution system.* Technical Report 098, CQM, Eindhoven.

Silver, E.A., Peterson, R. (1985): *Decision Systems for Inventory Management and Production Planning*, John Wiley and Sons, New York.

Tagaras, G. (1989): Effects of pooling on the optimization and service levels of two-location inventory systems, *IIE Transactions* **21**, 250–257.

Tagaras, G., Cohen, M.A. (1992): Pooling in two-location inventory systems with non-negligible replenishment lead times. *Management Science* **38**, 1067–1083.

Tijms, H.C. (1994): *Stochastic Models: an Algorithmic Approach*, John Wiley and Sons, Chichester.

Tüshaus, U., Wahl, C. (1997): Inventory Positioning in a Two-stage Distribution System with service Level Constraints. in: Fleischmann, B., Nunen, J. van, Speranza, G., Stähly, P. (eds.), *Advances in Distribution Logistics*, series Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Chapter 20.

A. Initial state of the system

The initial state of the system influences τ_t^i, since $\tilde{\tau}_t^i = \tilde{\tau}_{t+kl}^i$ for any integer value k (see equation (4.4)). In Sect. 4.1 it is assumed that τ_t^i is a stationair process, i.e., $E[\tau_t^i] = E[\tau_{t+1}^i]$ for any t. In this appendix we determine how to initialize the system such that, indeed, τ_t^i is a stationair process. Suppose the initial state of the system is characterized by $\{z_k^i\}_{k=0}^l$, where z_0^i denotes the echelon stock of retailer i, and z_k^i $(1 \leq k \leq l)$ denotes the number of products to arrive at retailer i in k periods. The pipeline and on hand stock of the CD are irrelevant. At the end of the first period holds

$$\begin{aligned} J_1^i &= z_0^i + z_1^i - D_{0,1}^i, \\ \hat{J}_1^i &= S_i' - p_i \left(\sum_{n=1}^{M} S_n' - \sum_{n=1}^{M} (z_0^n + z_1^n) + D_{0,1} \right). \end{aligned}$$

Since $\tau_1^i = \hat{J}_1^i - J_1^i$ we obtain

$$\tau_1^i = S_i' - p_i \left(\sum_{n=1}^{M} S_n' - \sum_{n=1}^{M} (z_0^n + z_1^n) + D_{0,1} \right) - (z_0^i + z_1^i) + D_{0,1}^i.$$

By definition $\tilde{\tau}_1^i = D_{0,1}^i - p_i D_{0,1} - \tau_1^i$. Hence,

$$\tilde{\tau}_1^i = (z_0^i + z_1^i) - p_i \sum_{n=1}^{M} (z_0^n + z_1^n) + p_i \sum_{n=1}^{M} S_n' - S_i'. \tag{A.1}$$

For $1 < t \leq l$ we know that $\tau_t^i = \hat{J}_t^i - J_t^i = \hat{J}_t^i - (\hat{J}_{t-1}^i - D_{t-1,t}^i + z_t^i)$. Substitution of $\hat{J}_t^i = S_i' - p_i(\sum_{n=1}^M S_n' - \sum_{n=1}^M \hat{J}_t^n)$ in the aforementioned equation yields

$$\tau_t^i = p_i \left(\sum_{n=1}^M \hat{J}_t^n - \sum_{n=1}^M \hat{J}_{t-1}^n \right) + D_{t-1,t}^i - z_t^i, \qquad 1 < t \leq l.$$

Note that $\sum_{n=1}^M \hat{J}_t^n - \sum_{n=1}^M \hat{J}_{t-1}^n$ equals the number of products arriving at all retailers at time t minus the demand at all retailers during $(t-1, t]$. Hence,

$$\tau_t^i = p_i \left(\sum_{n=1}^M z_t^n - D_{t-1,t} \right) + D_{t-1,t}^i - z_t^i, \qquad 1 < t \leq l.$$

Substitution of τ_t^i in the definition of $\tilde{\tau}_t^i$ yields

$$\tilde{\tau}_t^i = z_t^i - p_i \sum_{n=1}^M z_t^n. \tag{A.2}$$

Substitution of the equations (A.1) and (A.2) in $\tilde{\tau}_t^i = \Delta_i / l$, after replacing Δ_i by the expression given by (4.11), yields

$$\begin{cases} z_0^i + z_1^i &= \mu_i + p_i \sum_{n=1}^M (z_0^n + z_1^n - \mu_n) + S_i' - p_i \sum_{n=1}^M S_n', \\ z_t^i &= \mu_i + p_i \sum_{n=1}^M (z_t^n - \mu_n), \qquad 1 < t \leq l. \end{cases}$$

Possible initial states satisfying the above equalities are

$$z_0^i := S_i' - p_i \sum_{n=1}^M S_n', \qquad z_t^i := \mu_i, \quad 1 \leq t \leq l, \quad \text{or,}$$

$$z_0^i := \mu_i, \quad z_1^i := S_i' - p_i \sum_{n=1}^M S_n', \quad z_t^i := \mu_i, \quad 2 \leq t \leq l.$$

Unfortunately, due to imbalance we cannot guarantee the stationarity of τ_t^i.

B. Derivation of $\mathrm{E}[X - c]^+$ and $\mathrm{var}[X - c]^+$

In this section we derive a tractable expression for $\mathrm{E}[X - c]^+$, where X is distributed as a mixture of two Erlang distributions. Suppose X follows an E_{k_1,λ_1} distribution with probability θ_1, and an E_{k_2,λ_2} distribution with probability $\theta_2 := 1 - \theta_1$.
Then,

$$
\begin{aligned}
\mathrm{E}[X-c]^+ &= \int_c^\infty (x-c)\, dF_X(x) = \int_0^\infty 1 - F_X(x+c)\, dx \\
&= \int_c^\infty 1 - F_X(x)\, dx = \int_c^\infty \sum_{i=1}^{2} \theta_i \sum_{j=0}^{k_i-1} \frac{(\lambda_i x)^j}{j!} e^{-\lambda_i x}\, dx \\
&= \sum_{i=1}^{2} \frac{\theta_i}{\lambda_i} \sum_{j=0}^{k_i-1} \int_c^\infty \lambda_i^{j+1} \frac{x^j}{j!} e^{-\lambda_i x}\, dx \\
&= \sum_{i=1}^{2} \frac{\theta_i}{\lambda_i} \sum_{j=0}^{k_i-1} \sum_{n=0}^{j} \frac{(\lambda_i c)^n}{n!} e^{-\lambda_i c} \\
&= \sum_{i=1}^{2} \frac{\theta_i}{\lambda_i} \sum_{n=0}^{k_i-1} (k_i - n) \frac{(\lambda_i c)^n}{n!} e^{-\lambda_i c}.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}([X-c]^+)^2 &= \int_c^\infty (x-c)^2\, dF_X(x) = 2\int_0^\infty x(1 - F_X(x+c))\, dx \\
&= 2\int_0^\infty x \sum_{i=1}^{2} \theta_i \sum_{j=0}^{k_i-1} \frac{(\lambda_i(x+c))^j}{j!} e^{-\lambda_i(x+c)}\, dx \\
&= 2\int_0^\infty x \sum_{i=1}^{2} \theta_i \sum_{j=0}^{k_i-1} \frac{\lambda_i^j}{j!} e^{-\lambda_i c} \sum_{n=0}^{j} \binom{j}{n} c^{j-n} x^n e^{-\lambda_i x}\, dx \\
&= 2\sum_{i=1}^{2} \theta_i \sum_{j=0}^{k_i-1} \frac{\lambda_i^j}{j!} e^{-\lambda_i c} \sum_{n=0}^{j} \binom{j}{n} c^{j-n} \int_0^\infty x^{n+1} e^{-\lambda_i x}\, dx \\
&= 2\sum_{i=1}^{2} \theta_i \sum_{j=0}^{k_i-1} \frac{\lambda_i^j}{j!} e^{-\lambda_i c} \sum_{n=0}^{j} \binom{j}{n} c^{j-n} \frac{(n+1)!}{\lambda_i^{n+2}} \\
&= 2\sum_{i=1}^{2} \frac{\theta_i}{\lambda_i^2} \sum_{j=0}^{k_i-1} \sum_{n=0}^{j} (j+1-n) \frac{(\lambda_i c)^n}{n!} e^{-\lambda_i c}.
\end{aligned}
$$

From the first two moments of $[X-c]^+$ it follows that $\mathrm{var}[X-c]^+ = \mathrm{E}([X-c]^+)^2 - \mathrm{E}^2[X-c]^+$.

Reverse logistics and inventory control with product remanufacturing

E.A. van der Laan[1], M. Salomon[2], and J.A.A.E. van Nunen[1]

[1] Erasmus Universiteit Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

[2] Katholieke Universiteit Brabant, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract In this paper we investigate product remanufacturing in general, and the influence of product remanufacturing on production planning and inventory control. The contents of this paper are as follows: *(i)* we discuss some important issues concerning the application of product remanufacturing in industry, *(ii)* we review the literature on production planning and inventory control models that apply to the situation of remanufacturing, *(iii)* we investigate the steady-state behaviour of re-order point strategies with remanufacturing, and *(iv)* we indicate some new areas for future research in production planning and inventory control with remanufacturing.

Keywords: Production planning and inventory control, remanufacturing, re-order point strategies for inventory control, stochastic models, markov chains.

1 Introduction

The traditional approach of many manufacturers towards used products has been to ignore the issue (see Thierry et al. (1995)). Manufacturers did not feel responsible for handling their products after customer use. Consequently, most products were designed in such a way that while materials, assembly, and distribution costs were minimized, repair, reuse, and disposal costs were not taken into account. Traditionally, the majority of used products were landfilled or incinerated with considerable damage to the environment. Today, customers and authorities demand that manufacturers reduce the waste generated by their products. Some companies consider this as a threat to their business.

Clearly, complying with rapidly changing environmental regulations may require a fundamental change in doing business. However, there are large opportunities for companies that succeed in embodying current and future environmental demands in their business policy. For instance, by offering 'green and reusable products' companies may be able to attract and retain environmental-conscious customers. Furthermore, production of reusable

products may lead to improved product quality, which may also attract new customers.

Product recovery management may be defined as 'the management of all used and discarded products, components, and materials for which a manufacturing company is legally, contractually, or otherwise responsible' (Thierry et al. (1995)). In product recovery management several options exist to handle products after customer usage. The choice of the 'best' product recovery option in a particular practical setting depends on many factors, like environmental legislation, available technology in the production process, and costs. One possible recovery option which is central in this paper is *remanufacturing*. With remanufacturing, used products are recovered such that the quality standards are as strict as those for new products. After remanufacturing they can be sold or leased in the market of *new* products. Other possible options are:

Repair. Products are brought to working order. This implies that typically the quality standards of repaired products is less than those for new products. Usually repair requires minor (dis)assembly, since only the non-working parts are repaired or replaced.

Refurbishing. Products are upgraded to some prespecified quality standards. Typically these standards are less than those for new products but higher than those for repaired products.

Cannibalization. This involves selective disassembly of used products and inspection of potentially reusable parts. Parts obtained from cannabilization can be reused in the repair, refurbishing or remanufacturing process.

Recycling. Materials rather than products are recovered. These materials are reused in the manufacturing of new products.

Disposal. Products are disposed off after return from the customer.

Initiated by the arguments listed in the first paragraph, a growing number of industries is now becoming interested in remanufacturing. In some 'high-tech' industries, like in the aircraft industry, the automobile industry, the computer industry, and the medical instrument industry, remanufacturing has already been implemented. Table 1.1 lists some large companies within these industries that currently apply product remanufacturing.

Companies that apply product remanufacturing usually have to re-engineer a large number of processes within their organization. For instance, remanufacturing will have influences on processes related to product design, internal and external logistics (see e.g. this volume), information systems, marketing, quality control, production planning, and inventory control (see Thierry et al. (1995)). This paper is restricted to an investigation of the process related to *inventory control.*

Table 1.1. Some companies active in remanufacturing.

Company name	*Product*	*References*
De Vlieg-Bullard	Machine tools	Sprow (1992)
Abbott Laboratories	Medical diagnostic instruments	Sivinski and Meegan (1993)
Volkswagen Canada	Car engines	Brayman (1992)
Grumman	F-14 aircraft	Kandebo (1990)
Rank Xerox	Copiers	Thierry et al. (1995)
BMW	Car engines, starting motors, alternators	Vandermerwe and Oliff (1991)

The main objectives of this paper are to introduce the issue of remanufacturing and to demonstrate the practical relevance of research on remanufacturing (this section), and to define the processes that play an important role in inventory control with remanufacturing (Section 2). In Section 3 we review the literature on inventory control models with remanufacturing in Section 4 we show some of the effects of remanufacturing on the total costs of inventory control if the inventory system is controlled by a simple re-order point strategy. Finally, we indicate some interesting areas for future research in inventory control with remanufacturing in Section 5.

2 Important processes in remanufacturing

In studies that we carried out on remanufacturing we observed that a number of processes play an important role in inventory control systems (see Thierry et al. (1995)). A schematic representation of these processes and their flows for a single-product, two-component system is given in Figure 2.1.
The processes of Figure 2.1 are defined as follows:

The *demand* process generates the market demands for new products.

The *return* process generates the returns of used products from the market.

The *disassembly* process models the operations that are carried out to disassemble returned products into modules or components. Disassembly is sometimes necessary to test the quality and reusability of each individual module or component. Furthermore, disassembly provides the possibility to reuse some modules or components (cannibalization), and to dispose others.

The *testing* process models all operations that are necessary to test returned products, modules, or components. In general, if a returned product, module, or component passes the test, it may be remanufactured. If a returned product, module, or component fails at testing, it is disposed off.

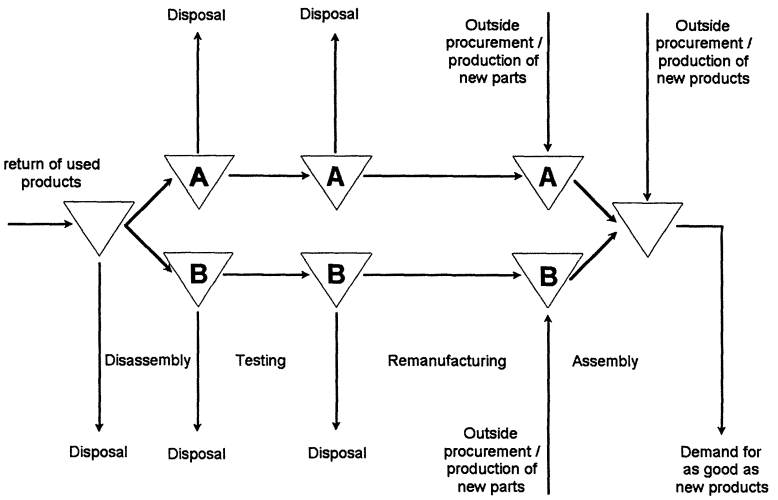


Figure 2.1. A schematic representation of the relevant processes and goods-flows in a hybrid manufacturing/remanufacturing system for one product consisting of two components A and B.

The *remanufacturing* process itself models all operations that are necessary to upgrade returned products. Output of the remanufacturing process is a product that has the same quality standards and other characteristics as a new product.

The *outside procurement/production* process, which models the outside procurement (external supply) or production (internal supply) of new products, modules, or components.

The *assembly* process models the in-house assembly of outside procured or inside produced products from modules or components.

The *inventory* process models the operations related to the inventory buffers in the production and inventory system.

The *disposal* process models the operations necessary to dispose products, modules, or components.

The extent to which the above processes and their interactions are modelled and controlled in mathematical inventory control models determines when and how these models are applicable in situations with remanufacturing.

3 Literature review

Many articles have appeared in the production planning and inventory control literature in which both the return process and the demand process are explicitly modelled. Excellent reviews can be found in Nahmias (1981) and in Cho and Parlar (1991). From the literature in which demand and return processes are considered simultaneously, we will focus the discussion on those models that directly apply to the situation of remanufacturing. As selection criteria we use that *in addition* to the demand and return process, the models must describe the remanufacturing process (either implicitly, or explicitly), *and* the production or outside procurement process.

We do not consider models in which demands for new products are generated by product failures only, i.e., product demands and product returns are perfectly correlated. These models are typical for the situation of spare part (repair) management, but do usually not apply to the situation with remanufacturing. Another difference between models for spare part management and remanufacturing lies in the objective: with spare part management the objective is to determine the *fixed* number of spare parts in the system, such that the associated long-run average costs are minimized. With remanufacturing the objective is to develop a policy on when and how much to remanufacture, dispose, and produce, such that some cost function is minimized. Essential is that with remanufacturing the number of products in the system may vary over time. Our selection criteria imply that the well-known family of METRIC models (see Sherbrooke (1968)) for spare-part management will not be considered here.

Before we review the literature on models that satisfy our selection criteria, we first list the most common assumptions that are made in these models with respect to the processes indicated in the previous sections.

Demand and return process. To model the demand and return process, assumptions are made on the inter-occurrence times, the demand quantity per occurrence, and the relation between the two processes (i.e., stochastically dependent or independent).

Disassembly process. This process is not considered in any model in the literature. The reason is, that all models apply to the situation of a *single* product, and each product is assumed to consist of a *single* component only. Clearly, in this situation no disassembly operations need to occur.

Testing process. This process is modelled by means of a single testing facility, where returned products are tested. Assumptions are made concerning the testing capacity, the testing time, and the variable testing costs.

Remanufacturing process. This process is modelled by means of a single remanufacturing facility, consisting of a number of parallel workcenters that

carry out the remanufacturing operations. Assumptions are made on the number of parallel workcenters, the remanufacturing time, and the variable remanufacturing costs,

Outside procurement/production process. This process is modelled in terms of an outside procurement source (external supply) or production resource (internal supply). With respect to this process assumptions are made on fixed and variable costs, on the lead-time in case of outside procurement, and on the production time and production capacity in case of internal production.

Assembly process. See *disassembly* process.

Inventory process. Two types of inventory are modelled. Type I inventory is the inventory of returned products that have passed the test and are waiting for remanufacturing or planned disposal. Type II inventory is the inventory of all *serviceable products*, i.e., products that were remanufactured or newly produced. For both types of inventories, assumptions are made on storage capacity and inventory holding costs,

Disposal process. The disposal process is modelled by means of a disposal center. Assumptions are made on fixed and variable disposal costs.

In addition to the above classification of processes, we also make a distinction between two types of *customer service*, i.e., customer service in terms of backorder costs per product per time unit, and customer service in terms of a service level. In case of periodic review this level, $\delta_b^{(n)}$ say, is defined as the maximum allowable probability of a stock-out occurrence in between two successive outside procurements or internal production runs in period n. In case of continuous review this level is defined as the long-run average maximum allowable probability of a stockout position.

In the following sections we separately discuss models with discrete planning periods (periodic review models, Section 3.1), models with continuous planning opportunity (continuous review models, Section 3.2), and a particular type of financial models that is related to the topic of remanufacturing (cash-balancing models, Section 3.3).

3.1 Periodic review models

In *periodic review* models the planning horizon is subdivided into a predetermined (in)finite number of planning periods. At the beginning of each planning period n decisions are taken according to the values of the following decision variables:

$Q_d^{(n)}$ = the quantity (batch-size) of products that is disposed of in planning period n,

$Q_p^{(n)}$ = the quantity (batch-size) of products that is procured outside or internally produced in planning period n,

$Q_r^{(n)}$ = the quantity (batch-size) of products that is remanufactured in planning period n.

All decision variables are assumed to be integer.

Objective in the periodic review models is to determine the values for the decision variables, such that total expected costs over the entire planning horizon are minimized. Some models also take the service level explicitly into account as a constraint.

Within the category of periodic review models, Simpson (1978) considers a model, with the following assumptions and characteristics:

Demand and return process. One demand and return occurrence per planning period, demand and return quantities are correlated and specified by means of a period dependent joint probability density function.

Testing process. No testing facility.

Remanufacturing process. No remanufacturing lead-time; the capacity of the remanufacturing facility is infinite.

Procurement/production process. No procurement lead-time; no fixed procurement costs.

Inventory process. Type I and Type II inventory buffers have infinite capacity; Type I and Type II inventory have (different) variable inventory holding costs.

Disposal process. No fixed or variable disposal costs.

Customer service. Modelled in terms of backorder costs,

Control strategy. At the beginning of each period n decisions are taken on $Q_d^{(n)}$, $Q_p^{(n)}$, and $Q_r^{(n)}$, such that the total expected costs over the planning horizon are minimized.

Simpson develops a dynamic programming based algorithm to determine the optimal values for the above decision variables. Also, an interesting structure on the optimal decisions is identified. It is proved that for each period n there exist three constants α_n, β_n, and γ_n, such that the optimal strategy is as follows:

- If at the beginning of period n Type II inventory is smaller than α_n, as many as possible products from Type I inventory will be remanufactured to increase Type II inventory to α_n.

- If Type I inventory is insufficient to increase Type II inventory to β_n ($< \alpha_n$), an outside procurement order is placed to increase Type II inventory to β_n.
- If after the previous steps the sum of Type I inventory and Type II inventory is larger than $\alpha_n + \gamma_n$, as many as possible products from Type I inventory are disposed of, such that after disposal the sum of Type I and Type II inventory is no less then $\alpha_n + \gamma_n$.

Kelle and Silver (1989) formulate a periodic review model, which differs somewhat from Simpson's model.

Demand and return process. One demand and return occurrence per planning period, but opposed to Simpson's model, demand and return quantities in each period are *independent* stochastic variables.

Testing process. No testing facility.

Remanufacturing process. No remanufacturing lead-time; infinite capacity of the remanufacturing facility; no remanufacturing costs (This model applies to containers, bottles, etc.).

Procurement/production process. No procurement lead-time; fixed and variable production costs.

Inventory process. No Type I inventory; Type II inventory buffer has infinite capacity,

Disposal process. No disposal,

Customer service. Modelled in terms of service level constraint,

Control strategy: at the beginning of each period n a decision is taken on $Q_p^{(n)}$, such that the total expected costs over the planning horizon are minimized.

Kelle and Silver formulate their model as a chance constraint integer program. The chance constraints state that the probability on a backlogging position at the end of period n may not be larger then $\delta_b^{(n)}$. They suggest an approximation procedure to solve the chance constraint integer program.

First step in the approximation procedure is to replace the stochastic inventory variables by their expectations, and to replace the probabilistic service level constraints by appropriate deterministic constraints on the minimum inventory level at the end of each period. These transformations yield a variant of the well-known (deterministic) Wagner-Whitin model for dynamic lotsizing. In this variant positive as well as negative demands are allowed to occur in each period. Second step in the approximation procedure is to transform this variant of the Wagner-Whitin model into an equivalent model

in which positive demands occur only. The latter model is then solved to optimality, using an appropriate dynamic programming based technique.

Inderfurth (1996) extends the work of Simpson to allow for non-zero remanufacturing lead-times:

Demand and return process. All returns and demands per period are continuous time-independent random variables. The inter-arrival distributions are arbitrary distribution functions, which may be stochastically dependent.

Testing process. No testing facility.

Remanufacturing process. The remanufacturing lead-time L_r is non-stochastic and equal to μ_{L_r}; the remanufacturing facility has infinite capacity and variable remanufacturing costs.

Procurement/production process. The procurement lead-time L_m is non-stochastic and equal to μ_{L_m}; there are variable production costs.

Inventory process. Both Type I and Type II inventory buffers have infinite capacity.

Disposal process. Variable disposal costs.

Service. Modelled in terms of backorder costs.

Control strategy: at the beginning of each period n decisions are taken on $Q_d^{(n)}$, $Q_p^{(n)}$, and $Q_r^{(n)}$, such that the total expected costs over the planning horizon are minimized.

Inderfurth considers several special cases, regarding the stocking policy of the returned products, and regarding the values of the manufacturing lead-time μ_{L_m} and the remanufacturing lead-time μ_{L_r}. For the case that returned items are not allowed to be stocked, for instance because the items are perishable, Inderfurth provides the following results.

- If $\mu_{L_m} = \mu_{L_r}$ the structure of the optimal policy can be formulated as a simple (L, U) policy:

$$\begin{array}{llll} Q_p^{(n)} = L^{(n)} - x_s, & Q_r^{(n)} = x_r, & Q_d^{(n)} = 0, \text{ for} & x_s < L^{(n)}, \\ Q_p^{(n)} = 0, & Q_r^{(n)} = x_r, & Q_d^{(n)} = 0, \quad L^{(n)} \leq & x_s \leq U^{(n)}, \\ Q_p^{(n)} = 0, & Q_r^{(n)} = x_r - (x_s - U^{(n)}), & Q_d^{(n)} = x_s - U^{(n)}, & x_s > U^{(n)}, \end{array}$$

Here, x_r is the remanufacturable inventory and x_s is the inventory position of serviceable products.

- If $\mu_{L_m} > \mu_{L_r}$ the structure of the optimal policy can be formulated as a three parameter $(L, U, \hat{U})$ policy.

- If $\mu_{L_m} < \mu_{L_r}$ the structure of the optimal policy is not of a simple form, even if the manufacturing lead-time and remanufacturing lead-time differ only one period.

Note that if returned products are not allowed to be stocked, all returned products will be remanufactured or disposed off at the end of each period. On the other hand, if returned products are allowed to be stocked there can be more interaction between the production, the remanufacturing and the disposal process. In the latter case Inderfurth derives the following results concerning the structure of the optimal policy.

- If $\mu_{L_m} = \mu_{L_r}$ the structure of the optimal policy can be formulated as a (L, U, M) policy:

$$
\begin{array}{llll}
Q_p^{(n)} = L^{(n)} - x, & Q_r^{(n)} = x_r, & Q_d^{(n)} = 0, & \text{for} \quad x < L^{(n)}, \\
Q_p^{(n)} = 0, & Q_r^{(n)} = x_r, & Q_d^{(n)} = 0, & L^{(n)} \leq x < M^{(n)}, \\
Q_p^{(n)} = 0, \; Q_r^{(n)} = M^{(n)} - x_s, & & Q_d^{(n)} = 0, & M^{(n)} \leq x \leq U^{(n)}, \quad x_s < M^{(n)}, \\
Q_p^{(n)} = 0, & Q_r^{(n)} = 0, & Q_d^{(n)} = 0, & M^{(n)} \leq x \leq U^{(n)}, \quad x_s \geq M^{(n)}, \\
Q_p^{(n)} = 0, \; Q_r^{(n)} = M^{(n)} - x_s, & & Q_d^{(n)} = x - U^{(n)}, & x > U^{(n)}, \quad x_s < M^{(n)}, \\
Q_p^{(n)} = 0, & Q_r^{(n)} = 0, & Q_d^{(n)} = x - U^{(n)}, & x > U^{(n)}, \quad x_s \geq M^{(n)}.
\end{array}
$$

  Here, x_r is the remanufacturable inventory, x_s is the inventory position of serviceables, and $x = x_s + x_r$.

- If $\mu_{L_m} \neq \mu_{L_r}$ the structure of the optimal policy is much more difficult to obtain and becomes very complex, even if the manufacturing lead-time and remanufacturing lead-time differ only one period.

3.2 Continuous review models

In *continuous review* models the time axis is continuous, and decisions are taken according to some predefined *control policy*. For the control policies considered in literature, the following integer valued decision variables are defined:

s_p = inventory position (i.e., the sum of Type I and Type II inventory) at which an outside procurement or production order is placed,

Q_p = the quantity (batch-size) that is outside procured or produced,

s_d = inventory position at which returned products are disposed of,

Q_d = quantity (batch-size) of returned products that is disposed of.

The objective is to determine values for the decision variables, such that the long-run average costs per unit of time are minimized. In some models also a service level constraint is explicitly taken into account.

Within the category of continuous review models, Heyman (1977) analyses a model with the following assumptions and characteristics:

Demand and return process. Demands and returns are independent. The inter-occurrence times and quantities are distributed according to general distribution functions.

Testing process. No testing facility.

Remanufacturing process. No remanufacturing lead-times; the capacity of the remanufacturing facility is infinite; variable remanufacturing costs.

Procurement/production process. No procurement lead-times; variable outside procurement costs.

Inventory: Type I inventory is not modelled; the Type II inventory buffer has infinite capacity; variable holding costs of Type II inventory are explicitly taken into account.

Disposal. Variable disposal costs,

Customer service. The system has perfect service, since backlogging never occurs due to zero remanufacturing and outside procurement lead-times.

Control strategy: the system is controlled by a single parameter s_d strategy: whenever the inventory position equals s_d, incoming remanufacturables are disposed of.

Heyman presents an expression for the disposal level s_d such that the sum of inventory holding costs, variable remanufacturing costs, variable outside procurement costs, and variable disposal costs is minimized. In case that the inter-arrival times of demands and returns are exponentially distributed, the *exact* expression is based on the analogy between this inventory model, and a

simple queueing model. Heyman proves that here the single parameter control rule dominates all other possible control rules in terms of total expected costs, i.e. no alternative control rule can ever result in lower expected costs.

In case that the inter-arrival times of demands and returns are generally distributed, Heyman derives an approximation procedure to determine the disposal level for which the total expected costs are minimal. The approximation procedure is based on diffusion processes. A small numerical study in his paper shows that the approximation procedure performs rather well.

Muckstadt and Isaac (1981) consider a model that extends Heyman's in the sense that a remanufacturing facility is explicitly modelled, and lead-times are non-zero. However disposal decisions are not taken into account:

Demand and return process. Demands and returns are independent; the inter-occurrence times are exponentially distributed; the demand and return quantities are always equal to one product per occurrence. To avoid unlimited growth of inventories it is assumed that the return rate is smaller than the demand rate.

Testing process. No testing facility.

Remanufacturing process. The remanufacturing lead-time is arbitrarily distributed; the capacity of the remanufacturing facility may be finite.

Procurement/production process. The procurement lead-time is constant; fixed outside procurement costs are considered.

Inventory: Type I and Type II inventory buffers have infinite capacity; inventory holding costs are taken into account for Type II inventory only.

Disposal process. No disposal.

Customer service. Service is considered in terms of backorder costs.

Control strategy: the system is controlled by an (s_p, Q_p) strategy. Whenever the inventory position equals s_p, an outside procurement order of size Q_p is placed. Returned products are remanufactured as soon as possible.

Muckstadt and Isaac present an approximation procedure to determine the control parameters s_p and Q_p, such that the sum of fixed outside procurement costs, inventory holding costs, and backordering costs is minimized. Their procedure is based on the fact that, with exponentially distributed demand and return inter-arrival times, the steady-state distribution for the inventory position can be computed exactly, by solving a continuous time Markov-chain model.

From the steady-state distribution of the *inventory position* an approximation on the distribution of the *net inventory* is derived. In this approximation the net inventory is assumed to behave as a Normal distribution

function, with mean and variance based on an approximation of the first two moments of the steady-state distribution of the net inventory. From the normal approximation of the net inventory expressions on the expected on-hand inventory position and on the expected backordering position are then derived. Using these expressions an approximation on the long run average costs per unit of time as function of the policy parameters s_p and Q_p is obtained. This cost function is then minimized, resulting in an expression for s_p (in closed-form) and an algorithm to determine Q_p numerically.

In the second part of their paper Muckstadt and Isaac consider a two echelon warehouse-retailer model, with an (s_p, Q_p) re-order policy for the warehouse, and an $(S^{(j)} - 1, S^{(j)})$ re-order level policy for the retailers. Here, $S^{(j)}$ is defined as the order up-to level for retailer j. Based on the results for the single-echelon case, an approximation procedure is developed to determine values for the policy parameters in the two echelon case, such that long-run average costs per unit of time are minimized.

An alternative approximation procedure for the same single-echelon model as formulated by Muckstadt and Isaac, was proposed by Van der Laan (1993). The main difference between the Muckstadt and Isaac procedure and the Van der Laan procedure lies in the nature of the approximation. In the latter procedure an approximation is used on the distribution of the net demand during the procurement lead-time, instead of an approximation on the distribution of net inventory.

A numerical comparison in Van der Laan (1993) has shown that in many cases this approach results in a more accurate approximation of the expected number of backorders, and hence in a better (lower cost) choice of the policy parameters s_p and Q_p. Furthermore, an extension of the single-echelon Muckstadt and Isaac model is given, in which customer service is considered in terms of a service level constraint, instead of backordering costs.

In Van der Laan (1993) and Van der Laan et al. (1996) two models are formulated in which remanufacturing and disposal decisions are considered simultaneously. The *first* model, proposed in Van der Laan et al. differs from the single echelon model proposed by Muckstadt and Isaac with respect to the following:

Inventory process. Type I (work-in-process) inventory capacity is limited to N; Type II inventory has infinite capacity; inventory holding costs are considered for Type II inventory only,

Disposal process. Variable disposal costs are considered,

Control strategy. The system is controlled by an (s_p, Q_p, N) strategy. This strategy is defined as follows: whenever the inventory position equals s_p, an outside procurement order of size Q_p is placed; whenever the number of products in Type I inventory equals N, every incoming remanufacturable product is disposed off before having entered the remanufacturing facility.

In Van der Laan et al. (1996) an approximation procedure is described to determine the policy parameters s_p, Q_p, and N simultaneously. The procedure is an extension of the approximation procedure in Van der Laan (1993) for the (s_p, Q_p) model. The *second* model, proposed in Van der Laan (1993), differs from the first model in that the system is controlled by an (s_p, Q_p, s_d) policy. With this policy, the disposal decision is based on the number of products in *inventory position*, rather than on the number of products in Type I inventory. The complete policy is as follows: whenever the inventory position equals s_p, an outside procurement order of size Q_p is placed; whenever the inventory position equals s_d, each additional incoming remanufacturable product is disposed off.

3.3 Cash balancing models

Alternative models that could serve as a starting point for remanufacturing models stem from finance: the cash-balancing models. The reason why we only briefly discuss these models here, is that many characteristics that are typical for a remanufacturing and production environment, like detailed modelling of the remanufacturing process itself, and non-zero lead-times for remanufacturing and production, are disregarded. Nevertheless, some of these models match our selection criteria, and may very well serve as a starting point from which models for remanufacturing and production could be developed further.

Cash balancing models usually consider a *local* cash of a bank with incoming money flows stemming from customer deposits, and outgoing money flows, stemming from customer withdrawals. The possibility exists to increase the cash-level of the local cash by ordering money from the central cash, or to decrease the cash level of the local cash by transferring money to the central cash. Objective in these models is to determine the time and quantity of the cash transactions, such that the sum of fixed and variable transaction costs, backlogging costs, and interest costs related to the local cash is minimized. There exist continuous review and periodic review cash-balancing models. An interesting result is, that for the continuous review model a four parameter (s_p, s_d, S_p, S_d) strategy is optimal in case that no remanufacturing or procurement lead-times exist (see Constantinides (1976), Constantinides and Richard (1978)).

The optimal strategy is as follows: if the inventory level at the local cash becomes less than s_p, an order is placed at the central cash to increase the local cash level to S_p. If the local cash level becomes higher than s_d, the local cash level is reduced to S_d by transferring money to the central cash. Note that according to our notation $Q_p = S_p - s_p$ and $Q_d = s_d - S_d$ if the demand and return quantities are always equal to one unit per transaction. An extensive overview of cash balancing models is given by Inderfurth (1982).

4 An analysis of product remanufacturing

Purpose of this section is to analyse some of effects caused by remanufacturing in case that no disposal takes place. It is assumed that all returned products pass the testing process, and no products that pass the test will be disposed off as part of the control policy. Our analysis is based on the steady-state behaviour of a production and inventory system under a continuous review (s_p, Q_p) policy. We have limited our analysis to this category of models because the time-independent steady-state system behaviour of continuous review models is easier to describe and to understand than the time-dependent system behaviour of periodic review models. Furthermore, in continuous review models remanufacturing and production/procurement lead-times are, in contrast to periodic review models, considered explicitly. Also, the (s_p, Q_p) policy closely resembles the control policies that are nowadays widely accepted and used in practice to control inventory systems in situations without remanufacturing.

The assumptions that we make here are the same as the assumptions made by Muckstadt and Isaac (1981) and by Van der Laan et al. (1996b) except that the remanufacturing facility is modeled as a standard queueing model with w servers, Coxian-2 (see Appendix A for a formal definition) or exponentially distributed inter-occurrence times with mean $\frac{1}{\gamma}$, and exponentially distributed service times with mean $\frac{1}{\mu}$. The unit demand is assumed to be a Poisson process with mean inter-occurence time $\frac{1}{\lambda}$. The fixed manufacturing leadtime equals τ. Objective is to determine the policy parameters s_p and Q_p, such that the long-run average costs per unit of time, defined as,

$$C(s_p, Q_p) = A_p E(P) + c_h E(O) + c_b E(B) \tag{1}$$

are minimized. In (1) the following notation is used:

$E(P)$	=	average number of outside procurement orders per unit of time,
$E(O)$	=	average number of products in on-hand Type II inventory per unit of time,
$E(B)$	=	average number of products in backordering per unit of time,
A_p	=	fixed ordering costs of a manufacturing batch,
c_h	=	inventory holding costs of serviceable products per product per time unit,
c_b	=	backorder costs per product per time unit.

To determine the parameters s_p and Q_p that minimize (1), we could have applied the approximation procedures suggested by Muckstadt and Isaac (1981), or by Van der Laan et al. (1996b). However, as we are interested in the *true* system behaviour as function of the input data, we have developed

an exact procedure to calculate the costs. In addition to the above notation, we use in our procedure the following notation: $I(t)$ is the inventory position of Type I and Type II inventory at time t, $N(t)$ is the net Type II inventory at time t, $R(t)$ is the number of products in the remanufacturing shop (Type I inventory plus the number of products in the remanufacturing facility) at time t, and $Z(t-\tau,t)$ is the *net demand* in the interval $(t-\tau,t]$. Here, the net demand is defined as the difference between the customer demand quantity in the interval $(t-\tau,t]$ and the number of products that leave the remanufacturing shop after repair in the interval $(t-\tau,t]$.

By definition, the net inventory at time t equals:

$$N(t) = I(t-\tau) - R(t-\tau) - Z(t-\tau,t) \tag{2}$$

Using (2) we compute the probability distribution of net inventory, i.e.,

$$\Pr\{N(t)=n\} = \sum_{\Omega} \Pr\{I(t-\tau)=i, R(t-\tau)=r, Z(t-\tau,t)=d\}, \tag{3}$$

where

$$\Omega = \{(i,r,z)|i-r-z=n\}.$$

It should be noted that the inventory position, the number of products in the remanufacturing shop, and the net demand during the procurement lead-time are not mutually independent. In order to facilitate our analysis, we rewrite (3) in terms of conditional probabilities:

$$\Pr\{N(t)=n\} = \sum_{\Omega} \Pr\{Z(t-\tau,t)=z|I(t-\tau)=i, R(t-\tau)=r\} \times \Pr\{I(t-\tau)=i, R(t-\tau)=r\}$$

To compute the limiting joint probability $\lim_{t\to\infty} \Pr\{I(t-\tau)=i, R(t-\tau)=r\}$ we observe that the inter-occurrence times of returns are Coxian-2 or exponentially distributed, and the repair times are exponentially distributed. Since in the resulting continuous-time Markov chain every state can be reached from every other state the chain is ergodic provided that $\gamma < min(\lambda, w\mu)$. Hence the steady-state probabilities $\pi_{i,r}$, which follow from solving the continuous-time Markov chain, are equal to $\lim_{t\to\infty} \Pr\{I(t-\tau)=i, R(t-\tau)=r\}$. For the case of *exponentially* distributed demand and return inter-occurrence times the state-space in the Markov-chain model is defined as $\mathcal{S} = \{(i,r)|i \geq s_p+1, r \geq 0\}$. Each state in the state space corresponds to the inventory position, and the number of products in the remanufacturing

shop. Given the state-space definition, the non-zero transition rates ν_{s_0,s_1} related to a transition from state $s_0 \in \mathcal{S}$ to state $s_1 \in \mathcal{S}$, are as follows.

$$\begin{aligned}
&\nu_{(i,r),(i,r-1)} = r\mu, && i \geq s_p + 1, \quad 0 < r \leq w,\\
&\nu_{(i,r),(i,r-1)} = w\mu, && i \geq s_p + 1, \quad r > w,\\
&\nu_{(i,r),(i-1,r)} = \lambda, && i > s_p + 1, \quad r \geq 0,\\
&\nu_{(s_p+Q_p,r),(s_p+1,r)} = \lambda, && r \geq 0,\\
&\nu_{(i,r),(i+1,r+1)} = \gamma, && i \geq s_p + 1, \quad r \geq 0.
\end{aligned}$$

We obtain the steady-state probabilities $\pi_{i,r}$ numerically, applying the well-known Gauss-Seidel procedure.

Remark: if demand and/or return inter-occurrence times are Coxian-2 distributed instead of exponentially distributed, the steady-state probabilities can be obtained using an extension of the continuous time Markov-chain model presented here (see Appendix B). In our computational experiments we have not considered Coxian-2 distributed demand inter-occurrence times.

Next, we calculate the conditional probability $\Pr\{Z(t-\tau,t) = z | I(t-\tau) = i, R(t-\tau) = r\}$. Note that the net demand $Z(t-\tau,t)$ depends on the number of products in the repair shop at time $t-\tau$, but does *not* depend on the inventory position $I(t-\tau)$, i.e., $\Pr\{Z(t-\tau,t) = z | I(t-\tau) = i, R(t-\tau) = r\} = \Pr\{Z(t-\tau,t) = z | R(t-\tau) = r\}$. The demand $D(t-\tau,t)$ in the interval $(t-\tau,t]$ is *not* dependent of the number of products in the repair shop at time $t-\tau$. Given that demand inter-occurrence times are exponentially distributed, the distribution of demand in the interval $(t-\tau,t]$ is given by

$$\Pr\{D(t-\tau,t) = d\} = \frac{\exp^{-\lambda\tau}(\lambda\tau)^d}{d!}$$

Using the distribution of demand, we derive the following expression on the distribution of net demand, conditioned on the inventory position and the number of products in the remanufacturing shop,

$$\begin{aligned}
&\Pr\{Z(t-\tau,t) = z | R(t-\tau) = r\} =\\
&\qquad \sum_{d=\max(0,z)}^{\infty} \Pr\{R^{out}(t-\tau,t) = d - z | R(t-\tau) = r\}\\
&\qquad\qquad \times P\{D(t-\tau,t) = d\}
\end{aligned}$$

where $R^{out}(t-\tau,t)$ is defined as the output of the remanufacturing shop in the interval $(t-\tau,t]$. Given exponentially distributed inter-occurrence times of demands, and exponentially distributed remanufacturing times, the

conditional probabilities $\Pr\{R^{out}(t-\tau,t) = d - z | R(t-\tau) = r\}$ can be computed for an arbitrary number of parallel remanufacturing facilities, using a uniformization technique to evaluate the transient behaviour of a Markov-chain model. The Markov-chain model and the uniformization technique will be discussed in Appendix C.

From the above we have

$$\begin{aligned} E(O) &= \lim_{t\to\infty} \sum_{n\geq 0} n \Pr\{N(t) = n\}, \\ E(B) &= -\lim_{t\to\infty} \sum_{n<0} n \Pr\{N(t) = n\}, \\ E(P) &= \frac{\lambda-\gamma}{Q_p}. \end{aligned}$$

Using these expectations we can numerically evaluate (1) for *fixed* values of s_p and Q_p. To determine the *optimal* values for the control parameters s_p and Q_p, we have implemented an *enumerative search* procedure. In the enumerative search procedure $C(s_p, Q_p)$ is evaluated for a large number of relevant (s_p, Q_p) combinations. In order to speed up our search procedure we use the result that for fixed Q_p the cost function (1) is convex in s_p (see Van der Laan (1993)).

In the remainder of this section we analyse the performance of the reorder point strategy with remanufacturing. In our analysis we focus on cost effects, and on the effects of remanufacturing on the choice of the optimal values for the control parameters. In setting up our analysis we have first created a *'base-case'* scenario (Table 4.1).

Table 4.1. *Parameter settings in the base-case scenario.*

$\frac{1}{\lambda}$	=	1.0
cv_D^2	=	1.0 (exponential)
$\frac{1}{\gamma}$	=	1.0
cv_R^2	=	1.0 (exponential)
$\frac{1}{\mu}$	=	0.5
w	=	∞
τ	=	10.0
A_p	=	10.0
c_h	=	1.0
c_b	=	100.0

In the other scenario's that we evaluate we have varied one or multiple of the input parameters from the 'base-case' scenario.

SCENARIO I: This scenario has been developed in order to evaluate the influence of changes in *(i)* the *magnitude* of the returns (i.e., the return rate γ)

and *(ii)* changes in the *uncertainty* of returns (i.e., the squared coefficient of variation cv_R^2) on costs (Figure 4.1a) and on the choice of the control parameters (Figure 4.1b–c). With respect to the analysis of uncertainty in returns we have modelled the return inter-occurrence times by a Coxian-2 distribution function, where the input parameters are chosen such that the squared coefficients of variation are $cv_R^2 = \frac{1}{2}, 1$, or 2,

SCENARIO II. In this scenario we have varied the backordering costs (c_b) and the variability in the return flow (cv_R^2) in order to investigate the influence of service and uncertainty on total expected costs (Figure 3a) and on the value of the control parameters (Figure 4.2b–c),

SCENARIO III. In this scenario we have varied the outside procurement lead-times (τ) in order to investigate the influence of these lead-times on total expected costs (Figure 4.3a) and on the value of the control parameters (Figure 4.3b–c),

SCENARIO IV. In this scenario we have varied the remanufacturing rate (μ) and the capacity of the remanufacturing shop (w) in order to investigate the effects of these two components on the total expected costs (Figure 4.4a) and on the control parameters (Figure 4.4b–c),

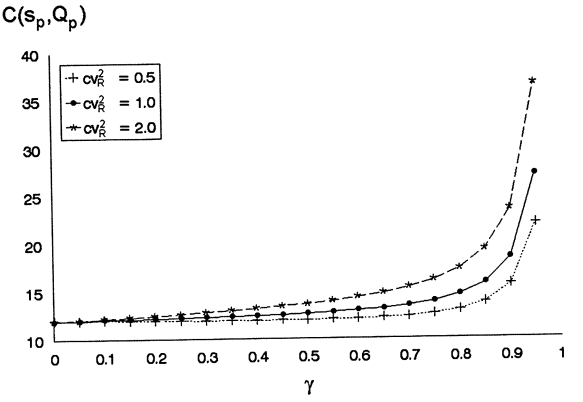
From the above experiments the following conclusions can be drawn:

In the base case scenario variable remanufacturing costs are lower than fixed and variable outside procurement costs. In this situation, remanufacturing of products may be worthwhile to consider.
Figure 4.1a indeed shows that, up to a certain *'critical'* level of the return rate ($\gamma \approx 0.7$ in this example), the minimal total expected costs $C(s_p, Q_p)$ very slowly changes when γ increases. However, beyond the critical level of the return rate, minimal total expected costs start to increase rapidly. The cost increase is mainly caused by the inventory holding costs, of which the effect now becomes dominant over the effects caused by the other cost components. Convexity of the minimal total expected costs in γ as shown in Figure 4.1a has been observed in many experiments and seems to be typical for the situation with remanufacturing. However, whether convexity is a structural property of the cost function is still an open research question. Convexity of the cost function is further interesting, because it suggests that disposal is an economically favourable option.

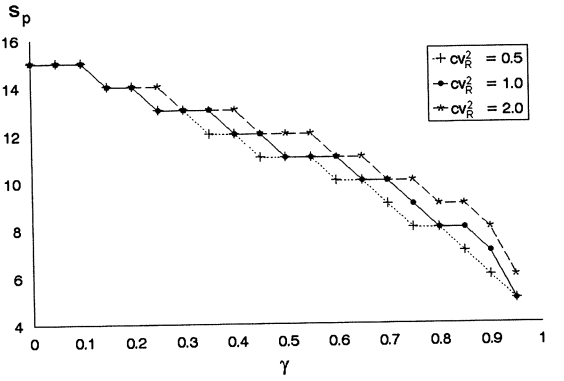
Figure 4.1a further shows that an increase in the variability of the return flow causes an increase in the minimal total costs. This effect has also been observed in many experiments, and becomes more dominant when the magnitude of the return flow increases.

Figure 4.1b shows that the re-order level s_p monotonously decreases when the return-rate γ increases. Furthermore, the higher the variability in the re-

(a)



(b)



(c)

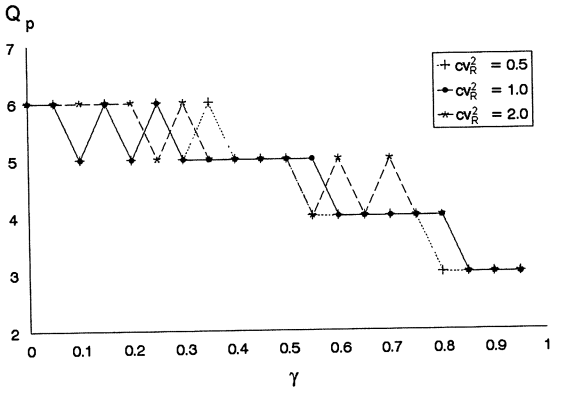


Figure 4.1a–c.

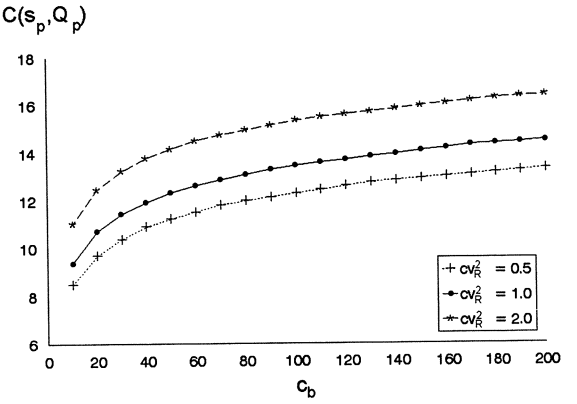
turns, the higher the re-order level. Figure 4.1c shows the somewhat counter intuitive effect that the re-order quantity Q_p *not* always decreases when the return rate γ increases. Also, a higher variability in the returns need not always to correspond to a higher re-order quantity. However, in many experiments we observed an interesting interaction between the re-order quantity and the re-order level, which may explain the above effect: when the return rate increased, an increase in the re-order quantity *always* corresponded to a decrease in the re-order level.

From Figure 4.2a it is observed that the minimal total expected costs behave as a concave function in the backordering costs. Furthermore, higher variability in returns causes larger effects on total costs if the backordering costs are increased. However, the re-order quantity seems to be rather insensitive for an increase in backordering costs, whereas the re-order level increases almost concave when backordering costs are increased (Figures 4.2b–c). An explanation for this effect is, that it is very likely that the distribution of the net-inventory has an exponentially decreasing lower tail. Consequently, the quantity that needs to be backordered decreases convex as function of the re-order level. Thus, an increase in the backordering costs needs to be less than proportionally compensated by an increase in the re-order level. Combining this with the observation that the re-order level has an almost linear effect on inventory holding costs may explain the behaviour of the minimal total cost function.

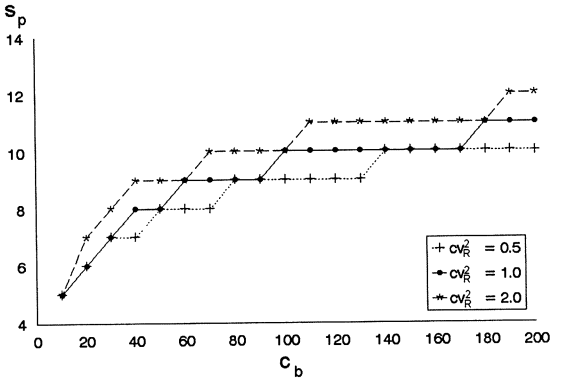
A well-known approximation for the (s_p, Q_p) model without remanufacturing is, to set the re-order level equal to the sum of the expected demand during the procurement lead-time plus a safety factor times the standard deviation of the demand during the procurement lead-time, and to set the re-order quantity equal to the Economic Order Quantity (EOQ) (see Silver and Peterson, 1985). According to this approximation, the expected demand during the procurement lead-time increases *linearly* in the procurement lead-time and as a *square root* in the standard deviation of demand during the procurement lead-time. This could explain the somewhat concave relationship between the procurement lead-time and re-order level (Figure 4.3b), and, derived from it, the somewhat concave relationship between the procurement lead-time and the minimal total expected costs (Figure 4.3a). As in the case without remanufacturing, the procurement quantity Q_p is hardly effected by changes in the procurement lead-time (Figure 4.3c).

Figure 4.4a shows that when μ becomes larger (here, $\mu \approx 0.55$) *fewer* remanufacturing facilities result in *lower* minimal total expected costs. This observation is clearly counter intuitive. However, the effect occurs mainly due to a model limitation: in (1) we consider inventory holding costs only for Type II inventory What actually happens when fewer remanufacturing facilities are present, is, that Type I inventory becomes relatively large, whereas

(a)



(b)



(c)

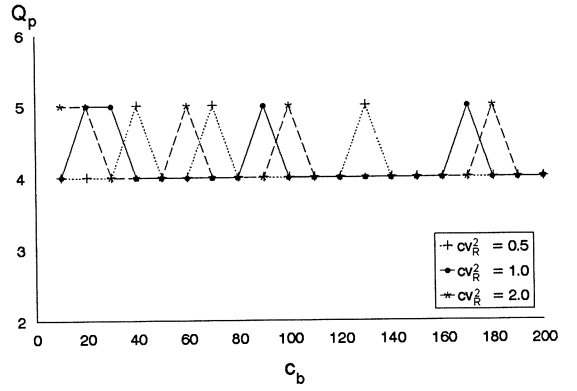
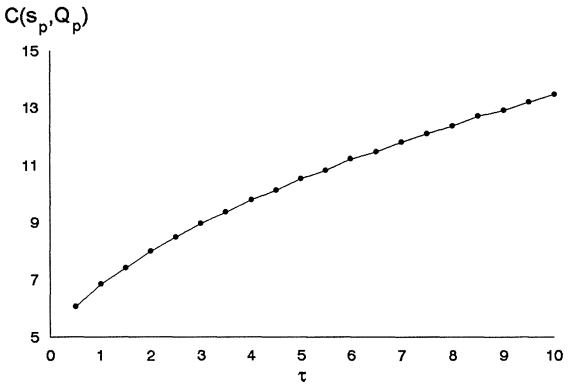
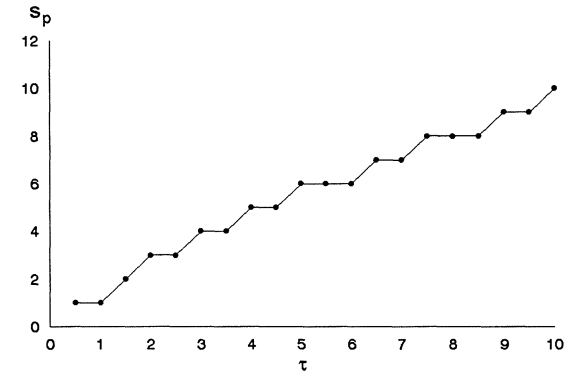


Figure 4.2a–c.

(a)



(b)



(c)

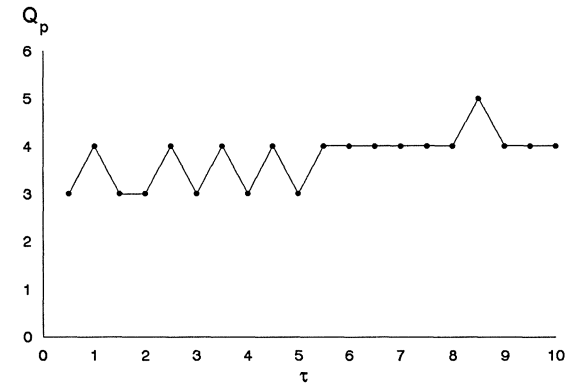
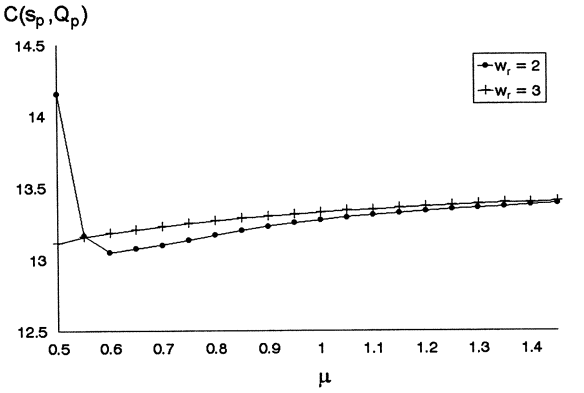
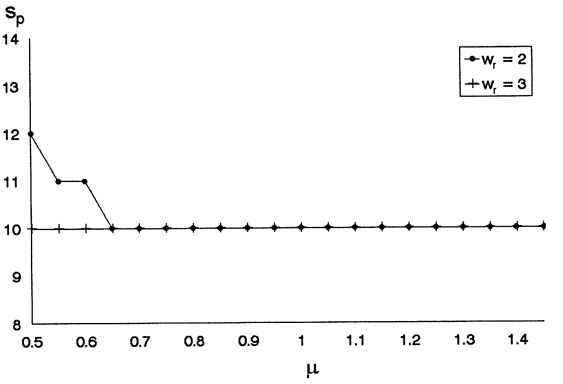


Figure 4.3a–c.

(a)



(b)



(c)

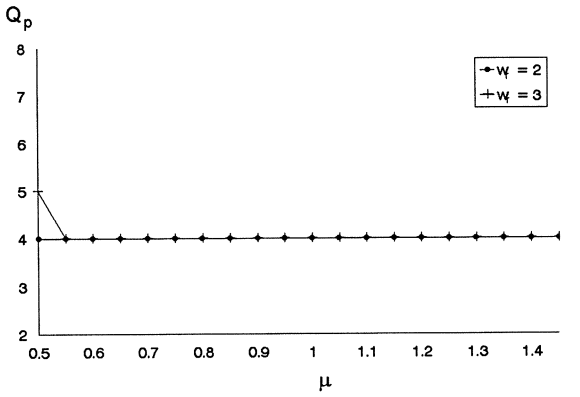


Figure 4.4a–c.

Type II inventory remains relatively small. Consequently, total inventory holding costs related to Type II inventory may very well be smaller with fewer remanufacturing facilities. However, when μ becomes too small, remanufacturing capacity becomes insufficient when fewer remanufacturing facilities are present. This capacity problem causes long waiting times in the queue in front of the remanufacturing shop, resulting in an increased number of backlogging and outside procurement occurrences, and a strong increase in total costs. Figures 4.4b–c further show that the re-order level as well as the re-order quantity appear to be rather insensitive for changes in both μ and w.

Another limitation which makes it difficult to compare total costs between two situations in which a different number of remanufacturing facilities are present is, that no fixed costs related to the remanufacturing facilities itself are taken into account.

It should be noted that the above conclusions are based on a limited number of instances (scenario's). Although we observed similar effects when evaluating alternative scenario's, we do not claim that exactly the same effects will *always* occur in presence of remanufacturing. The examples are given to create awareness, and to explain the effects that *might occur* in practice.

5 Conclusions and directions for further research

In this paper we have investigated the processes that are relevant in inventory control with remanufacturing. From the literature on mathematical models for inventory control we made a selection of models that seem applicable in the context of remanufacturing. However, even for this selected set of models we found that a number of relevant processes are not modelled, or modelled to a limited extent only. Some of the missing model components are listed below:

Multi-level product structures (i.e., multiple products or products that consist of more than one component). All models assume a single level product structure (i.e. one product that consists of a single component only). Therefore, disassembly of returned products and assembly of new products is not considered. Consequently, the applicability of these models in an MRP environment is only limited. Research on how to determine appropriate lot-sizes and safety-stock levels in a multi-level environment with remanufacturing seems to be very useful (see also Brayman (1992)),

Dependency of demand and return processes. In the models considered, the model of Simpson (1978) is the only one in which market demands and mar-

ket returns are not either completely independent, or completely dependent. In most practical situations, some demands and returns occur because of product replacements (in which case there is perfect correlation between returns and demands), whereas the other returns and demands are not caused by replacements (in which case no correlation exists between returns and demands). In the light of this, it seems interesting to further model and investigate the (economic) effects of mixes between correlated and uncorrelated demand and return occurrences,

Demand and return quantities. All continuous review models, except Heyman (1977), consider return and demand quantities to be equal to one product per occurrence. Any more general assumption would complicate model analysis considerably. However, as unit demands and unit returns do not always occur in practice, further research on models in which this assumption is relaxed seems worthwhile,

Lead-times. In the periodic review and in the cash-balancing models remanufacturing lead-times and procurement lead-times are not considered. As these lead-times might be significant in practice, the periodic-review and the cash-balancing models are in some settings of little practical value. Extensions of periodic review and cash-balancing models to include lead-times is therefore relevant,

Interactions between remanufacturing capacity and production capacity. Most models consider outside procurement instead of internal production. Therefore, the capacity interactions between remanufacturing and production are disregarded. However, these capacity interactions may be significant in manufacturing companies that produce internally. Consequently, any research to analyse these capacity effects would be very interesting,

Inventories. In continuous review models all products that pass the test are *immediately* remanufactured. Disposals may occur before testing (i.e., as part of the control strategy), or after testing, for products with a negative test outcome. Consequently, in continuous review models Type I inventory occurs only because of products waiting for remanufacturing (work-in-process). Furthermore, inventory holding costs for Type I inventory are disregarded. In periodic review models the situation is different: not all returned products are immediately remanufactured. In this way, Type I inventory can be used to create additional flexibility: some products will be disposed off, whereas others will be remanufactured. Decisions on disposal and remanufacturing are part of the control strategy. Furthermore, inventory holding costs for Type I inventory are explicitly taken into account. Extensions of the continuous review models to use Type I inventory in the same way as it is used in the periodic review models (i.e., to allow for more flexibility) seems worthwhile to investigate,

Service. All models, except the periodic review model of Kelle and Silver (1989) and the continuous review model by Van der Laan (1993), consider service in terms of backlogging costs. However, in practice backlogging costs are difficult to specify. Modelling service in terms of a, usually easier to specify, service level seems therefore a fruitful research topic,

Control policy. Simpson (1978) has carried out research to identify an optimal control strategy for a periodic review model, under the assumption of zero remanufacturing and procurement lead-times, and zero fixed procurement costs. Heyman (1977) has proved optimality of a continuous review single parameter disposal strategy, also under the assumptions of zero remanufacturing and procurement lead-times, and zero fixed procurement costs. Inderfurth (1996) investigated the structure of the optimal policy for a more general model, but still any research in this direction would be very useful,

Disposal costs. Except in cash-balancing models, no other models consider fixed disposal costs, and related to this, no other models consider the problem of determining the disposal batch size.

After the literature review we have investigated some of the effects that remanufacturing may cause on inventory control. First, we analysed the effects if the system is controlled by a continuous review (s_p, Q_p) strategy. In order to analyse the effects accurately, we derived an exact method to evaluate the costs. The cost evaluation method was then integrated with an enumerative search procedure to find the optimal values for the control parameters.

One of the most interesting observations that we made during the analysis of the (s_p, Q_p) operating strategy is, that the cost function seems to behave convex in the return-rate γ. From this, we may conclude that remanufacturing is not only from an environmental point of view, but also from an economical point of view an option worthwhile to consider. Furthermore, convexity of the cost function also suggests that remanufacturing should be applied in combination with disposal, especially when the product return rate becomes higher.

This paper is meant as a first attempt to structure the literature on remanufacturing, and to obtain some insight into the effects that remanufacturing may cause on production planning and inventory control. However, as may be clear from the discussion above, a lot of research remains to be done in this new and challenging field.

References

Brayman, R.B. (1992): How to implement MRP II successfully the second time: getting people involved in a remanufacturing environment. In: APICS Remanufacturing Seminar Proceedings, September 23–25, 82–88.

Cho, D.I. / Parlar, M. (1991): A survey of maintenance models for multi-unit systems. In: European Journal of Operational Research, 51:1–23.

Constantinides, G.M. (1976): Stochastic cash management with fixed and proportional transaction costs. In: Management Science, 22(12):1320–1331.

Constantinides, G.M. / Richard, S.F. (1978): Existence of optimal simple policies for discounted-cost inventory and cash management in continuous time. In: Operations Research, 26(4):620–636.

Heyman, D.P. (1977): Optimal disposal policies for a single-item inventory system with returns. In: Naval Research Logistics Quarterly, 24:385–405.

Hoadley, B. / Heyman, D.P. (1977): A two-echelon inventory system with purchases, disposition, shipments, returns, and transhipments. In: Naval Research Logistics Quarterly, 24:1–19.

Inderfurth, K. (1982): Zum Stand der betriebswirtschaftlichen Kassenhaltungstheorie. In: Zeitschrift fuer Betriebswirtschaft, 3:295–320 (in German).

Inderfurth, K. (1996): Simple Optimal Replenishment and Disposal Policies for a Product Recovery System with Leadtimes. Preprint Nr. 7, Fakultät für Wirtschaftswissenschaft, Otto-von-Guericke-Universität Magdeburg, Germany.

Kandebo, S.W. (1990): Grumman, U.S. Navy under way in F-14 remanufacturing program. In: Aviation Week & Space Technology, December 44–45.

Kelle, P / Silver, E.A. (1989): Purchasing policy of new containers considering the random returns of previously issued containers. In: IIE Transactions, 21(4):349–354.

Muckstadt, J.A. / Isaac, M.H. (1981): An analysis of single item inventory systems with returns. In: Naval Research Logistics Quarterly, 28:237–254.

Nahmias, S. (1981): Managing repairable item inventory systems: a review. In: TIMS Studies in the Management Sciences, 16:253–277. North-Holland Publishing Company, The Netherlands.

Sherbrooke, C.C. (1968): METRIC: a multi-echelon technique for recoverable item control. In: Operations Research, 16:122–141.

Silver, E.A. / Peterson, R. (1985): Decision Systems for Inventory Management and Production Planning. (Wiley & Sons) New York.

Simpson, V.P. (1978): Optimum solution structure for a repairable inventory problem. In: Operations Research, 26:270–281.

Sivinski, J.A. / Meegan, S. (1993): Case study: Abbott labs formalized approach to remanufacturing. In: APICS Remanufacturing Seminar Proceedings, May 24–26, 27–30.

Sprow, E. (1992): The mechanics of remanufacture. In: Manufacturing Engineering, March 38–45.

Thierry, M.C. / Salomon, M. / Van Nunen, J.A.E.E. / Van Wassenhove, L.N. (1995): Strategic production and operations management issues in product recovery management. In: California Management Review, 37(2):114–135.

Tijms, H.C. (1986): Stochastic Modelling and Analysis: A Computational Approach. (Wiley & Sons) Chichester.

Van der Laan, E.A. (1993): On Inventory Control Models where Items are Remanufactured or Disposed. Unpublished Master's Thesis. Erasmus Universiteit Rotterdam, The Netherlands.

Van der Laan, E.A. / Dekker, R. / Salomon, M. / Ridder, A. (1996): An (s,Q) inventory model with remanufacturing and disposal. In: International Journal of Production Economics, 46–47:339–350.

Vandermerwe, S. / Oliff, M.D. (1991): Corporate challenges for an age of reconsumption. In: Columbia Journal of World Business, Fall vol., 7–25.

Appendix A: Definition of the Coxian-2 distribution

A random variable (r.v.) X has a Coxian-2 distribution (also called K_2) if,

$$X = \begin{cases} X_1 & \text{with probability } q \\ X_1 + X_2 & \text{with probability } 1-q \end{cases},$$

where X_1 and X_2 are independent and exponentially distributed r.v.'s with parameters γ_1 and γ_2 respectively and $0 \leq q \leq 1, \gamma_1 > 0, \gamma_2 > 0$.

Notice that the Coxian-2 distribution reduces to an exponential distribution if $q = 1$ and to an Erlang-2 distribution if $q = 0$. Fitting a Coxian-2 distribution on the first two moments leaves some degrees of freedom in the choice of the parameters. We applied a so-called gamma normalization in which the parameters are chosen such that the Coxian-2 distribution has the same third moment as a gamma distribution with the same first two moments. Such a fit is always possible (provided that the squared coefficient of variation, cv_X^2,

is larger than $\frac{1}{2}$ (see Tijms 1986, p.399-400). The values of γ_1, γ_2 and q are given by

$$\begin{aligned}
\gamma_1 &= \tfrac{2}{E(X)}\left(1+\sqrt{\left(\tfrac{cv_X^2-\frac{1}{2}}{cv_X^2+1}\right)}\right), \\
\gamma_2 &= \tfrac{4}{E(X)} - \gamma_1, \\
q &= (1-\gamma_2 E(X)) + \tfrac{\gamma_2}{\gamma_1},
\end{aligned}$$

Appendix B: Calculation of steady-state probabilities in case of Coxian-2 distributions

In this case we add an extra component to the state description of the continuous-time Markov chain to indicate the phase of the Coxian-2 distribution. In principle also Erlang distributions can be incorporated in this way. As we are only interested in the effect of a reduced or increased variability and each extra state variable increases the computational effort we will only consider the Coxian-2 distribution for either the demand or the return inter-occurrence times. Below we specify the case for the return process; the case for demand process is analogous. Let the r.v. $A(t)$ denote the phase of the return process at time t and assume that the Coxian-2 distribution has parameters γ_1, γ_2 and q. The state space now becomes $S = \{(i,r,a) | i \geq s_p + 1, r \geq 0, a = 0 \text{ or } 1\}$. Transition rates are identical to those for the corresponding states in the exponential case, except for those related to the returns. The latter are

$$\begin{aligned}
\nu_{(i,r,a),(i+1,r+1,0)} &= q\gamma_1, && i \geq s_p+1, \quad r \geq 0, \\
\nu_{(i,r,a),(i+1,r+1,1)} &= (1-q)\gamma_1, && i \geq s_p+1, \quad r \geq 0, \\
\nu_{(i,r,a),(i+1,r+1,1)} &= \gamma_2, && i \geq s_p+1, \quad r \geq 0.
\end{aligned}$$

The same numerical procedure can be used to determine the steady-state probabilities of the extended Markov chain. Summing over the a-component of the state description now yields the steady state probabilities for I and R.

Appendix C: Calculation of the conditional probability $P(R^{out}(t-\tau,t) = d | R(t-\tau) = r)$

In case the return inter-occurrence times are exponentially distributed and there are w servers the remanufacturing shop can be modelled as an $M/M/w$ queue. Transient analysis of this queue now yields the desired probability for

starting state r. To this end we applied a uniformization technique (see e.g. Tijms (1986)).

In case the return inter-occurrence times are Coxian-2 distributed we have to analyse the transient behaviour of the *Coxian-2*/M/w queue. Again we formulate a continuous-time Markov chain, now with an extra state component to indicate the phase of the Coxian-2 distribution. The same has to be done with the joint Markov chain of the inventory position and the number of products in the remanufacturing shop (see section A.2). For the latter we first calculate the limiting probabilities for the case there are r products in the repair shop. Summing over the inventory position then gives us the stationary probabilities of being in the first or second phase of the Coxian-2 distribution. Returning to the transient behaviour of the remanufacturing shop we now calculate the probability of ending with d items while starting with r items given a certain phase of the Coxian-2 distribution. Weighting these probabilities with the stationary probabilities now gives the desired result.

Forecasting Techniques in Logistics

Sander de Leeuw[1], Karel van Donselaar[2], Ton de Kok[2]

[1] Center for Technology, Policy and Industrial Development, Massachusetts Institute of Technology, Cambridge, MA 02138, USA (sanderl@mit.edu) and: Management Division, Babson College, Babson Park, MA 02157, USA.
[2] School of Technology Management, Eindhoven University of Technology, P.O Box 513, 5600, MB Eindhoven, The Netherlands.

Abstract. This article deals with the selection of distribution control techniques and discusses one element of distribution control, the type of forecasting techniques. It first goes into the detail of a classification of distribution control decisions, the bigger framework underlying this article which can be used to select appropriate distribution control techniques. The type of forecasting technique is one of these distribution control decisions, and this is the topic of discussion for the rest of the article. The results of case study research and literature research on the application of forecasting techniques are described. A simulation model is presented, which is used to research the usefulness of forecasting techniques. As opposed to most research on forecasting, the relation between characteristics of products, processes and markets on different types of forecasting techniques is investigated from a logistics perspective in this simulation. The effect of forecasting techniques is assessed based on the impact the techniques have on the inventory level, not on a forecast accuracy measure. Two techniques are discriminated: techniques that can incorporate demand patterns and techniques that can not incorporate these patterns. It appears that the combination of a seasonal pattern or a trend in demand with demand uncertainty has a significant impact on the choice between these two techniques. Only if demand uncertainty is low and if demand contains clear patterns, forecasting techniques that can incorporate demand patterns outperform those that can not.

1.1 Introduction

Physical distribution control is concerned with all activities needed to co-ordinate the place and timing of demand for and supply of products and capacities in such a way that objectives regarding products, markets and the distribution process are met (De Leeuw, 1996a,b). Several standard techniques for physical distribution control - we will refer to these as distribution control techniques - have been de-

veloped and applied. A classic technique is the reorder point (ROP) technique, also referred to as Statistical Inventory Control (SIC). According to SIC, each warehouse orders a batch - fixed or variable in size - each time a pre-specified inventory level is passed. The level of this so called reorder point is dependent upon variables such as the mean and deviation of the supply lead time and the mean and deviation of the demand rate. In more advanced systems, orders are placed when echelon inventory levels - i.e. the inventory of the warehouse considered plus the inventory in all its downstream warehouses - instead of local (also called installation) stock levels are passed. Base Stock Control is an example of such a technique that uses echelon stock norms (see for example Silver and Peterson (1985)). More recent techniques use replenishments that are not triggered by realisations of customer demand (reactive) but by future demand forecasts (proactive). Again, replenishments can be based on either local stock norms (Distribution Requirements Planning) or on echelon stock norms (Line Requirements Planning (Van Donselaar, 1990)).

Many times, the literature introducing new concepts in the area of distribution as well as production does not discuss application restrictions, as can be seen in the work by Martin (1993) or Orlicky (1975). Theoretical and practical evidence, however, shows that the application of these standard control techniques is not always equally successful (Van Donselaar, 1989; Masters et al., 1992). Only a few attempts have been made so far to assess the applicability of distribution control techniques (for an example, see Masters et al. (1992). Moreover, it is hardly discussed in literature how a framework for physical distribution control should be set up. It was therefore judged necessary to enlarge the knowledge on the successful application of physical distribution control techniques. In the next section, we will first discuss a classification of distribution control decision that encompasses the essential characteristics of a distribution control technique. After that, we will focus on one of these decisions - the type of forecasting technique.

1.2 Approach and classification

1.2.1 Contingency characteristics

This paper is the result of a study which had the objective to draw conclusions on the relationship between on the one hand company and environmental characteristics and on the other the selection of distribution control techniques. This approach is often referred to as a *"contingency"* approach. Contingency theory has been widely applied in management science and states that a system will only be effective if there is a balance between this system and its relevant environment. From a systems theory point of view, characteristics of the input of the distribution system and the requirements imposed on the output determine the design of a

distribution control framework. The parameters we use in this respect are characteristics of the processes used, the products distributed and the markets served (see Grunwald & van der Linden (1980) and Hoekstra and Romme (1993)). This is graphically represented in Figure 1.

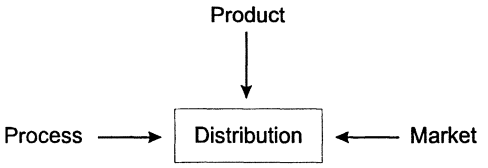


Figure 1. Characteristics of processes, products and markets influence distribution control design

The investigation of the relation between process, product and market characteristics on the one hand and the control technique on the other is relatively complex. For this reason, distribution control techniques are first classified according to their typical characteristics. The next section further discusses classifications of control techniques.

1.2.2 Classifications for distribution control

Several classifications of distribution control techniques have been presented in the past, with different purposes. A frequently used discriminatory parameter to explain differences between distribution control techniques is "push" versus "pull" (see Brown (1967), Christopher (1985)). However, there is a wide variety of associations commonly linked to the terms "push" and "pull" (Pyke & Cohen, 1990). Silver (1981) gives a classification of inventory models, which has been elaborated by Prasad (1994). Both classifications, however, are set up to classify inventory control theory by means of theoretically oriented aspects such as the type of demand processes assumed or the stock out policy. They only consider the theoretical capabilities of mathematical inventory control models and are therefore omitted from the discussion.

More practically oriented classifications are for example those by Jenniskens (1986) and Rosenfield and Pendrock (1980). Jenniskens (1986) discriminates between integral and local stock norms and between reactive and proactive planning. Rosenfield and Pendrock (1980) use a differentiation between coupled and independent systems with centralised or decentralised control.

The classifications of Jenniskens (1986) and Rosenfield and Pendrock (1980) are oriented at a specific part of distribution control only. The classification of Jenniskens incorporates the question which type of data should be used for the reorder calculation but does not give an explanation about who initiates a reorder

in a distribution system. Rosenfield and Pendrock incorporate the initiation of reorders but do not discuss for example different methods for calculating a reorder.

1.3 Classification of control decisions

The classification used in this paper is a mixture of the classifications discussed above. It contains four elements, which are represented in Table 1[1] and discussed in Sections 1.3.1 to 1.3.4. The four elements of distribution control presented in Table 1 are referred to as the distribution control decisions. Each control technique can be expressed in terms of these control decisions. Statistical Inventory Control, for example, is a technique which is characterised by non time phased reorder planning, local status information, central stock is assumed and the allocation is coordinated locally. Characteristics of products, processes and markets are used to determine the outcome of the control decisions, which in turn is used for the selection of a distribution control technique.

Table 1. Classification of distribution control

Control decision	Deals with...
Type of reorder planning	The ability to incorporate a time phased pattern in the planning of the independent demand and the dependent demand
Status information	The use of local or integral information (for example about inventory) for reorder purposes
Central stock function	Having both central stock and local stock or only local stock
Allocation co-ordination	Having a centrally or a locally (i.e., in the local Distribution Centre - abbreviated DC) co-ordinated allocation

1.3.1 Type of reorder planning

The type of reorder planning is discriminated into two parts: the planning of the independent and of the dependent demand. The independent demand concerns the demand of the independent customer, the dependent demand consists of the product requirements from the local Distribution Centres (DCs) as faced by the central DC.

The independent demand planning deals with the question which technique should be used to forecast the demand. A difference is made between two types of

[1] This classification is actually a result of literature research and of case study research

techniques. The first type can incorporate patterns in demand and is referred to as a *forecasting technique which can incorporate a demand pattern.* The forecast $F_{t,t+i}$, which is the forecast made at time t for period $t+i$, may be different for each i at a specific moment t. The second type is a technique that can not incorporate patterns, called a *forecasting technique which can not incorporate a demand pattern.* As a result, all $F_{t,t+i}$ are the same for all values of i at a specific moment t.

Regarding the method to deal with dependent demand from the local DCs, there are two possibilities. In a time phased planning technique, it is attempted to predict the moment on which a new order is generated by the lower echelon. The order is planned in such a way that the stock is available only just before it is needed (see Figure 2). The other possibility is to discard the pattern of reorders from the local DCs over time and to replenish up to a specific level based on the reorders of the local Distribution Centres. The effect on central stock levels for time phased and non time phased planning is shown in Figure 2.

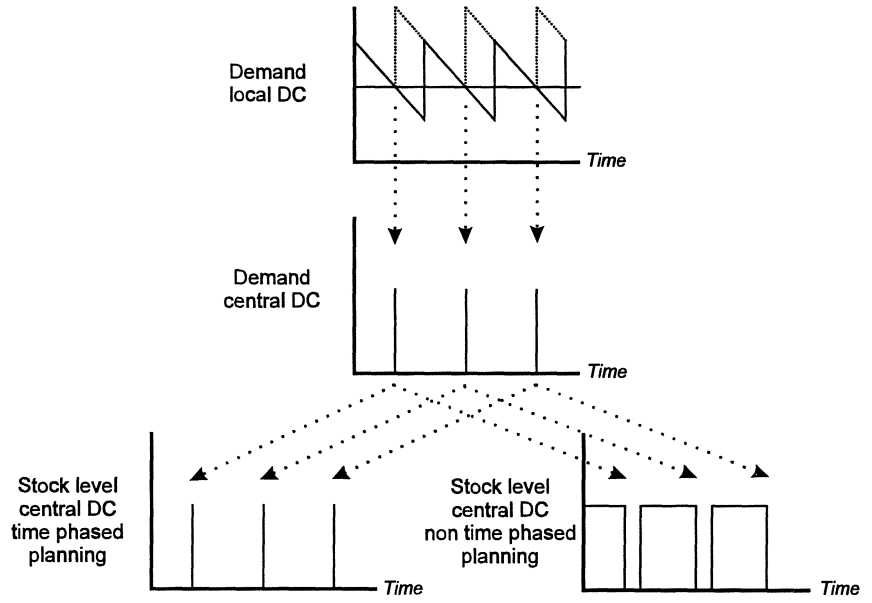


Figure 2. Difference between time phased and non time phased planning

1.3.2 Status information

Status information is information about demand and stock levels in the distribution system. This applies to the replenishment of goods only. The status information that is used in the replenishment calculation (see Jenniskens (1986)) can be either local or integral (i.e., echelon information).

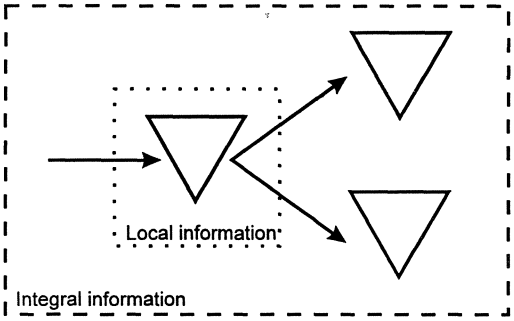


Figure 3. Local vs. integral status information

Local status information refers to information about stock norms for the inventory of one DC only – also called installation stock norms – and about local demand, i.e., the demand of the link next downstream. Integral (echelon) information is information about stock norms for the inventory of a complete system and about integral demand information, i.e. information about the demand of the independent customer (see Figure 3).

1.3.3 Central stock function

The central stock decision deals with the question whether a central depot function is needed for storing the replenished items or whether it is possible to use a central DC as a cross docking point only. At a cross docking point, goods are not stored but immediately shipped to the local DCs after receipt at the central DC. Figure 4 depicts a situation with a central stock function and one without.

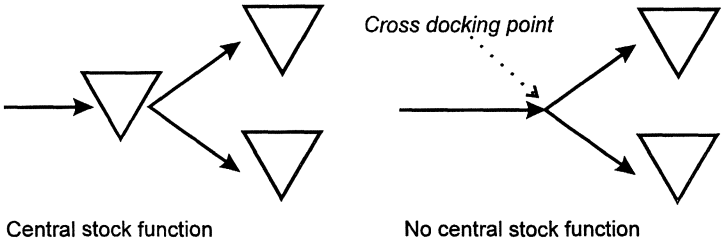


Figure 4. Central stock function vs. no central stock function

As this research deals with distribution control in multi-echelon inventory systems, we assume that the local DCs always carry inventory.

1.3.4 Co-ordination of the allocation process

The co-ordination of the allocation process refers to the degree of centralised control (see Rosenfield and Pendrock (1980), Lee and Billington (1993) for a discussion on centralised control). Locally co-ordinated allocation means that the local DCs order goods at the central facility at their own initiative. The frequency of allocation is dependent on the review frequency of the local stock.

Centrally co-ordinated allocation implies that the decision on the timing and the amount of the shipments of goods to the local DCs is not made by the local DCs but by a central department. The replenishment batch may be completely allocated in one time to the local DCs or it may be allocated in fractions. The α-policy (Erkip, 1984) is a technique according to which the replenishment batch is allocated in two times. According to this policy, a fraction α is allocated to the local DCs and a fraction $(1-\alpha)$ retained at the central depot each time a new replenishment batch is received. As soon as there is a local need for a shipment, the remaining fraction $(1-\alpha)$ is allocated to the local DCs in a centrally co-ordinated way. It is also possible that only the initial allocation is centrally co-ordinated and that the remaining stock is allocated to the local DCs in a locally co-ordinated way. In that situation, it should be determined when the next allocation will take place.

The terms centrally co-ordinated and locally co-ordinated allocation are relatively similar to the concept of push and pull. Push is generally related to central control of inventory (Christopher, 1985). Pull, on the other hand, is related to local control of inventory (Christopher, 1985). Brown (1967) uses a similar differentiation between push and pull. Generally speaking, however, there seems to be a certain disagreement or misunderstanding about what push and pull really consist of. Pyke and Cohen (1990) argue that it is not possible to label an entire production or distribution system as push or pull. They therefore introduce a framework to differentiate between elements in distribution and production that are push and that are pull. As the discussion on push and pull tends to be somehow ambiguous, the use of these terms may be confusing. We will therefore avoid the use of the terms push and pull as much as possible.

1.4 Use of classification in practice

The control decisions comprise the core elements of distribution control. Distribution control techniques such as Distribution Requirements Planning or Statistical Inventory Control can be represented in terms of these decisions. With this classification it is possible to design distribution control systems through relating the control decisions to the relevant PPM[2]-characteristics. Distribution control techniques can then be selected based on the decisions.

---

[2] PPM means Process Product Market (see Section 1.2.1)

This process of relating distribution control decisions to the relevant characteristics should be done for groups of products with homogeneous PPM-characteristics. Using the same control technique for all products within a company may be sub-optimal for some product groups due to a misfit of the control technique with the specific characteristics of the products, the processes used or the markets served. If distribution control decisions are made per group of homogeneous products in terms of their product, process and market characteristics - these groups are called the distribution control situations - this misfit may be avoided.

1.5 Research methodology

The research problem discussed here is to relate the relevant PPM characteristics to the four control decisions. There are two parts that should be considered:

- Which PPM-characteristics are relevant for each of the distribution control decisions?
- To what extent are the PPM-characteristics relevant and which are the important characteristics?

The first part is an exploratory question. Besides literature review, exploratory case study research has been chosen as a method to address this part of the research problem. Three case studies have been conducted in different types of environment (one in a department store, and two in different manufacturing environments). To address the second question, quantitative research (computer simulation) is needed.

To avoid a lengthy discussion on all four control decisions, the exploratory and the quantitatively oriented questions have been investigated only for one decision. The decision on the type forecasting technique was selected for further research, since this is generally a simply identifiable and changeable function in a company. Results of this research may therefore easily be applied in businesses. We will therefore focus at the forecasting decision in the remainder of this paper. We first discuss our case study and literature research in this area. The simulation will be discussed after that.

1.6 Case study results on forecasting

Three case studies have been performed. The first study took place within Walker Europe, a manufacturer of exhaust systems. The second was within Vroom and Dreesmann, a department store company and the third was within EMI Compact Disc Manufacturing. The companies have been chosen in such a way that both a retail and a manufacturing environment were researched, with diverse characteristics of products, processes and markets.

In the Walker case study, the relatively low *demand uncertainty* together with the presence of a seasonal *demand pattern* of A-items - the fast movers - appeared to be key factors for the selection of a forecasting technique which can incorporate demand patterns. For the B-items, the less stable demand and the relatively high *amount of Stock Keeping Units (SKU's)* lead to the selection of a forecasting technique without a pattern. For C-items - the slow movers - the use of forecasts that can incorporate patterns is less useful due the relatively large *production batch size*, which results in a large batch size stock. As a result, improved accuracy of forecasting techniques and hence lower safety stock levels only has a minor effect on the total stock level. A technique which can not incorporate a demand pattern is thus preferred.

Within V&D, the number of SKU's is relatively high. For this reason, a forecast without a pattern is preferred for the allocation decision. Two product groups have been researched, suit cases and drug store items. For the suitcase distribution study, the decision to select a forecasting technique without a pattern is predominantly determined by the large supply batch size. Also during promotional action periods, such a forecasting technique is preferred for both the replenishment and the allocation decision. Due to the short term of an action period, it is often not meaningful to take demand patterns in an action period into account in the replenishment decision. If the *lead time* is short, which is the case for some drug store products, it is preferred to use a constant forecasting technique without any demand patterns. The ability to incorporate demand patterns during these short lead times is not expected to add value, as it is difficult to discern a pattern during such a short time period.

The sales pattern of new items at EMI is too unpredictable to use a forecasting technique that can incorporate patterns. For this reason, a forecast without a pattern has been selected. For mature items, a simple technique is preferred as well due to the unpredictable sales pattern. For the old items, the large production batch size is an important reason to select a forecast without a demand pattern. The high value density and the low *demand rate* enables the use of small distribution batches for the old items. As a result, the local stock norms can be low and the allocation can simply equal the quantity sold.

The relevant PPM-characteristics and their effect on the forecasting decision are summarised in Table 2.

Table 2. PPM-characteristics and their effect on the type of forecasting technique; W=Walker, V=Vroom and Dreesmann, E=EMI)

PPM-characteristic	Value of characteristic	Type of forecast	Case study
Lead time	Short lead times	Without a pattern	V
Production batch size	Large batch sizes	Without a pattern	W,E
Amount of SKU's	Large amount of SKU's	Without a pattern	V,E
Demand uncertainty	Low uncertainty	With a pattern	W
Demand rate	Low demand rate	Without a pattern	E
Demand pattern	Seasonal demand pattern	With a pattern	W

1.7 Literature overview: forecasting in logistics perspective

1.7.1 Forecasting accuracy

Whether it is possible to forecast demand is generally understood to be related to the size of the forecast error. The performance of forecasting techniques is dependent on the circumstances under which they are used. Makridakis and Hibon (1979), Makridakis (1988) and Armstrong (1985) draw the conclusion that simple techniques often outperformed the more sophisticated ones. This is supported by Alstrøm and Madsen (1994), who have investigated the effect of different types of exponential smoothing forecasting techniques by means of simulation for seasonal demand patterns. However, what they judge to be simple methods, may yet be relatively complex methods for practitioners.

Many articles use some form of forecasting accuracy as a measure to judge forecasting techniques. Assuming that there is enough data available for forecasting purposes, the possibility to forecast will generally become a problem if demand gets non-stationary – i.e., the mean and standard deviation of demand are changing in the course of time. Jacobs and Whybark (1992) conclude that if demand has become non-stationary and thus uncertain, SIC - which uses a constant forecasting technique - outperforms a more complex approach with MRP. For this reason, demand uncertainty influences the choice for a forecasting technique.

1.7.2 Other measures than forecasting accuracy

Ritzman and King (1993) have investigated the effects of forecast errors and concluded that when the objective is to have low inventories, the effect of having a good forecast is less important than the effect of having small lot sizes. Silver and

Peterson (1985) argue that for cheap items and slow moving items, forecasts are not useful. The reason for this is that it is hard to achieve any sizable absolute savings in the costs of these items. The guideline for these types of items should therefore be to keep procedures simple. Product value and the demand rate are thus expected to influence the choice for a forecasting technique.

The question whether a specific forecasting technique is useful is not only related to forecast accuracy. It can be argued that if some form of forecasting is possible, it is always useful to apply forecasting, but this argument leaves out the costs involved with forecasting. Makridakis and Wheelwright (1978, 1979) give an overview of forecasting techniques and their costs. They state that the best forecasting technique for a situation is dependent on the pattern of the data, the time horizon of the forecast, the cost, and the ease of application. They also report that exponential smoothing methods have generally been found to be superior in short term forecasting to other techniques.

1.8 Conclusion

Although the evaluation of forecasting techniques has received much attention in the literature, the discussions are mainly oriented towards the evaluation of forecasting techniques by means of statistical criteria (see Makridakis and Wheelwright (1978, 1979), Makridakis (1986, 1988), Makridakis and Hibon (1979), Armstrong (1985)). However, there seems to be a lack of agreement on the question which statistical criterion should be used as the measure of forecast error (Armstrong, 1985).

From literature, we can conclude that at least the following aspects play a dominant role in the question what type of forecasting technique should be used:

Table 3. Influence of PPM-characteristics on forecasting (literature summary)

PPM-characteristic	Forecasting type	Literature Reference
Seasonal demand	Forecast with pattern	Alstrøm & Madsen, 1994
Non-stationary demand	Forecast without pattern	Jacobs & Whybark, 1992
Low product value	Forecast without pattern	Silver & Peterson, 1985
Low demand rate	Forecast without pattern	Silver & Peterson, 1985

In this paper, we take a different direction at forecasting than most other types of forecasting research. We assess the effect of forecasting techniques from a logistics perspective instead of a statistical perspective. As opposed to most previous forecasting research, we are not primarily interested in statistical measures of forecast errors as a criterion to assess the performance of a forecasting technique.

For this reason, we have conducted a simulation experiment. In this experiment, which will be discussed in the next section, the forecasting problem will be approached from a logistical point of view by investigating the effect of forecasting techniques on logistical aspects.

1.9 Simulation experiment

1.9.1 Research approach

Fildes & Beard (1992) argue that forecasting techniques in production and inventory control should be chosen in line with data characteristics. Unfortunately, demand data often have little structure and a relatively high level of randomness. A further characteristic is the low stability of the forecasting performance over time (Fildes & Beard, 1992). The benefits of matching a forecasting technique to a homogeneous subset of data may therefore yield substantial accuracy benefits.

It is not only the characteristics of the demand data – such as the mean and the variance – that have an influence. Also characteristics of the processes used – lead time and batch size – are expected to have an effect. For example, the larger the lead times are, the more difficult it may be to make a reliable forecast. As discussed earlier, distribution control techniques should be selected based on the characteristics of products distributed, markets served and processes used to serve these markets. In line with this, forecasting techniques are also assumed to depend on characteristics of products, processes and markets. Case study research has been used to reveal the characteristics to be used in the simulation experiment (see Section 1.6 and De Leeuw (1996a)). Based on these studies and the literature discussed earlier, it has been decided to investigate the effect of the following characteristics on forecasting techniques in more detail:

- The length of the inventory review period.
- The size of the supply batch.
- The length of the supply lead time.
- The presence of a trend in the demand.
- The presence of seasonality in the demand.
- The stability of demand.

Only market and process characteristics appeared to be relevant, product characteristics are hence omitted in this discussion. As said earlier, the effect of different forecasting techniques will be assessed in this paper based on the logistics impact. The average level of physical stock in a stocking location, measured in units, will be used for this (comparable to Alstrøm and Madsen (1994), although they also incorporate capacity aspects in their performance measures).

1.10 Description of the simulation model

Given the practical direction of the research, a main requirement for the forecasting techniques researched was that they be used in practice. In line with the classification presented earlier in Table 1, two types of forecasting techniques have been distinguished: techniques that can incorporate demand patterns and techniques that can not. The Holt Winters exponential smoothing technique (abbreviated *HW*), as described by Silver and Peterson (1985) has been selected to serve as an example of a technique that can incorporate demand patterns. Moving average (abbreviated *MA*) and single exponential smoothing (abbreviated *Exp*) have been selected to serve both as an example for a technique that can not incorporate demand patterns. All three techniques are used in practice. For reasons of simplicity, a single echelon system has been used for the simulation (see Figure 5).

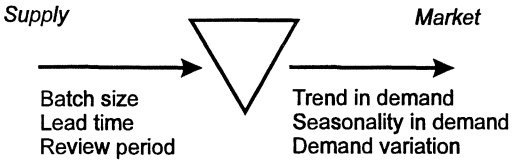


Figure 5. Single echelon system and relevant PPM-characteristics

As a result of the short life cycles, products are often phased out before there is enough data available to establish a stable forecast. The use of long simulation runs, where forecasts can be based on a large amount of data, is therefore judged inapplicable. To account for the effect of shortening life cycles, many short runs are used in the experiment. In each run, 2 years were simulated. Different situations of demand stability have been researched, ranging from stable demand to uncertain demand. The measure for demand uncertainty is the standard deviation of the forecast error divided by the average demand. The reorder level is recalculated every simulation run, based on the average forecast during the lead time plus review period. To ensure that the results of the simulation experiments are comparable, the simulation runs are stopped if a fill rate of 95% has been reached. In order to attain this fill rate, the average forecast during the lead time plus review period has been multiplied by a safety factor equal to (lead time + review period + safety time) / (lead time +review period). In the simulations, the safety time has been varied until a fill rate of 95% was reached (+/- 0.01%). For this adjustment, a simple procedure is used which starts with a safety time value, calculates the fill rate and adjusts the safety time based on the fill rate value[3]. This adjusted safety

[3] If the fill rate is lower than 95%, the safety time is adjusted upward, otherwise it is adjusted downward. If a safety time entailing in a fill rate higher than 95% and one with

time is used for subsequent simulation runs and recalculated after each run. This procedure is different from Alstrøm and Madsen (1994), who calculate the reorder levels based on a lost-sales-model with buckets of one day.

To cover a broad range of possibilities, each PPM-characteristic has different settings. These settings are summarised in Table 4.

Table 4. Settings of the forecast parameters used

Characteristic	Values researched
Review period	1 week or 4 weeks
Batch size	Lot For Lot (LFL), 400 and 1600 units (average demand is 100 units per week)
Lead time	4 and 16 weeks
Trend	Trend (demand increase of 1% per week) and no trend
Seasonality	Season or no season
Standard deviation of forecast error divided by average expected demand[4] (abbreviated: DU)	0.1, 0.5 and 1

If demand is seasonal, the demand increases linearly from week 1 on and reaches its top in week 24. From week 24 onwards, it declines and reaches its minimum in week 72. Then demand increases again according to the same cycle. The peak of the demand is 50% above the starting level and the minimum level is 50% below the starting level (if there is no trend in demand). In case demand is non-seasonal, the seasonal indices in HW are set to zero. For each forecasting technique, two settings have been used: one which uses a long history of data and the other which uses a short history. For Moving Average, a history of N=48 resp. N=8 periods is used. Comparable values for the forecasting parameters of the other two techniques have been calculated using Silver and Peterson (1985).

a fill rate lower than 95% have been found, interpolation between those two points is used to estimate the safety time which renders a fill rate closer to the desired fill rate. This interpolated safety time is used in a subsequent simulation run and recalculated each run.

[4] This is a measure of demand uncertainty; we will abbreviate this measure in the remainder of this paper as DU (Demand Uncertainty)

1.11 Conclusions of the simulation research

Below, we will summarise the conclusions of the simulation research. For details on the simulation, we refer to De Leeuw (1996a).

1.11.1 The demand uncertainty has a significant impact on the forecasting performance

The performance of techniques that incorporate demand patterns deteriorates quickly if the demand uncertainty (as measured by DU) increases. These forecasting techniques quickly become inaccurate due to the property that all noise is translated into a pattern, particularly if demand contains seasonality as well. The performance of techniques that incorporate demand patterns deteriorates further if long lead time apply. A technique not containing a demand pattern, such as moving average, is preferable in that case due to the better stability of the forecast. It is also beneficial in case of a high demand uncertainty to use data over a long period to establish a forecast. This reinforces the stability of the forecast. It only seems beneficial to incorporate demand patterns in a forecast if demand uncertainty is low and if there is season and/or a trend in demand involved. These techniques are then better able to follow the pattern in demand, leading to lower stock levels. If there is no trend nor a season and if demand uncertainty is low, the differences between the forecasting techniques are small.

1.11.2 Large batch sizes and long review periods dampen the variation of demand

For large uncertainty, the stock levels decrease if large batch sizes or long review periods are used. The reason for this is that the number of reorder moments decreases; as a result, the number of occasions where forecast mistakes can have an influence on the stock level is smaller. As a result, if the demand variation is rather large, the use of large batch sizes or long review periods stabilises the forecast. Therefore, if the demand uncertainty is large (the measure DU exceeds 0.5), it is preferable to use large batch sizes or a long review period if forecasts that incorporate a pattern are used.

1.11.3 It is nearly always preferred to use a long data history

The use of long data history for the forecast reinforces the stability of the forecast and can hence result in lower stock levels. This holds especially if the environment is 'complex', because of factors such as demand variation, long lead times or a pattern. It should be noticed, however, that this may depend on the type of

demand distribution used in the simulation model (the demand distribution used is not characterised by changes in the average demand level).

1.11.4 A forecast without a pattern is often the best to choose

Forecasting techniques which incorporate demand patterns only perform well under very specific circumstances (i.e., limited uncertainty combined with seasonality). However, if the environment gets complex (i.e., the demand uncertainty measure DU > 0.5 and demand contains seasonality), the good performance of these techniques quickly deteriorates and turns much worse than the performance of techniques that do not incorporate demand patterns. Real life is often relatively uncertain and demand shows large variations. Because of the vulnerability of forecasting techniques that incorporate patterns to uncertainty, especially if demand contains seasonality and long lead times apply, the use of techniques which do not incorporate demand patterns are often preferable. Batch sizes may dampen the negative effect of the techniques that incorporate patterns, but simple techniques are still preferred.

1.12 Differences of simulation with earlier findings

From above, we can conclude that the use of forecasting techniques that can incorporate demand patterns is confined to very stable situations with patterns in demand. Table 5 again summarises the literature findings on forecast research.

Table 5. Influence of PPM-characteristics on forecasting (literature summary)

PPM Characteristic	Forecasting type
Low product value	Forecast without pattern
Non-stationary demand	Forecast without pattern
Low demand rate	Forecast without pattern
Seasonal demand	Forecast with pattern

It should be noted that this research is oriented at the analysis of multi-dimensional criteria, as opposed to much literature. From literature it is not particularly clear for example what to do in case demand is uncertain and contains a seasonal pattern as well (if demand is uncertain, a forecast without a demand pattern is preferred, but if there is seasonality, a forecast with a pattern is preferred according to literature). As such, the literature conclusion that forecast techniques

which can incorporate a pattern should be used in case of seasonal demand needs additional detail. High demand uncertainty in combination with a seasonal demand pattern favours the use of techniques that can not incorporate demand patterns. The statement that low demand rate - and therefore high demand uncertainty as a result because low demand is often very erratic demand - favours forecasts without demand patterns is supported by this simulation. The literature conclusion on product value has not been researched. We have only investigated the total level of inventory and omitted the cost aspects of it, but we support the observation of Silver and Peterson (1985] that if product value is low, the effort put in forecasting should be minimal.

As far as the case study conclusions are concerned (see Table 6), findings can be made more precise based on the simulation study.

Table 6. Case study conclusions on forecasting

Value of characteristic	Type of forecast
Short lead times	Without a pattern
Large batch sizes	Without a pattern
Large amount of SKU's	Without a pattern
Low uncertainty	With a pattern
Low demand rate	Without a pattern
Seasonal demand pattern	With a pattern

The simulation study showed that forecasts which can incorporate a pattern may be preferable in case of short lead times if demand is relatively certain and there is a seasonal pattern in demand, but the differences with the simpler forecast which can not incorporate a pattern appeared to be small (10%).

The use of forecasting techniques seems relatively insensitive to batch sizes. The Demand Uncertainty appeared to have a more dominant role in the choice for a forecasting technique. Batch sizes may dampen out fluctuations in demand. If forecasting techniques which incorporate patterns are applied in uncertain environments, large batch sizes may improve their performance as a result. We have not tested the statement about the amount of SKU's forecasted as only one product has been simulated. The last three statements can be supported, although as mentioned earlier, if demand is seasonal and uncertain, a forecast technique which can not incorporate patterns is preferred.

1.13 References

Alstrøm, P., Madsen, P. (1994): Evaluation of forecast models used for inventory control during a product's life cycle: a simulation study. in: International Journal of Production Economics 35, 191-200

Armstrong, J. Scott (1985): Forecasting by extrapolation: conclusions from 25 years of research, in: Interfaces 14 no 16, 52-66.

Brown, R.G. (1967): Decision rules for inventory management. (Holt, Rinehart and Winston) New York

Christopher, M. (1985): The strategy of distribution management. (Gower Press) London

De Leeuw, S.L.J.M. (1996a): The selection of distribution control techniques in a contingency perspective. Ph.D. dissertation, Eindhoven University of Technology, Eindhoven

De Leeuw, S.L.J.M. (1996b): Distribution control at Exhaust Systems Europe. in: International Journal of Physical Distribution and Logistics Management 11 no 8,

Erkip, N.E. (1984): A restricted class of allocation policies in a two-echelon inventory system. Technical report 628, School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York

Fildes, R., Beard, C. (1992): Forecasting systems for production and inventory control. in: International Journal of Operations and Production Management 12 no 5, 4-27

Jacobs, F.R., Whybark, D.C. (1992): A comparison of reorder point and materials requirements planning inventory control logic. in: Decision Sciences 23, 332-342

Jenniskens, F. (1986): Meer zekerheid bij het omgaan met onzekerheden. Master's thesis, Eindhoven University of Technology, Eindhoven

Lee, H.L., Billington, C. (1993): Material management in decentralised supply chains. in: Operations Research 41 no 5, 835-847

Makridakis, S. (1986): The art and science of forecasting. in: International Journal of Forecasting 2, 15-39

Makridakis, S. (1988): Metaforecasting; ways of improving forecasting accuracy and usefulness. in: International Journal of Forecasting 4, 467-491

Makridakis, S., Hibon, M. (1979): Accuracy of forecasting: an empirical investigation. in: Journal of the Royal Statistical Society 142 part 2, 97-145

Makridakis, S., Weelwright, S.C. (1978): Forecasting: methods and applications. (John Wiley and Sons) New York

Makridakis, S., Wheelwright, S.C. (1979): Forecasting: framework and overview. in: Makridakis, S., Wheelwright, S.C. (eds.): Forecasting, TIMS Studies in the Management Sciences 12, North Holland Publishing Company, Amsterdam

Martin, A.J. (1993): Distribution Resource Planning: the gateway to true quick response and continuous replenishment (Oliver Wight Companies) Essex Junction

Masters, J.M., Allenby, G.M., La Londe, B.J., Maltz, A. (1992): On the adoption of DRP. in: Journal of Business Logistics 13 no 1, 47-67

Orlicky, J (1975): Material Requirements Planning (Mc Graw Hill) New York

Prasad, S (1994): Classification of inventory models and systems. in: International Journal of Production Economics 34, 209-222

Pyke, D., Cohen. M.A. (1990): Push and pull in manufacturing and distribution systems. in: Journal of Operations Management 9 no 1, 24-43

Ritzman, L.P., King, B.E. (1993): The relative significance of forecast errors in multistage manufacturing. in: Journal of Operations Management 11, 51-65

Rosenfield, D.B., Pendrock, M.E. (1980): The effects of warehouse configuration design on inventory levels and holding. in: Sloan Management Review, summer, 21-33

Silver, E.A. (1981): Operations Research in inventory management: a review and critique. in: Operations Research 29 no 4, 628-645

Silver, E.A., Peterson, R. (1985): Decision systems for inventory management and production planning (John Wiley and Sons) New York

Van Donselaar, K.H. (1989): Material coordination under uncertainty. Ph.D. dissertation, Eindhoven University of Technology, Eindhoven

Van Donselaar, K. (1990): Integral stock norms in divergent systems. in: European Journal of Operational Research 45, 70-84

Inventory Positioning in a Two-Stage Distribution System with Service Level Constraints

Ulrich Tüshaus[1] and Christoph Wahl[2]

[1] Universität der Bundeswehr Hamburg, 22039 Hamburg, Germany
[2] Universität St. Gallen, 9000 St. Gallen, Switzerland

Abstract. A 1-warehouse, n-retailer system is considered in which periodic customer demand only occurs at the lower echelon. Inventory may be stored at both echelons. Transshipments between stockpoints are excluded. All stockpoints replenish inventory by means of local (T, S)-policies which are assumed to share a simple nested schedule. An approximate mathematical representation of the considered distribution system is introduced which lends itself to a use in performance measurement or optimization.

Numerically attractive expressions for stock on hand and backlog are based on a cycle-oriented approach which allows an easy handling of the complexity arising from stochastic customer demands. Besides, other performance measures like the well-known fill rate, a γ-service and the so-called lead time index are approximated. The analytical link between system states at the upper and lower echelons is achieved by modeling customer waiting times as coupling variables.

Numerical studies show a high correspondance between simulated and analytical quantities. Moreover, numerical results underline the appropriateness of including average waiting time estimates as an additional time factor in formulas for stock on hand and backlog at an arbitary retailer.

1 Introduction

In this paper, a core problem concerning cost-efficient inventory positioning in multi-echelon distribution systems is addressed: The derivation of robust and numerically inexpensive approximate formulas for performance measures like average stock on hand or customer service levels.

Consider a 1-warehouse, n-retailer system in which periodic customer demand only occurs at the lower echelon. Inventory may be stored at both echelons. Lateral transshipments between stockpoints are excluded. All stockpoints replenish inventory by means of local (T, S)-policies which are assumed to share a simple nested schedule. With local (T, S)-policies, the inventory position IP at a stockpoint – defined as stock on hand plus outstanding orders minus backlog – is reviewed periodically, i. e. every T time units. In case IP is below an order-up-to-level S, an order of $Q = S - IP$ is triggered immediately. Obviously, the order quantity Q is a stochastic variable depending on the demand process during the preceding review period of length T.

An approximate mathematical representation of the considered distribution system is introduced which lends itself to a use in performance measurement or optimization. Numerically attractive expressions for stock on hand and backlog are based on a cycle-oriented approach which allows an easy handling of the complexity arising from stochastic customer demands. Besides, other performance measures like the well-known fill rate, a γ-service and the so-called lead time index are approximated. The analytical link between system states at the upper and lower echelons is achieved by modeling average customer waiting times as coupling variables. Customer waiting time estimates are based on a state-dependent, retailer specific approach. The approximation concept used here seems to overcome some of the criticism concerning the use of fixed waiting times in periodic review models.

Numerical studies show a high correspondance between simulated and analytical quantities. Moreover, numerical results underline the appropriateness of including average waiting time estimates as an additional time factor in formulas for stock on hand and backlog at an arbitrary retailer.

The structure of this paper is as follows: In Sec. 2 a brief overview for multi-echelon inventory systems with periodic review is given. Fundamental assumptions and modeling concepts applied here are described in Sec. 3. In Sec. 4 and 5 estimates for important performance measures at the upper and lower echelon are presented and their use is motivated. Computational results in Sec. 6 underline the validity of estimates. Finally, in Sec. 7 the main findings are summarized.

2 Literature Review

Base Stock Systems with Periodic Review. Most literature about multi-echelon inventory models with periodic review concentrates on systems with base stock control. In inventory models with base stock control, policies at echelons up are modeled as functions of policies at echelons down; all replenishment decisions are made on the basis of echelon-information about stock levels, current demand and delivery processes. On the one hand, base stock systems require a considerable degree of coordination to be successfully implemented which, until today, might cause significant technical and/or organizational problems. On the other hand, using base stock control can increase system performance considerably and simultaneously keep costs low since real time data of lower echelons is exploited for decisions at higher echelons. Therefore, from an economic point of view, it does not come with surprise that the major part of recent scientific papers focusses on this area (for a thorough overview for centralized inventory control see Federgruen (1993), Ch. 3).

In their fundamental article on inventory management in multi-echelon systems Clark and Scarf (1960) show the economic dominance of centralized inventory control using the *echelon stock* as a planning basis. The echelon

stock is a bottom-down definition of system-wide net stock, which requires all relevant information to be always available. In contrast to this centralized concept one finds *installation stock control* which is restricted to local information about demand process and stock levels only.

Clark and Scarf (1960) show that for divergent multi-echelon inventory systems an intricate allocation problem arises when there is insufficient stock at supplying stockpoints. In such a case an efficient allocation to the next lower stockpoints must be made. Here, Clark and Scarf (1960) could not derive an optimal replenishment strategy for divergent systems, i. e. a critical number policy as for serial systems. To be able to derive near-optimal policies, the authors introduce the so-called *balance assumption.* According to the balance assumption, scarce available stock at supplying points is allocated in such a way that no stockpoint in the system is discriminated with respect to economic constraints like, for example, customer service levels, i. e. stock levels should be balanced. A mathematical analysis of the impact of imbalance situations can be found in Zipkin (1984). Until today, no optimal allocation rule has been derived, but there exists a rich literature about reasonable rationing policies in echelon stock systems; see e. g. Eppen and Schrage (1981), Federgruen and Zipkin (1984), Jackson (1988), Lagodimos (1992), van der Heijden et al. (1996).

In particular, centralized $(1, S)$-policies, where each period an order of variable size is triggered, have attracted major research interests (see e. g. Langenhoff and Zijm (1990), de Kok (1990), de Kok et al. (1994), Johnson et al. (1995), Verrijdt and de Kok (1996)). The attractiveness to plan on a daily basis might be partially explained by the following factors: Significant improvements in data interchange (EDI, intranets etc.) now allow the use of centralized planning schemes at relatively low costs with real time data or at least frequently updated information available. From an economical point of view, fixed order costs are often negligible in distribution planning so that ordering periodically often proves to be advantageous.

But, several reasons may make favorable a use of the more general (T, S)-policies: Given non-negligible fixed order costs, lower ordering frequencies at all stockpoints are economically sensible. Clearly, (T, S)-policies seem to offer many opportunities of coordinating a multi-echelon replenishment schedule; in particular replenishment orders from lower stockpoints can be pooled at higher echelons over a longer review period. Atkins and Iyogun (1988) emphasize opportunities for exploiting economies of scale when pooling replenishment orders of similar articles. Contracts with carriers often provide fixed reorder intervals. Furthermore, the nervousness of stocking policies at lower echelons can be expected to reduce with decreased planning frequencies at higher echelons (see Jensen (1996)).

Virtual Models. The class of virtual models introduced by Graves (1989) is composed of both centralized and decentralized planning features. The replenishment order process is nested: When the supplying stockpoint receives

an order all retailers place orders. Furthermore, stock is allocated *virtually*, i. e. orders for individual units at the supplying stockpoint are filled in the same sequence as the original demands at the supplied stockpoints. The virtual allocation assumption is only approximate, but it allows the analysis of periodic review models with tools known from systems with continuous policies. In Graves (1996) a possible extension to systems with stochastic lead times is discussed. For a two-echelon inventory system, Axsäter (1993) derives a recursive optimization procedure which allows an exact evaluation of different (T, S)-policies.

Installation Stock Systems with Periodic Review. Only few literature exists for installation stock systems with periodic review, where only local information is used at each stockpoint. It is obvious that local stock control tends to produce higher current inventory costs than centralized planning schemes. On the other hand control mechanisms for installation stock systems are cheaper, less complicated and easier implemented. In practice, organizational reasons often forbid applications of centralized control systems.

Rosenbaum (1981b) reports on a successful application of an approximate two-echelon distribution model of the pull-type at The Eastman Kodak Company. At the lower echelon, regional distribution centers apply local (s, Q)-policies where each time the inventory position IP falls below the reorder point s a fixed quantity Q is ordered. At the upper echelon, two central distribution centers replenish their inventory position according to a (T, S)-policy. The formal model is derived in Rosenbaum (1981a) for normally distributed customer demands. Approximate fill rate expressions at each stockpoint and an easy-to-handle heuristic for system-wide safety stock minimization are suggested.

Simultaneously, both Matta and Sinha (1991) as well as Rogers and Tsubakitani (1991) present similar two-echelon optimization models for a pull-system with local $(1, S)$-policies at both echelons and normally distributed customer demands. While Matta and Sinha (1991) consider a typical 1-warehouse, n-retailer system, Rogers and Tsubakitani (1991) concentrate on a simple divergent production system with n final products and one common component. Both approaches make use of a fixed additional delay increasing deterministic lead times. The derivation of the delay grounds on Little's formula (see Little (1961)) and is only approximate. For $(S-1, S)$-models a similar use of Little's formula was suggested by Sherbrooke (1968), already. Note that with $(S-1, S)$-policies the inventory position is reviewed continuously and in case of (unit-)demand a corresponding order is triggered. Deuermeyer and Schwarz (1981) stated acceptable results, too, when using Little's formula to derive the average waiting time in a two-echelon, (s, Q)-distribution system with identical retailers and Poisson demand.

Approximate performance measures for a multi-echelon distribution system with local (T, S)-policies are derived in van der Heijden (1993), Ch. 7. Customer demands are assumed to follow a compound Poisson process and

stochastic lead times have a stationary probability density function. Numerical results for both a 2-echelon and a 3-echelon system show the adequacy of the approximate formulas which are evaluated by means of the PDF-method suggested in de Kok (1991a), (1991b). Customer waiting time formulas are based on an approach for a one-stage inventory system as suggested in van der Heijden and de Kok (1992). In fact, if lead times are deterministic, the waiting time distribution derived in van der Heijden and de Kok (1992) is exact.

For the even more general case of customer demand following a compound renewal process Chen and Zheng (1992) derive an exact waiting time distribution for a one-stage inventory system with (T, S)-control. When customer demand can be expected to follow a compound Poisson process and the lead time is constant, Chen and Zheng (1992) obtain easy-to-handle formulas for the waiting time distribution.

Matta and Sinha (1995) discuss a special two-echelon distribution system of the pull-type. At the retailer level, $(1, S)$-policies are applied. The warehouse uses a periodic (s, S)-policy. Customer demands are assumed to be normally distributed. From standard results in renewal theory Matta and Sinha (1995) apply approximate expressions for the mean and variance of waiting times which are required when determining moments of the (effective) lead time demand at each retailer. This two-moment approach for modeling waiting times is strongly based on an earlier paper of Svoronos and Zipkin (1988) for $(S-1, S)$-models. Computational results underline the appropriateness of the approximate model in case of high service levels and low coefficients of variation of periodic customer demand.

3 Two-Echelon Inventory System

Consider a two-echelon distribution network with stockpoints both at the retailer level (echelon 1) and the warehouse level (echelon 2). There are n retailers supplied by one warehouse. Define $\mathcal{I}^c = \{c, 1, \ldots, n\}$ as the set of all stockpoints and $\mathcal{I} = \{1, \ldots, n\}$ as the subset of retailers. Echelons 1 and 2 are linked by transportation processes modeled as deterministic lead times L_i for all $i \in \mathcal{I}$. The central warehouse replenishes its stock by periodically ordering from a production location which, by assumption, always has sufficient capacity. Deliveries arrive at the warehouse a constant lead time L_c after the corresponding order was triggered. Figure 1 illustrates the network structure described above.

Pull System. A pull-system is considered, i. e. local planning schemes are applied for inventory positioning in a multi-echelon setting. In a decentralized planning environment, each stockpoint controls its inventory position independently from system states at other stockpoints, i. e. only local information about current demand, inventory levels and outstanding orders is used for

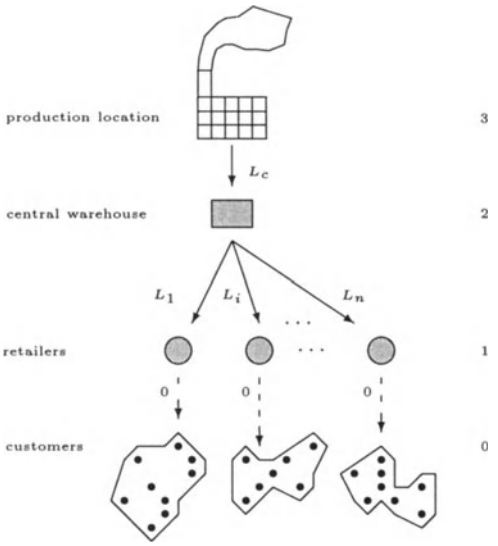


Fig. 1. Two-echelon network

decision making. Note that in real world distribution systems such schemes make up a significant part of planning systems. Simultaneously, only few work on locally controlled distribution systems exists. Therefore, this paper contributes to a barely analyzed, but practically important area. In subsequent paragraphs, we describe conditions and assumptions under which the system is expected to work.

Monitoring Policy. The monitoring policy defines the points in time at which model variables are registered. We will use discrete monitoring which can be defined as follows: Define t_m, $t_m \in \mathbb{N}_0$, the time between two consecutive moments in time at which the stock level is counted. Furthermore, let us assume review periods $T = m_T \cdot t_m$ and fixed lead times $L = m_L \cdot t_m$ to be integer multiples of the monitoring interval t_m. Without loss of generality, for further analysis, the monitoring interval will be set to $t_m = 1$.

Review Policies. All stockpoints $i \in \mathcal{I}^c$ follow local (T_i, S_i)-policies. Reorder cycles $T_i \in \mathbb{N}$ have been determined by management in advance according to a long term schedule, for example. Consequently, the set of order-up-to-levels $\mathcal{S} = \{S_c, S_1, \ldots, S_n\}$ represents all decision variables for optimization problems which underlie operational tasks of inventory positioning in distribution networks. Clearly, for a moderate number n of retailers, optimal T_i can be found by systematic enumeration.

Nested Schedule. To assure stationarity of the *derivative* (= internal) demand process at the upper echelon we assume a (wide-spread used) nested schedule:

The fixed reorder interval T_c of stockpoint c must be a multiple integer of the reorder intervals T_i at the lower echelon:

$$k_i = T_c/T_i \quad k_i \in \mathbb{N}, \text{ for all } i \in \mathcal{I} \ . \tag{1}$$

By *derivative* we denote the fact that internal demand is *derived* from preceding orders of stockpoints $i \in \mathcal{I}$.

Order Sequencing. Customer demands and orders are satisfied by stock on hand on a first-come first-served basis. In case of a stockout at stockpoint $i \in \mathcal{I}^c$ demand is (partially) *backlogged.* Whether allowing for backlogging is reasonable or not depends, for example, on the competitiveness of a considered branch, the substitutability of goods or simply the liability of customers.

Allocation Rule. At stockpoint c, order-splitting is allowed as to achieve a higher customer service. Clearly, orders must be splitted in case of scarce stock on hand $I_{c,t}$ at time t, only. An allocation takes place when current order processes $Q_{i,t'}$ from stockpoints $i \in \mathcal{I}$ occur at stockpoint c or, alternatively, when a current delivery $Q_{c,t'}$ arrives at stockpoint c.

We assume scarce physical stock $I_{c,t'}$ to be allocated *proportionally* over stockpoints at echelon 1 where t' defines the point in time when an allocation of stock on hand to stockpoints $i \in \mathcal{I}$ realizes. The assumption of proportional allocation seems to be a direct impact of the local planning environment since, at least intuitively, there is no better way for stockpoint c than to distribute products proportionally.

The proportional allocation rule can be defined as follows: At time t', stockpoint c reviews its order queues $\mathcal{Q}_{i,t}$ for all $i \in \mathcal{I}$ with respect to recent and outstanding orders $Q_{i,t} \in \mathcal{Q}_{i,t}$, $t \leq t'$. By assumption, stock on hand $I_{c,t'}$ at time t' is scarce. Therefore, a rationing policy must be used for the first sequence of orders Q_{j,t^0} for all $j \in \mathcal{J}$ and t^0 which cannot be fully satisfied. Note that $\mathcal{J} \subset \mathcal{I}$ denotes the subset of stockpoints $j \in \mathcal{J}$ which triggered an order Q_{j,t^0} at time t^0. Define with q_{j,t^0} the fraction of remaining stock on hand which is distributed to stockpoint j. Stock on hand at stockpoint c is said to be allocated proportionally if the following condition holds for each fraction q_{j,t^0}:

$$q_{j,t^0} = Q_{j,t^0} \ / \sum_{k \in \mathcal{J}} Q_{k,t^0} \ . \tag{2}$$

Periodic Demand. We assume periodic iid customer demand $D_{i,t}$ for all $i \in \mathcal{I}$, $t \in \mathbb{N}$. It is assumed that an appropriate stationary cumulative distribution function (cdf) $F_i(d_{i,t})$ with probability density function (pdf) $f_i(d_{i,t})$ and $d_{i,t}$ a realization of $D_{i,t}$ exists. Mean demand and standard deviation per unit of time are denoted by μ_i and σ_i, respectively. Interarrival times λ_i of customer demands D_i at stockpoint $i \in \mathcal{I}$ are assumed to be one unit of time with probability 1.

Internal demand processes $D_{c,t}$ at stockpoint c are the result of previous customer demands, which themselves were transformed into batch orders $Q_{i,t}$ at stockpoints $i \in \mathcal{I}$ at each review $t = m \cdot T_i$, $m \in \mathbb{N}$ and $i \in \mathcal{I}$. Evidently, by means of an accurate analytical description of ordering processes $Q_{i,t}$ at stockpoints $i \in \mathcal{I}$ at time t it is possible to derive (gross) internal demand processes $D_{c,t}$ at time t.

Decomposition Scheme. We follow the wide-spread approach of analytically decomposing both echelons. The only direct analytical connection of estimates at the upper echelon with some at the lower echelon is achieved by introduction of waiting time expressions. Such a concept is known as the decomposition scheme and allows an easier modeling (see e. g. Deuermeyer and Schwarz (1981), Svoronos and Zipkin (1988)). For an illustration of the decomposition approach used by the authors the reader might refer to App. A.

Cycle Based Approximation. Since we initially assumed discrete monitoring, model quantities are measured at discrete points in time, i. e. our approximation is based on supporting points assigned to times $t \in \mathbb{N}$. Moreover, expected model quantities are estimated during a so-called *cycle*. Such an approach deviates from other analytical concepts, for example, the approach of van der Heijden (1993) who describes system state at an arbitrary point in time $t > 0$ and, next, derives approximate expressions by letting $t \longrightarrow \infty$. By the expression *cycle* the authors mean a specific time interval of constant length which can be observed after a system has reached statistical stationarity. Those cycles subsequently occur and share a regular pattern which can be described mathematically. In inventory theory, there exist two well-known (steady state) cycles: First, the *replenishment cycle* embraces system states between two subsequent deliveries. Second, the *reorder cycle* embraces system states between two subsequent replenishment orders. Both cycle approaches will be used here. For an illustration of the cycle based approximation scheme refer to App. B.

Basic Notation. In Tab. 1 basic notation used in this paper is listed. Additional definitions are introduced when needed.

4 Approach for the Upper Echelon

In the following, formulas for expected model quantities at the upper echelon are presented. Those quantities are required for performance measurement and optimization, for example. The upper echelon is estimated based on a *reorder cycle approach* instead of the replenishment cycle approach. This deviation from (conventional) replenishment cycle based approximations allows to model the impact of waiting times in a more accurate way. Expressions for stock on hand and backlog at the upper echelon at an arbitrary point in time t are presented in subsections 4.1 and 4.2, respectively. Moreover, to

Table 1. Basic notation

$\mathcal{I}^c$	set of all stockpoints, $\mathcal{I}^c = \{c, 1, \ldots, n\}$
$\mathcal{I}$	set of lower stockpoints, $\mathcal{I} = \{1, \ldots, n\}$
c	index of warehouse
PU	unit of product
TU	unit of time
MU	unit of money
S_i	order-up-to-level at stockpoint $i \in \mathcal{I}^c$ [PU]
T_i	reorder cycle at stockpoint $i \in \mathcal{I}^c$ [TU]
L_i	deterministic lead time to stockpoint $i \in \mathcal{I}^c$ [TU]
$W_{i,t}(S_c)$	average waiting time of a customer at stockpoint $i \in \mathcal{I}$ at time t [TU]
$H_{i,t}$	specific time interval of stockpoint $i \in \mathcal{I}^c$ at time t [TU]
$IP_{i,t}$	inventory position at stockpoint $i \in \mathcal{I}^c$ at time t [PU]
$NI_{i,t}(S_c, S_i)$	net inventory at stockpoint $i \in \mathcal{I}^c$ at time t [PU]
$I_{i,t}(S_c, S_i)$	stock on hand at stockpoint $i \in \mathcal{I}^c$ at time t [PU]
$B_{i,t}(S_c, S_i)$	backlog at stockpoint $i \in \mathcal{I}^c$ at time t [PU]
$Q_{i,t}$	order quantity of stockpoint $i \in \mathcal{I}^c$ at time t [PU]
$Q^-_{i,t}$	quantity delivered to stockpoint $i \in \mathcal{I}^c$ at time t [PU]
$D_{i,t}$	demand at stockpoint $i \in \mathcal{I}^c$ at time t [PU]
$D_i[a, b]$	gross demand during time interval $[a, b]$ at stockpoint $i \in \mathcal{I}^c$ [PU]
$X_{i,t}$	cumulative net demand until time t at stockpoint $i \in \mathcal{I}^c$ [PU]
$F_{i,t}$	cdf of cumulative net demand at time t at stockpoint $i \in \mathcal{I}^c$ [PU]
F_i	cdf of customer demand per unit of time at stockpoint $i \in \mathcal{I}$ [PU]
$DS_{c,t}(S_c)$	average duration of a shortage at stockpoint c [TU]
DS_c^0	target duration of a stock out required for stockpoint c
$\beta_i(S_c, S_i)$	fill rate at stockpoint $i \in \mathcal{I}$
β_i^0	target fill rate required for stockpoint $i \in \mathcal{I}$
$\gamma_i(S_c, S_i)$	γ-service measure at stockpoint $i \in \mathcal{I}$
γ_i^0	target γ-service required at stockpoint $i \in \mathcal{I}$
$LTI_i(S_c)$	lead time index for stockpoint $i \in \mathcal{I}$

measure internal reliability, it is recommended to apply the average duration of a stock out at the upper echelon, for which an easy-to-handle formula is introduced in subsection 4.3.

4.1 Stock on Hand

Two-Moment Approximation. Consider a reorder cycle of length T_c at the warehouse where transactions are monitored at discrete points in time $t = 1, \ldots, T_c$. To be able to approximate the average behavior of warehouse quantities we have to model three stochastic components which are (1) starting stock quantities, (2) cumulative gross internal demand and (3) quantities delivered to the warehouse. From the assumption of periodic review it is evident that all required quantities can be put down to the original, triggering

customer demand processes $D_{i,r}$ from stockpoint i at time r, $r > 0$. Now, a technique is searched to model the net impact of previous customer demand processes on system state at a specific point in time t. The authors found a two-moment based aggregation technique to be very useful. The two-moment approach used here can be characterized briefly as follows: For each of those stochastic components mentioned above determine the underlying, in general cumulative, demand process. Note that for each demand process important central or non-central moments must be derived. Then, interpreting the arrival of a delivery as negative demand it is possible to summarize those three single cumulative demand processes in a new variable – cumulative net internal demand – for which moments can be easily calculated on account of approximate independence of the three stochastic components (for a brief explanation see App. C; a more thorough foundation is given in Tüshaus and Wahl (1997)).

Starting Condition. In Tüshaus and Wahl (1997) it was motivated that stock quantities at the beginning of a (steady state) reorder cycle are conditioned by cumulative gross demand during m_c preceding reorder cycles where factor m_c is a function of both reorder cycle T_c and lead time L_c. One obtains the following definition of factor m_c at the warehouse:

$$m_c = \begin{cases} 1 + \lfloor L_c/T_c \rfloor & T_c < L_c \ \wedge \ L_c \bmod T_c \neq 0, \\ L_c/T_c & T_c < L_c \ \wedge \ L_c \bmod T_c = 0, \\ 1 & T_c \geq L_c \ . \end{cases} \tag{3}$$

In the following, the more comfortable formula $m_c = (L_c + (T_c - L_c) \bmod T_c)/T_c$ is preferred over (3).

Note that on account of the nested schedule assumption (1) cumulative gross internal demand $D((u-1)T_c; uT_c]$, $u \in \mathbb{N}$, at the warehouse equals the sum of ordering processes $Q_{i,t}$, $(u-1)T_c < t \leq uT_c$, for all $i \in \mathcal{I}$. From the assumption of (T_i, S_i)-policies it follows that $Q_{i,t}$ results from cumulative gross customer demands $D(0; T_i]$ at retailer i during its reorder interval of length T_i. Using this relation yields the two central moments $\mu_c(\cdot)$ and $\sigma_c(\cdot)$ of cumulative gross internal demand conditioning starting stock quantities:

$$\mu_c(m_c T_c) = m_c T_c \mu_c \tag{4}$$

with $\mu_c = \sum_{i=1}^{n} \mu_i$ and

$$\sigma_c(m_c T_c) = \sqrt{m_c T_c}\, \sigma_c \tag{5}$$

with $\sigma_c^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}$ where σ_{ij} is the covariance between stockpoints i and j.

Cyclical Demand. The second stochastic component is given by current cumulative gross internal demand $D_c(0; t]$ until time t, $0 < t \leq T_c$. The magnitude

of variation in demand $D_c(0;t]$ depends on the relative planning frequency k_i at the retailers. In this context, let us define the indicator function of a review $l(t,T_i)$ as follows:

$$l(t,T_i) = \begin{cases} 1 & t \bmod T_i = 0, \\ 0 & \text{otherwise} \ . \end{cases} \tag{6}$$

Obviously, moments of cumulative gross internal demand $D_c(0;t]$ until time t are easily obtained when taking expectations of retailer orders weighted by $l(t,T_i)$:

$$\mu_c(t) = \sum_{r=1}^{t} \sum_{i=1}^{n} l(t,T_i) T_i \mu_i = \sum_{i=1}^{n} \lfloor t/T_i \rfloor T_i \mu_i \tag{7}$$

and

$$\sigma_c(t) = \sqrt{\sum_{r=1}^{t} \sum_{i=1}^{n} \sum_{j=1}^{n} l(t,T_i) l(t,T_j) \sqrt{T_i T_j} \sigma_{ij}} \ . \tag{8}$$

Impact of a Delivery. Finally, let us consider the third stochastic component, i. e. the impact of a quantity $Q^{-}_{c,t}$ possibly delivered at time t. By assumption, orders from the warehouse are always satisfied by the production location. Therefore, if a delivery arrives $Q^{-}_{c,t}$ equals the quantity $Q_{c,t'}$ of the triggering order. It is easy to see that the quantity ordered can be expected to be $\mu_c(T_c) = \mathrm{E}(Q_{c,t'}) = T_c \mu_c$ with standard deviation $\sigma_c(T_c) = \mathrm{SD}(Q_{c,t'}) = \sqrt{T_c}\sigma_c$.

Aggregating Moments. From the assumption of independence of those three stochastic components it follows that aggregate moments $\mu_c[t]$ and $\sigma_c[t]$ of cumulative net internal demand $X_{c,t}$ at time t, $0 < t \leq T_c$, can be easily summarized. To begin with, let us denote by $t_i^\star = T_c - (T_c - L_c) \bmod T_c$ for all $i \in \mathcal{I}^c$ the time a delivery arrives at stockpoint i (see Tüshaus and Wahl (1997) for a derivation of $t_i^\star$). Moreover, define the indicator function $l(t,t_i^\star)$ for the timing of a delivery as

$$l(t,t_i^\star) = \begin{cases} 1 & t \geq t_i^\star, \\ 0 & t < t_i^\star \end{cases} \tag{9}$$

where (9) can be rewritten $l(t,t_i^\star) = \lfloor t/t_i^\star \rfloor$, alternatively. Then, central moments of $X_{c,t}$ at time t read as follows:

$$\mu_c[t] = (m_c - \lfloor t/t_c^\star \rfloor) \cdot \mu_c(T_c) + \mu_c(t) \tag{10}$$

and

$$\sigma_c[t] = \sqrt{(m_c - \lfloor t/t_c^\star \rfloor) \cdot \sigma_c(T_c)^2 + \sigma_c(t)^2} \ . \tag{11}$$

As motivated in App. C an adequate pdf is fitted to moments $\mu_c[t]$ and $\sigma_c[t]$.

Stock on Hand Formula. Now, let us denote with $f_{c,t}(x_{c,t})$ an appropriately fitted pdf of cumulative net internal demand $X_{c,t}$ at time t and $F_{c,t}(x_{c,t})$ the corresponding cdf. Expressions for stock on hand $I_{c,t}(\cdot)$ at time t, $1 \leq t \leq T_c$, during a reorder cycle are based on the well-known one-period formulas which are used in Newsboy-style inventory problems (see e. g. Hadley and Whitin (1963), pp. 297). One obtains:

$$I_{c,t}(S_c) = \begin{cases} \int_0^{S_c} (S_c - x_{c,t}) f_{c,t}(x_{c,t})\, dx_{c,t} & \mu_c[t] > 0 \\ S_c & \mu_c[t] = 0 \wedge t \geq t_c^\star \ . \end{cases} \tag{12}$$

The second case of (12) can be explained as follows: Simple analysis shows that for inventory systems with $T_c \geq L_c$, $T_c > 1$, and where, furthermore, the first review at locations $i \in \mathcal{I}$ realizes after the arrival of a delivery at location c at time $t_c^\star$, moments for stochastic demands $\mu_c[t]$ and $\sigma_c[t]$ might vanish. Obviously, if $\mu_c[t_c^\star] = 0$ location c receives a delivery and no current demand can be observed. Hence, stock on hand will amount to $I_{c,t} = S_c$. Of course, for times $t \geq t_c^\star$ having $\mu_c[t] = 0$ stock on hand remains unchanged, that is $I_{c,t} = I_{c,t-1} = S_c$.

Finally, average stock on hand $\bar{I}_c(S_c)$ during a reorder cycle at the warehouse is obtained as follows:

$$\bar{I}_c(S_c) = T_c^{-1} \cdot \sum_{t=1}^{T_c} I_{c,t}(S_c) \ . \tag{13}$$

4.2 Backlog

In analogy to the stock on hand formula (12), expected backlog at time t, $t = 1, \ldots, T_c$, can be determined as follows:

$$B_{c,t}(S_c) = \begin{cases} \int_{S_c}^{\infty} (x_{c,t} - S_c) f_{c,t}(x_{c,t})\, dx_{c,t} & \mu_c[t] > 0 \\ 0 & \mu_c[t] = 0 \wedge t \geq t_c^\star \ . \end{cases} \tag{14}$$

Again, an estimate for average backlog $\bar{B}_c(\cdot)$ during a reorder cycle at the warehouse is given by:

$$\bar{B}_c(S_c) = T_c^{-1} \cdot \sum_{t=1}^{T_c} B_{c,t}(S_c) \ . \tag{15}$$

4.3 Duration of a Stock Out

As an internal performance measure the average duration of a *stock out* (abbr.: SO) $DS_c(S_c)$ is suggested. We will find an approximate expression for

$DS_c(S_c)$ by means of estimates $DS_{c,t}(S_c)$ which can be observed in steady state at times $t = 1, \ldots, T_c$. During a reorder interval at the warehouse, orders $Q_{i,r}$ from retailers might face a stock out at times r, $(u-1)T_c < r \leq uT_c$, $u \in \mathbb{N}$. The duration of such a stock out situation depends, first, on both arrival dates and quantities of future deliveries, and, second, on the time span until a next order can be placed by the warehouse. In the following, a delivery $Q^-_{c,t}$ to the warehouse is assumed to be large enough to satisfy all orders outstanding from review periods preceding the current one. Clearly, for (very) low values S_c cumulative backlog stemming from previous review periods won't be fully eliminated and, hence, the assumption of large $Q^-_{c,t}$, which is required for $DS_c(\cdot)$ to hold, is no longer valid.

To begin with, let us consider a 1-warehouse, 1-retailer system where an order $Q_{i,t_{so}}$ is triggered at time t_{so} and induces a stock out at the warehouse. In Fig. 2 a possible scenario is illustrated.

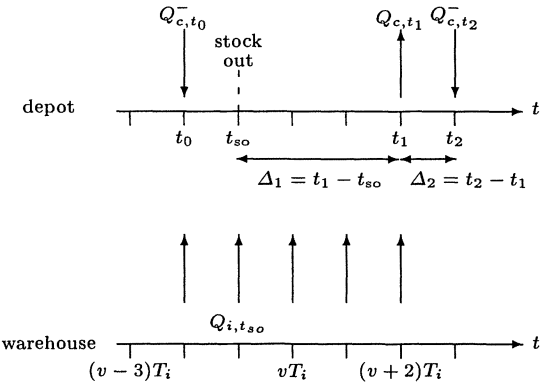


Fig. 2. Duration of a stock out at the upper echelon

Obviously, in our example, order $Q_{i,t_{so}}$ can be at most partially fulfilled. Hence, the remaining part of $Q_{i,t_{so}}$ is backlogged and waits until time t_2 when a new delivery Q^-_{c,t_2} comes in, which itself was triggered at time $t = t_2 - L_c$. Expressions $DS_{c,t}$ are determined as follows: First, the expected duration of a stock out – denoted as $\Delta_{c,t} = \Delta_1 + \Delta_2$ (see Fig. 2) – for an order arriving at time t is quantified. Second, the probability for an order facing a stock out at time t – denoted as $P(\mathrm{SO} = \mathrm{true}, t)$ – is estimated.

Simple analysis shows that the nominal duration of a stock out $\Delta_{c,t}$ at the warehouse at time t is obtained by:

$$\begin{aligned} \Delta_{c,t} &= l(t, t^\star_c) \cdot T_c + t^\star_c - t \\ &= \lfloor 1 + t/t^\star_c \rfloor \cdot T_c + (T_c - L_c) \mod T_c - t \ . \end{aligned} \tag{16}$$

From Equation (16) it can be easily seen that $\Delta_{c,t}$ is restricted onto the range $[0; T_c]$ (which follows from the assumption of sufficiently high quantities $Q_{c,t_c^\star}$ delivered). Hence $\Delta_{c,t}$ yields only approximate results, in general.

Next, we approximate the probability of a stock out $P(\mathrm{SO} = \mathrm{true}, t)$ by means of cdfs $F_{c,t}(S_c)$ which denote the probability of cumulative net internal demands to be at most S_c, i. e. the theoretical maximum of stock on hand. One obtains:

$$P(\mathrm{SO} = \mathrm{true}, t) \approx 1 - F_{c,t}(S_c) \ . \tag{17}$$

Now, an expression for the expected duration of a stock out occuring at time t, $1 \leq t \leq T_c$, is given (for sufficiently high S_c) with

$$DS_{c,t}(S_c) = (\lfloor 1 + t/t_c^\star \rfloor \cdot T_c + (T_c - L_c) \mod T_c - t) \cdot (1 - F_{c,t}(S_c)) \ . \tag{18}$$

Denote with $\mathcal{R} \subset \mathcal{T} = \{1, \ldots, T_c\}$ the subset of points in time when orders $Q_{i,t}$ from retailers arrive at the warehouse. Then, we have an expression for the average duration of a shortage $DS_c(S_c)$ at the warehouse:

$$DS_c(S_c) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} DS_{c,r}(S_c). \tag{19}$$

Obviously, expression $DS_c(S_c)$ can be used as an internal performance measure given a target level of, say, $DS_c^0 = v$, $v \geq 0$.

5 Approach for the Lower Echelon

In the following, the lower echelon is estimated by means of the *replenishment cycle approach*. An expression for stock on hand and backlog at a retailer at an arbitrary time during a replenishment cycle is derived. Then, three performance measures are presented: the fill rate, γ-service measure and the so-called lead time index. They can be used as to measure customer satisfaction and/or overall systems performance.

5.1 Customer Demands

Remember that we assumed iid customer demand $D_{i,t}$ per unit of time with stationary cdf $F_{i,t}(d_{i,t}) = F_i(d_{i,t})$ and pdf $f_{i,t}(d_{i,t}) = f_i(d_{i,t})$ for all $i \in \mathcal{I}$ and $t \in \mathbb{N}$. Furthermore, let us denote by $F_i^{(a)}(x_{i,t})$ the cdf which results from an a-fold convolution of cdf $F_i(d_{i,t})$ with $f_i^{(a)}(x_{i,t})$ the corresponding pdf. Obviously, it is assumed that closed-form convolutions exist and analytical results are numerically tractable. In the contrary case, it is recommended to fit a substitutive pdf on moments $\mu_i^{(a)}$ and $\sigma_i^{(a)}$ derived from the sum of a iid random variates. This yields an approximate expression for the convolved density.

In the usual case of customer demands being modeled by means of non-negative variables, there exist some excellent methods to derive quite accurate (substitutive) pdfs and cdfs which, for example, represent mixtures of tractable distributions (see e. g. Tijms (1986)). Besides, Badinelli (1996) recently summarized promising methods for approximating convolutions of arbitrary probability densities which are based on orthogonal polynomials. For the following analysis, pdf $f^{(a)}(\cdot)$ is assumed to be either exact or approximate.

5.2 Stock on Hand

The Impact of Waiting Times. Estimates for expected stock on hand $I_{i,t}(\cdot)$ at retailer i at time t, $1 \leq t \leq T_i$, are required for calculation of average holding costs during an arbitrary replenishment cycle in steady state. Note that, since lead times L_i are deterministic, the length of a replenishment cycle is constant and, hence, at least intuitively, depicting system states $t = 1, \ldots, T_i$ during one single steady state cycle should yield quite accurate results.

Indeed, if delays, induced by a possible stock out at the warehouse, are negligible, such an approach might be satisfactory. But, since we assumed the warehouse to have restricted stocking capacities, customer orders arriving at time t at the lower echelon might face an additional, in general non-negligible waiting time $W_{i,t}$ until complete satisfaction. Therefore, customer waiting times are modeled explicitly, here. Doing that it is possible to couple system states at both echelons which is favorable with respect to the accuracy of estimates since, then, the impact of a stocking policy at the upper echelon is reflected at the lower echelon. Simultaneously, analytical problems arise when including $W_{i,t}$.

Subcycles. In general, the reorder cycle length T_c is expected to exceed the length T_i of a replenishment cycle at retailer i. Consequently, a customer order $D_{i,t}$ arriving at retailer i at time $t = 1, \ldots, T_i$ during an arbitrary replenishment cycle must be characterized furthermore by its relative location on the time scale at the warehouse which is $r = 1, \ldots, T_c$. Note that the integer multiple $k_i = T_c/T_i$ defined (1) gives the number of reorder cycles at the retailer being covered by one reorder cycle at the warehouse. Thus, a customer order $D_{i,t}$ is characterized by a pair (t, s) where t is conditioned on s, i. e. $(s-1)T_i \leq t \leq sT_i$, $s = 1, \ldots, k_i$. Let us denote with *subcycle* a reorder cycle at an arbitrary retailer covered by a reorder cycle at the warehouse. Next, let $W^s_{i,t}$ be the average waiting time of a customer order arriving at time t in subcycle s. For the moment, let us skip an explicit definition of expression $W^s_{i,t}$. In later analysis we will present approximate expressions for customer waiting times (see subsection 5.5).

Average Stock on Hand. Again, all estimates $I^s_{i,t}(\cdot)$ for stock on hand at time t during subcycle s are based on single-period formulas that arise in Newsboy-problems. As an extension of the approach in Rogers and Tsubakitani (1991)

we model the impact of waiting times in a simple, but, in general, highly accurate manner: Instead of deriving the distribution function of waiting times, only supporting points $W_{i,t}^s$ are modeled which represent estimates of the average waiting time observable for an arbitrary customer arriving at time t, $1 \leq t \leq T_i$, in subcycle s. One obtains the following formula of average stock on hand $\bar{I}_i(\cdot)$ for retailer i:

$$\bar{I}_i(S_c, S_i) = \frac{1}{k_i T_i} \cdot \sum_{s=1}^{k_i} \sum_{t=1}^{T_i} \int_0^{S_i} (S_i - x_{i,t}) f_i^{(L_i + W_{i,t}^s + t)}(x_{i,t})\, dx_{i,t} \ . \tag{20}$$

Note that, for the lower echelon, expressions for stock on hand are functionally influenced by both local and central control parameters S_i and S_c, respectively. The influence of S_i can be seen directly whereas the impact of S_c is comprised in the waiting time expression $W_{i,t}$ although, in favor of an easier notation, this functionality is not depicted here by writing $W_{i,t}(S_c)$.

5.3 Backlog

Service constraint formulas require the calculation of expected backlog $B_{i,t}^s(\cdot)$ at retailer i at time t, $1 \leq t \leq T_i$, in subcycle s. Again, applying the well-known Newsboy-formulas, one obtains

$$B_{i,t}^s(S_c, S_i) = \int_{S_i}^{\infty} (x_{i,t} - S_i) f_i^{(L_i + W_{i,t}^s + t)}(x_{i,t})\, dx_{i,t} \ . \tag{21}$$

Note the fundamental relation $B_t = I_t - NI_t$ between expected backlog B_t, stock on hand I_t and net stock NI_t at time t. Additionally, for non-negative customer demand, net stock equals $NI_{i,t}(S_c, S_i) = S_i - \mathrm{E}(X_{i,t})$ which yields:

$$B_{i,t}^s(S_c, S_i) = I_{i,t}^s(S_c, S_i) - S_i + \mathrm{E}(X_{i,t}) \ . \tag{22}$$

5.4 Service Constraints

Fill Rate. A wide-spread performance measure is given by the fill rate β which can be defined as the long-run fraction of demand satisfied directly from stock on hand. Let us denote the target fill rate required at retailer i by β_i^0, $0 \leq \beta_i^0 \leq 1$. A well-known approximation for the fill rate in subcycle s is obtained using $\beta_i^s(S_c, S_i)$ (see e. g. de Kok (1991b) for a thorough analysis of the fill rate and alternative performance measures):

$$\beta_i^s(S_c, S_i) = 1 - \frac{B_{i,T_i}^s(S_c, S_i) - B_{i,0}^s(S_c, S_i)}{T_i \cdot \mu_i} \tag{23}$$

with $B_{i,0}^s(S_c, S_i)$ expected backlog immediately after the arrival of a delivery, i. e. backlog at the beginning of a replenishment cycle s, and with $B_{i,T_i}^s(S_c, S_i)$

expected backlog just before the next delivery at the end of replenishment cycle s. An obvious property of $\beta_i^s(S_c, S_i)$ is that double counting of positive backlog is avoided by subtracting an expression $B_{i,0}^s(S_c, S_i)$. Expressions of expected backlog are obtained according to Eqn. (21). When using $\beta_i^s(S_c, S_i)$, a decision maker requires retailer i to guarantee a predetermined service level at system states in subcycles s where stock on hand reaches its theoretical minimum.

γ-Service Measure. The γ-service measure represents another way to measure system performance. The γ_i-measure can be described as the fraction of average demand per unit of time exceeding average backlog per unit of time at retailer i. Let γ_i^0 be the target γ-measure required for retailer i. Deriving an expression for $\gamma_i(S_c, S_i)$ yields:

$$\gamma_i(S_c, S_i) = \frac{\bar{B}_i(S_c, S_i)}{\mu_i} \quad . \tag{24}$$

Note the average backlog $\bar{B}_i(S_c, S_i)$ is obtained in analogy to Eqn. (20). When using $\gamma_i(S_c, S_i)$ as a performance measure, the average readiness of retailer i to satisfy an incoming (average) customer order of amount μ_i is of interest. Evidently, values $\gamma_i \geq 1$ seem to be disadvantageous because an arbitrary customer is expected to wait always a positive time span.

5.5 Waiting Times

Approximate Waiting Times. In the preceding subsections formulas for model quantities at the lower echelon have been presented with values $W_{i,t}^s$ still undefined. In the following, $W_{i,t}^s$ will be approximated applying standard results from queueing theory. Imagine a single-server system with an infinite queue size, with customer interarrival times and customer service times following a continuous Markov process. In such a $M|M|1$-system, the average waiting time of a customer can be quite accurately expressed: the average waiting time equals the fraction of unserved customers with respect to the average number of incoming customers (see Little (1961)).

In a distribution system, the warehouse represents the server and orders from retailers can be seen as single customers. Define the number of unserved orders by $NB_{c,t}$ and the number of incoming retailer orders $NO_{c,t}$ at time t. Applying Little's formula yields the following expression for the average waiting time $\mathrm{E}(W_t)$:

$$\mathrm{E}(W_t) = \frac{\mathrm{E}(NB_{c,t})}{\mathrm{E}(NO_{c,t})} \quad . \tag{25}$$

Note that, depending on the demand processes, expression $\mathrm{E}(W_t)$ underestimates or overestimates the actual waiting time. In case of non-unit demand, we have to approximate expressions $\mathrm{E}(NB_{c,t})$ and $\mathrm{E}(NO_{c,t})$. The expected

number of unserved orders at time t is substituted by expected backlog $B_{c,t}(\cdot)$ and the number of incoming orders is replaced by the mean demand rate $\mu_{c,t}$. One obtains an approximate formula $W[t]$:

$$W[t] = \frac{B_{c,t}(S_c)}{\mu_{c,t}} \tag{26}$$

where $\mu_{c,t} = \sum_{i=1}^{n} l(t, T_i) T_i \mu_i$. Hence, the waiting time of a retailer order will be determined by means of an average value $W[t]$ which can be observed during a replenishment cycle. As mentioned before, during a reorder interval at the warehouse level, retailer orders $Q_{i,t}$, $1 \leq t \leq T_c$, arrive according to their planning frequency given by $1/T_i$. In general, the planning frequency at the lower echelon is expected to be higher, i. e. $T_c > T_i$ for some $i \in \mathcal{I}$. Note that we have assumed T_c to be an integer multiple k_i of T_i to ensure stationarity of aggregated demand processes. In such a case, one reorder interval at the warehouse covers k_i reorder intervals (read: subcycles) at a retailer. In Fig. 3 an example with $k_i = 3$ subcycles is illustrated for a 1-warehouse, 1-retailer system.

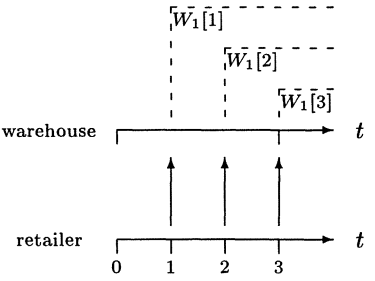


Fig. 3. Waiting time regions for retailer 1

Obviously, each reorder cycle is defined by a specific waiting time region where $W_1[s]$ holds. Now, for retailer i in a general 1-warehouse, n-retailer system, let us specify expression $W_i[s]$ given reorder cycle T_i and the number of subcycles k_i:

$$W_i[s] = W[sT_i], \quad s = 1, \ldots, k_i \tag{27}$$

where $W[sT_i]$ is obtained from (26).

Modified Model Quantities. Formulas for stock on hand, backlog and service measures are easily modified by substituting general expressions $W_{i,t}^s$ with

$W_i[s]$ which yields:

$$I_{i,t}^{s}(S_c, S_i) = \int_0^{S_i} (S_i - x_{i,t}) f_i^{(L_i + W[sT_i] + t)}(x_{i,t})\, dx_{i,t} \tag{28}$$

for expected stock on hand at time t in subcycle s. Expected Backlog now reads:

$$B_{i,t}^{s}(S_c, S_i) = \int_{S_i}^{\infty} (x_{i,t} - S_i) f_i^{(L_i + W[sT_i] + t)}(x_{i,t})\, dx_{i,t} \ . \tag{29}$$

Modifying $\beta_i^s(S_c, S_i)$ and $\gamma_i(S_c, S_i)$ is straightforward and, hence, omitted here. The subcycle oriented approach sketched above seems to overcome some of the main arguments which are brought up against a direct use of an expected waiting time expression based on Little's formula in models with periodic review (see e. g. van der Heijden (1993), pp. 108): First, according to the above formulas, each retailer i has individual average waiting time expressions resulting from its ordering frequency during one long reorder cycle at the warehouse. Second, since one reorder cycle at the warehouse may cover several replenishment cycles at a retailer, the latter faces a long cycle with several subcycles having an individual average waiting time estimate, each. By that, the dependence of the actual waiting time on system states in preceding subcycles is analytically reflected in a reasonable manner. For a detailed discussion of the use of Little's formula in statistical inventory control refer to Diks et al. (1996), for example.

Effective Lead Times. It is easy to see that a straightforward estimate for the average waiting time $\overline{W}_i$ of an arbitrary order of retailer i is given by:

$$\overline{W}_i = \frac{1}{k_i} \cdot \sum_{s=1}^{k_i} W[sT_i] \ . \tag{30}$$

Consequently, an approximate formula for the effective average lead time $\bar{L}_i^{\text{eff.}}$ to retailer i yields:

$$\bar{L}_i^{\text{eff.}} = L_i + \overline{W}_i \ . \tag{31}$$

Lead Time Index. We suggest an additional performance measure, the lead time index LTI, which measures the relative magnitude of average delays with respect to lead times and which is defined as follows:

$$LTI_i(S_c) = \bar{L}_i^{\text{eff.}} / L_i \ . \tag{32}$$

Analyzing $LTI_i(S_c)$, first, gives a hint at the economic importance of delays with regard to objectives like customer satisfaction or time to market, and, second, permits a decision maker to judge a more accurate modeling waiting time to be valuable or not for system analysis.

6 Computational Results

The approximate two-echelon distribution model has been tested for a large set of problems. It has been found that the overall validity of analytical approximations is excellent for both the upper and lower echelon.

To validate analytical results an analogous discrete event simulation model was used. In one simulation run 50'000 customer orders per retailer were generated. Following classical heuristics for determining the length of the warm-up phase, the first 10'000 customers were skipped with regard to statistical evaluations. Several simulation runs were carried out with different starting values for the random number generator. We used a random number generator proposed by Härtel (1994) which guarantees a sufficient length of period. The confidence level for mean estimates was set to 95 %.

6.1 Some Preliminaries

All customer populations are drawn from normal populations. Therefore, the occurence of negative customer demands has positive probability which, at a first glance, seems unrealistic. However, negative demands can be interpreted ecomonically, too, insofar they could represent customers preferring to return delivered goods. Furthermore, even if such transactions can be excluded for a specific distribution system, it depends on the coefficient of variation to make a normal approximation reasonable. Let us claim a maximum cumulative probability of $\alpha = 0.025$ for realisations of negative demand. Assuming normally distributed customer demands per unit of time, the maximum coefficient of variation $\mathrm{CV} = \sigma/\mu$ inducing $\alpha = 0.025$ is obtained as follows:

$$\Phi^{-1}(-\mathrm{CV}^{-1}) = \Phi^{-1}(-\frac{\mu}{\sigma}) = -1.9599$$

with $\Phi(\cdot)$ the standard-normal cdf. Obviously, for mean demand rates μ being at least $1.96\times$ the standard deviation σ, i.e. for values $\mathrm{CV} \leq 0.51$, the normal model works in 97.5 % of all cases. Although in earlier studies we truncated the normal approximation for negative demands, in the following test studies the full support of normally distributed variates was included. In this manner, it was possible to conlude if the distribution model works well for high values CV, too. (Note that by introducing truncation, some additional methodological biases occur when comparing simulated and analytical quantities.)

In subsection 6.2, the main statistical results for both echelons are listed. To measure the robustness of our approximate estimates we assumed policies (T_c, S_c) to have *low* values S_c, in general, which tends to increase waiting times, and moderate policies S_i for all $i \in \mathcal{I}$. By *low* we mean that, first, values S_c are determined given *low* coefficients of variation CV_c of aggregate demand and negative safety factors r_c. Obviously, in all cases, safety stock at

the upper echelon is negative. Second, safety stock for cases with low coefficients of variation applies to cases with higher values CV_c, too, which tends to yield decreased system performance for test problems with high variation in demand. In this context the reader should note that previous numerical studies have shown even the *optimized* upper echelon to hold negative safety stock in many cases (see Graves (1996) for similar results).

All order-up-to-levels S_i have been determined based on a simple safety stock planning approach which works as follows: To begin with, note that S_i can be decomposed in a lot-sizing problem (determine Q_i) and a risk time coverage problem (determine s_i), i. e. :

$$S_i = Q_i + s_i \ . \tag{33}$$

Now, fix Q_i by the average demand during one reorder cycle of length T_i. This yields $\bar{Q}_i$:

$$\bar{Q}_i = T_i \cdot \mu_i \tag{34}$$

Next, the coverage problem can be further separated in an average lead time demand and a safety stock planning problem which yields $\bar{s}_i$:

$$\bar{s}_i = L_i \cdot \mu_i + r_i \cdot \sqrt{T_i + L_i} \cdot \sigma_i \tag{35}$$

with $r_i \in \mathbb{R}$ a safety factor. The impact of waiting times $W_{i,t}$ on system performance has been omitted here which might yield underestimated safety stocks at the lower echelon. For the lower echelon, values S_i were calculated given a moderate value $\mathrm{CV}_i = 0.25$ and $r_i \approx 2$. For the upper echelon, values S_c were calculated given $\mathrm{CV}_c = 0.25$ and $r_c < 0$. All values S_i were rounded off, additionally.

6.2 Numerical Results

Problem Set The test problems for identical retailers are based on the following input data:

1. number of identical retailers: $n = 5$,
2. customer demands: $\mu_i = 100$, $\sigma_i = (5, 85)$,
3. lead times: $L_i = (1, 2, 3, 4)$,
4. reorder intervals: $T_i = (1, 2, 3, 4)$.

In the following tables, some results are listed for selected performance measures like, for example, mean physical stock, backlog and the duration of a stock out at the upper echelon. To simplify the notation we omitted braces $(\cdot)$ in all analytical performance measures. In tables 2, 3 and 4 the 95%-ranges of simulated values are listed. Note that deviations between analytical and simulated estimates are expressed as a percentage $\Delta(\%) = 100 \cdot (simulated\_value - approximate\_value)/simulated\_value$ in choosing the nearest simulated value embraced in the corresponding (empirical) confidence interval.

Numerical Findings We summarize the following numerical results: Even for high coefficients of variation CV_i of customer demands D_i estimates for stock on hand, backlog, effective lead time and the duration of a stock out are approximated with a high degree of accuracy (these findings are valid for the fill rate and the γ-measure, too). It seems that the (absolute) frequency of ordering $1/T_i$ has no impact on the approximation validity. But, the relative length of lead times L_i/T_i influences the accuracy of estimates, i. e. one can observe that the accuracy slightly moves down with increasing L_i/T_i. This might be explained by the fact that the probability of *imbalances* at the lower echelon can be expected to increase, too. By *imbalance* we mean a situation of unbalanced stock at the lower echelon inducing insufficient customer service at some stockpoints and too much stock on hand at other stockpoints.

The occurence of imbalances tends to be largely supported by, first, the non-existence of more sophisticated rationing policies than the myopic proportional rule implemented for our model, and, second, the considerably low values S_c determined for each test problem. In the current literature about multi-echelon inventory systems of the *push-type* several rationing strategies have been succesfully implemented (see e. g. van der Heijden et al. (1996) for an overview). Until now no such rationing strategies are known to the authors for systems of the *pull-type*. It therefore might be fruitful to test other rationing strategies that take into account the frequency of ordering, the relative length of leadtimes and coefficients of variation in customer demand.

All approximate formulas work well even for low order-up-to-levels S_c. At the same time, it must be mentioned that in case of extremely low order-up-to-levels S_c at the warehouse, estimates for the mean duration of a stock out DS_c and average waiting times $\bar{W}_i$ become worse (see Tab. 4). Now, it depends on the relative importance of waiting times that such inaccuracies have a significant impact on estimates for model quantities at the lower echelon. The relative importance of possible waiting times can be measured by the lead time index LTI_i suggested above (see Equ. (32)). To see this, let us consider the six cases with highest values LTI_i, i. e. $(T_i, T_c, L_i, L_c) = (1, 4, 1, 3)$ given $\sigma_i = 85$, $(T_i, T_c, L_i, L_c) = (1, 4, 1, 3)$ given $\sigma_i = 5$ and so on. Those six cases altogether make up 87.5 % of cases with relative deviations for stock on hand and 42.9 % of cases with relative deviations for backlog being both significantly different from zero.

Next, let us analyze the obviously bad estimate for DS_c when the inventory system is run on a $(T_i, T_c, L_i, L_c) = (1, 1, 2, 4)$-basis. Now, for a low value $S_c = 1500$ the formula $DS_c(\cdot)$ presented in Equ. (19) does not work any more. But, this does not come with surprise since a use of DS_c requires the assumption of previous backlog being eliminated by a current delivery to be approximately fulfilled. Clearly, this is not the case for very low order-up-to-levels S_c. For very low values S_c, backlog can be expected to cumulate over more than just one review period. For the given example, values DS_c were simulated for varying order-up-to-levels S_c to approximate the cut-off point which ensures Equ. (19) to hold. It was found that the cut-off point can be

located around values S_c slightly more than 1500 units, say 1540 or 1580. To see this, consider the sequence of values for configurations in the range $S_c = [1400; 2100]$ listed in Tab. 5. Evidently, all of our formulas work well for low internal system performance at the upper echelon, i. e. , for $\gamma_c \approx 0.84$ which says that approximately 84 % of incoming customer orders won't be satisfied immediately. Note that there is a solid region for the duration of a stock out with $S_c = [1580; 2000)$ and $DS_c(S_c) = 1$ both preceded and succeeded by a non-linear range with rapidly decreasing values DS_c.

In Fig. 4 fractions of relative deviations in *total* stock underline the high reliability of the analytical model. Further analysis shows that for 90 % of the cases with $\sigma_i = 5$, estimates of stock on hand are nearly exact (see percentage deviations in Tab. 2). Figure 5 displays the distribution of fractions for *total* backlog and confirms the high degree of backlog formulas. Finally, a brief analysis of Tab. 2 and 3 yields that in many of the observed cases errors in estimation at the upper and lower echelon compensate each other which yields far smaller deviations in systemwide expressions than in stock specific expressions (see Fig. 4 and 5).

Table 2. Estimated stock on hand

T_i	T_c	L_i	L_c	σ_i	S_i	S_c	$\bar{I}_i^{\text{anal.}}$	$\bar{I}_i^{\text{sim.}}$	$\Delta(\%)$	$\bar{I}_c^{\text{anal.}}$	$\bar{I}_c^{\text{sim.}}$	$\Delta(\%)$
1	1	1	1	5	270	600	70	70	0.0	100	[100;101]	0.0
1	1	1	1	85	270	600	86	[85;87]	0.0	136	[135;137]	0.0
1	1	2	4	5	390	1500	1	1	2.3	0	0	0.0
1	1	2	4	85	390	1500	62	[62;66]	0.9	17	[17;19]	1.8
1	2	2	1	5	390	900	80	80	0.0	200	200	0.0
1	2	2	1	85	390	900	104	[103;106]	0.0	233	[232;235]	0.0
1	2	2	4	5	390	1800	25	25	0.0	0	0	0.0
1	2	2	4	85	390	1800	66	[66;70]	0.4	41	[40;43]	0.0
1	3	1	4	5	270	2400	40	40	0.0	133	133	0.0
1	3	1	4	85	270	2400	64	[65;68]	1.7	192	[199;204]	3.6
1	3	2	3	5	390	1800	47	47	0.0	100	100	0.0
1	3	2	3	85	390	1800	81	[80;84]	0.0	138	[137;140]	0.0
1	4	1	1	5	270	1500	52	52	0.0	377	377	0.0
1	4	1	1	85	270	1500	75	[75;77]	0.0	413	[419;423]	1.4
1	4	1	3	5	270	2100	35	35	0.0	175	175	0.0
1	4	1	3	85	270	2100	61	[61;63]	0.8	214	[212;215]	0.0
2	2	1	1	5	390	900	130	130	0.0	450	450.00	0.0
2	2	1	1	85	390	900	141	[138;141]	0.0	482	[482;483]	0.0
2	2	2	3	5	500	1500	100	100	0.0	250	250	0.0
2	2	2	3	85	500	1500	131	[127;132]	0.0	260	[259;262]	0.0
2	4	2	4	5	500	3500	150	150	0.0	1'000	1'000	0.0
2	4	2	4	85	500	3500	166	[163;166]	0.0	1'017	[1'014;1'018]	0.0

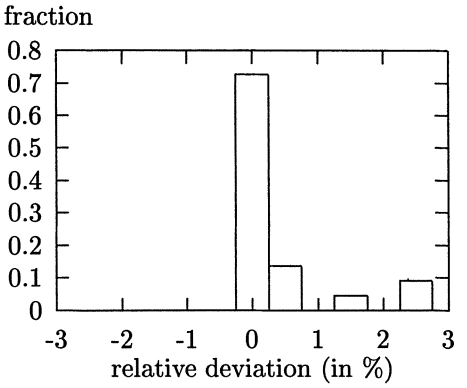


Fig. 4. Distribution of relative deviations in total physical stock

Table 3. Estimated backlog

T_i	T_c	L_i	L_c	σ_i	S_i	S_c	$\bar{B}_i^{\text{anal.}}$	$\bar{B}_i^{\text{sim.}}$	$\Delta(\%)$	$\bar{B}_c^{\text{anal.}}$	$\bar{B}_c^{\text{sim.}}$	$\Delta(\%)$
1	1	1	1	5	270	600	0	0	0.0	0	0	0.0
1	1	1	1	85	270	600	24	[23;24]	−0.5	36	[36;37]	0.0
1	1	2	4	5	390	1500	11	11	0.0	500	500	0.0
1	1	2	4	85	390	1500	75	[73;77]	0.0	517	[513;522]	0.0
1	2	2	1	5	390	900	0	0	0.0	50	50	0.0
1	2	2	1	85	390	900	31	[30;31]	0.0	83	[82;84]	0.0
1	2	2	4	5	390	1800	25	25	0.0	450	450	0.0
1	2	2	4	85	390	1800	74	[74;77]	0.2	491	[494;504]	0.8
1	3	1	4	5	270	2400	17	17	0.0	233	[233;234]	0.0
1	3	1	4	85	270	2400	52	[51;54]	0.0	292	[285;293]	0.0
1	3	2	3	5	390	1800	17	17	0.0	300	300	0.0
1	3	2	3	85	390	1800	58	[57;61]	0.0	338	[342;350]	1.4
1	4	1	1	5	270	1500	8	[7;8]	0.0	127	127	0.0
1	4	1	1	85	270	1500	38	[37;39]	0.0	163	[161;163]	0.0
1	4	1	3	5	270	2100	30	30	0.0	325	325	0.0
1	4	1	3	85	270	2100	63	[63;66]	0.0	364	[370;377]	1.8
2	2	1	1	5	390	900	0	0	0.0	50	50	0.0
2	2	1	1	85	390	900	18	[15;17]	−5.1	82	[82;83]	0.0
2	2	2	3	5	500	1500	0	0	0.0	250	250	0.0
2	2	2	3	85	500	1500	33	[29;32]	−3.4	260	[257;262]	0.0
2	4	2	4	5	500	3500	0	0	0.0	0	0	0.0
2	4	2	4	85	500	3500	18	[18;19]	0.0	17	17	0.0

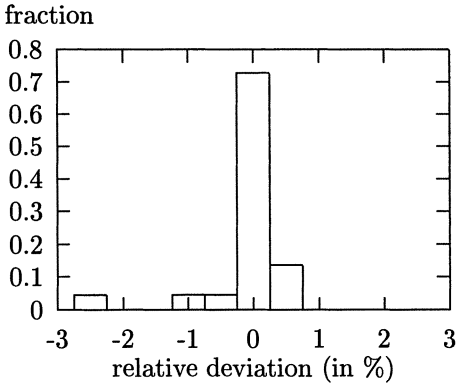


Fig. 5. Distribution of relative deviations in total backlog

Table 4. Estimated service times

T_i	T_c	L_i	L_c	σ_i	S_i	S_c	$\bar{L}_i^{\text{eff}}$(anal.)	$\bar{L}_i^{\text{eff.}}$(sim.)	LTI_i	$DS_c^{\text{anal.}}$	$DS_c^{\text{sim.}}$
1	1	1	1	5	270	600	1.00	1.00	1.00	0.00	0.00
1	1	1	1	85	270	600	1.07	1.05	1.07	0.30	0.33
1	1	2	4	5	390	1500	3.00	3.00	1.50	1.00	[1.49;1.50]
1	1	2	4	85	390	1500	3.03	[2.98;2.99]	1.52	0.91	[1.44;1.45]
1	2	2	1	5	390	900	2.10	2.10	1.05	0.50	0.50
1	2	2	1	85	390	900	2.17	2.14	1.09	0.34	[0.35;0.36]
1	2	2	4	5	390	1800	2.90	2.90	1.45	1.50	1.50
1	2	2	4	85	390	1800	2.98	[2.91;2.92]	1.49	1.18	[1.37;1.39]
1	3	1	4	5	270	2400	1.47	1.47	1.47	1.00	1.00
1	3	1	4	85	270	2400	1.58	[1.52;1.53]	1.58	0.84	[0.83;0.85]
1	3	2	3	5	390	1800	2.60	2.60	1.30	1.00	1.00
1	3	2	3	85	390	1800	2.68	[2.63;2.64]	1.34	0.97	[0.99;1.00]
1	4	1	1	5	270	1500	1.25	1.25	1.25	0.50	0.50
1	4	1	1	85	270	1500	1.33	[1.29;1.30]	1.33	0.50	[0.50;0.51]
1	4	1	3	5	270	2100	1.65	1.65	1.65	0.75	0.75
1	4	1	3	85	270	2100	1.73	[1.68;1.69]	1.73	0.99	[1.00;1.02]
2	2	1	1	5	390	900	1.10	1.10	1.10	1.00	1.00
2	2	1	1	85	390	900	1.17	1.14	1.17	0.65	[0.66;0.67]
2	2	2	3	5	500	1500	2.50	2.50	1.25	1.00	1.00
2	2	2	3	85	500	1500	2.52	[2.49;2.50]	1.26	0.91	[0.98;1.00]
2	4	2	4	5	500	3500	2.00	2.00	1.00	0.00	0.00
2	4	2	4	85	500	3500	2.02	2.03	1.01	0.14	0.15

Table 5. Internal Performance Measures with $(T_i, T_c, L_i, L_c) = (1, 1, 2, 4), S_i = 390, \sigma_i = 5$

S_c	$DS_c^{\text{anal.}}$	$DS_c^{\text{sim.}}$	$\gamma_c^{\text{anal.}}$	$\gamma_c^{\text{sim.}}$
1400	1.000	2.000	1.200	1.200
1480	1.000	1.850	1.040	1.041
1500	1.000	1.508	1.000	1.002
1520	1.000	1.151	0.960	0.962
1580	1.000	1.000	0.840	0.840
1600	1.000	1.000	0.800	0.800
1700	1.000	1.000	0.600	0.600
1800	1.000	1.000	0.400	0.400
1900	0.999	1.000	0.200	0.200
2000	0.500	0.507	0.018	0.018
2100	0.000	0.000	0.000	0.000

7 Summary

In this paper approximate procedures for the numerical evaluation of a two-echelon distribution system with local control were presented. The whole distribution system is modeled for arbitrary periodic (iid) customer demand processes. A use for other types of periodic demand than normally distributed one, e. g. Gamma or Poisson demands, is straightforward. One major advantage of the proposed approximation techniques is their easy implementation and a very low numerical effort even for large systems, say with a number of retailers $n > 20$. Numerical results show a high degree of accuracy for all performance measures even for small distribution systems with negative safety stock at the warehouse level. The robustness of the tested approximate model lends itself to a use in performance measurement or optimization. Improvements in estimation quality could be reached by integrating better estimates for waiting times of retailer orders than the state-dependent Little's formula which was used here.

The two-echelon inventory model can be easily extended to a n-echelon system by applying the aggregation technique used for the warehouse to stockpoints at higher system levels. Furthermore, in case of non-negligible fixed order costs, the distribution model can be run for several reasonable review periods to find the most advantageous schedule. Thus, lot sizing aspects like cost degression can be included theoretically in an adequate way. An extension of the model to stochastic lead times at the retailer level is possible and will be presented in a forthcoming paper. Recent numerical studies motivate a use of our periodic demand model to approximate systems with compound Poisson demand. Relative deviations between those models are found to be negligible.

Finally, a nice property of the two-echelon model presented here is that it can be easily reformulated as a cost minimization problem under service constraints for which a relaxed version can be optimized. For the special case of local $(1, S)$-policies an effective solving procedure was demonstrated in an earlier paper (see Ott et al. (1996)). For the general class of local (T, S)-policies that procedure will yield only approximate results since it reduces the problem to an extreme value problem. One possible way to improve the solving procedure is to extend it to a mixed extreme value problem. For example, it seems promising to minimize system-wide costs by restricting analysis to the mixture of worst-case and best-case scenarios for both inventory costs and customer performance.

Acknowledgement: This research was supported in part by SNF (Schweizerischer Nationalfonds zur Förderung der wissenschaftlichen Forschung) under grant 1214-042225.

References

Atkins, D. R. / Iyogun, P. O. (1988): Periodic versus 'can-order' policies for coordinated multi-item inventory systems. *Management Science*, 34:791–796.

Axsäter, S. (1993): Optimization of order-up-to-S policies in two-echelon inventory systems with periodic review. *Naval Research Logistics*, 40(2):245–253.

Badinelli, R. D. (1996): Approximating probability density functions and their convolutions using orthogonal polynomials. *European Journal of Operational Research*, 95:211–230.

Chen, F. / Zheng, Y.-S. (1992): Waiting time distribution in (T, S) inventory systems. *Operations Research Letters*, 12(3):145–151.

Clark, A. J. / Scarf, H. E. (1960): Optimal policies for a multi-echelon inventory problem. *Management Science*, 6:475–490.

De Kok, A. G. (1990): Hierarchical production planning for consumer goods. *European Journal of Operational Research*, 45:55–69.

De Kok, A. G. (1991): Basics of inventory management: Part 1 – Renewal theoretic background. Working Paper 520, Department of Econometrics, Tilburg University, Warandelaan 2 PO. Box 90153, 5000 LE Tilburg.

De Kok, A. G. (1991): Basics of inventory management: Part 2 – The (R, S)-model. Working Paper 521, Department of Econometrics, Tilburg University, Warandelaan 2 PO. Box 90153, 5000 LE Tilburg.

De Kok, A. G. / Lagodimos, A. G. / H. P. Seidel (1994): Stock allocation in a two-echelon distribution network under service constraints. Working Paper TUE/BDK/LBS/94-03, Faculty of Technology Management, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven.

Deuermeyer, B. L. / Schwarz, L. B. (1981): A model for the analysis of the system service level in warehouse–retailer distribution systems: The identical retailer case. In (1981), pages 163–193.

Diks, E. B. / de Kok, A. G. / Lagodimos, A. G. (1996): Multi-echelon systems: A service measure perspective. Working Paper TUE/TM/LBS/96-02, Department of Operations Planning and Control, Graduate School of Indus-

trial Engineering and Management Science, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven.

Eppen, G. D. / Schrage, L. (1981): Centralized ordering policies in a multi-warehouse system with lead times and random demand. In (1981), pages 51–67.

Federgruen, A. (1993): Centralized planning models for multi-echelon inventory systems under uncertainty. In Stephen C. Graves, Alexander H. G. Rinnooy Kan, and Paul H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*, pages 133–173. North-Holland, Amsterdam.

Federgruen, A. / Zipkin, P. H. (1984): Approximations of dynamic multilocation production and inventory problems. *Management Science*, 30(1):69–84.

Graves, S. C. (1989): A multi-echelon inventory model with fixed reorder intervals. Working Paper 3045-89-MS, Sloan School of Management.

Graves, S. C. (1996): A multiechelon inventory model with fixed replenishment intervals. *Management Science*, 42(1):1–18.

Hadley, G. / Whitin, T. M. (1963): *Analysis of Inventory Systems.* Prentice-Hall, Inc., Englewood Cliffs NJ.

Härtel, F. (1994): *Zufallszahlen für Simulationsmodelle – Vergleich und Bewertung verschiedener Zufallszahlengeneratoren.* Dissertation Nr. 1600, Hochschule St. Gallen, Dufourstrasse 50, CH-9000 St. Gallen.

Jackson, P. L. (1988): Stock allocation in a two-echelon distribution system or what to do until your ship comes in. *Management Science*, 34(7):880–895.

Jensen, T. (1996): *Planungsstabilität in der Material-Logistik*, volume 10 of *Schriften zur Quantitativen Betriebswirtschaftslehre.* Physica-Verlag, Heidelberg.

Johnson, M. E. / Lee, H. L. / Davis, T. / Hall, R. (1995): Expressions for item fill rates in periodic inventory systems. *Naval Research Logistics Quarterly*, 42:57–80.

Lagodimos, A. G. (1992): Multi-echelon service models for inventory systems under different rationing policies. *International Journal of Production Research*, 30 (4):939–958.

Langenhoff, L. J. G. / Zijm, W. H. M. (1990): An analytical theory of multi-echelon production/distribution systems. *Statistica Neerlandica*, 44:149–174.

Little, J. D. C. (1961): A proof for the queueing formula: $L = \lambda W$. *Operations Research*, 9(3):383–387.

Matta, K. F. / Sinha, D. (1991): Multi-echelon (R, S) inventory model. *Decision Sciences*, 22(3):484–499.

Matta, K. F. / Sinha, D. (1995): Policy and cost approximations of two-echelon distribution systems with a procurement cost at the higher echelon. *IIE Transactions*, 27:638–645.

Ott, S. Ch. / Tüshaus, U. / Wahl, C. (1996): Mehrstufige Distributionssysteme mit periodischer Kontrolle: Optimierung eines Lagerhaltungsmodells mit Servicegradrestriktion und normalverteilter Nachfrage. Arbeitsbericht, Institut für Unternehmensforschung (Operations Research), Universität St. Gallen, Bodanstrasse 6, CH-9000 St. Gallen.

Rogers, D. F. / Tsubakitani, S. (1991): Newsboy-style results for multi-echelon inventory problems: Backorders optimization with intermediate delays. *The Journal of the Operational Research Society*, 42:57–68.

Rosenbaum, B. A. (1981): Service level relationships in a multi-echelon inventory system. *Management Science*, 27:926–945.

Rosenbaum, B. A. (1981): Inventory placement in a two-echelon inventory system: An application. In (1981), pages 195–207.

Schwarz, L. B. editor (1981): *Multi-Level Production/Inventory Control Systems: Theory and Practice*, volume 16 of *Studies in the Management Sciences*. North-Holland, Amsterdam.

Sherbrooke, C. C. (1968): METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, 16:122–141.

Svoronos, A. / Zipkin, P. H. (1988): Estimating the performance of multi-level inventory systems. *Operations Research*, 36(1):57–72.

Tijms, H. C. (1986): *Stochastic Modeling and Analysis: A Computational Approach*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester New York Brisbane Toronto Singapore.

Tüshaus, U. / Wahl, C. (1997): Approximations of periodically controlled distribution systems. Working paper, Institut für Unternehmensforschung (Operations Research), Universität St. Gallen, Bodanstrasse 6, CH–9000 St. Gallen.

Van der Heijden, M. C. / Diks, E. B. / de Kok, A. G. (1996): Allocation policies in general multi-echelon distribution systems with (R, S) order-up-to-policies. Working Paper 96-23, Department of Mathematics and Computer Science, Graduate School of Industrial Engineering and Management Science, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven.

Van der Heijden, M. C. (1993): *Performance Analysis for Reliability and Inventory Models*. Dissertation, Vrije Universiteit te Amsterdam, Amsterdam.

Van der Heijden, M. C. / de Kok, A. G. (1992): Customer waiting times in an (R, S) inventory system with compound Poisson demand. *Zeitschrift für Operations Research*, 36(4):315–332.

Verrijdt, J. H. C. M. / de Kok, A. G. (1996): Distribution planning for a divergent depotless two-echelon network under service constraints. *European Journal of Operational Research*, 89:341–354.

Zipkin, P. H. (1984): On the imbalance of inventories in multi-echelon systems. *Mathematics of Operations Research*, 9:402–423.

Appendices

A The Approximation Scheme

In Fig. 6 our approach is visualized for a 1-warehouse, 1-retailer system.

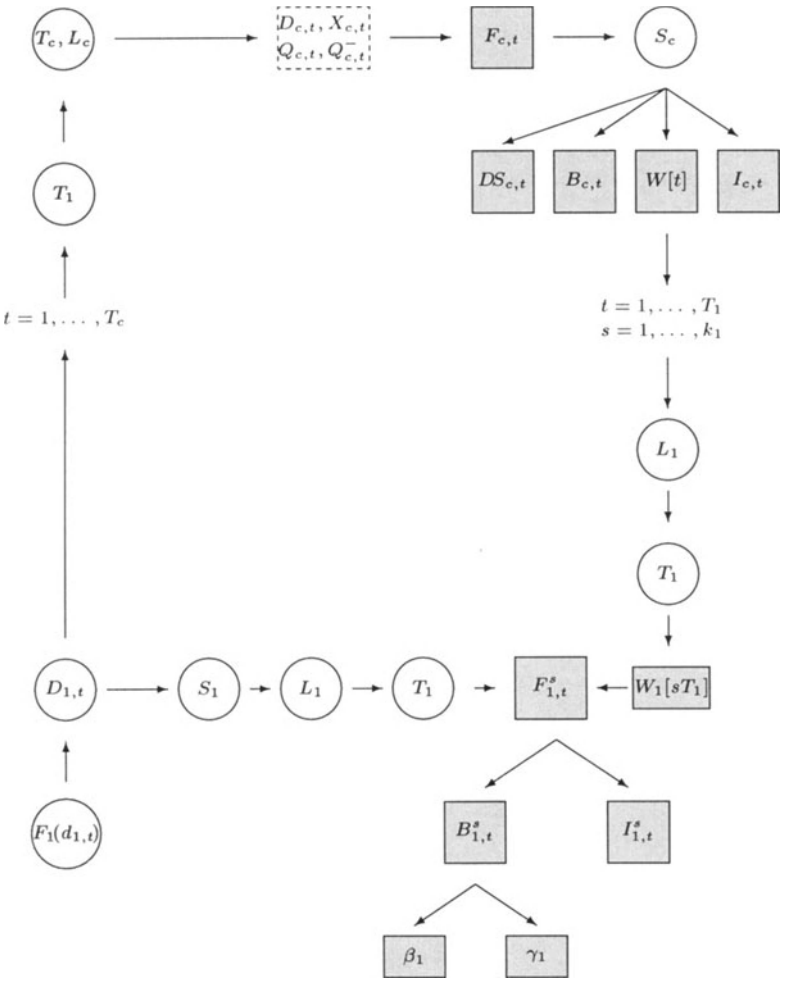


Fig. 6. Approximation concept

Expressions in circles represent model inputs. In the dashed box one finds intermediate warehouse quantities which serve to derive cdfs $F_{c,t}$ of aggregate demand. The model output is emphasized by shaded boxes. Note that there

are two individual planning levels: the planning level of the warehouse and that of the retailer. This distinction leads to a transformation process of planning intervals which is illustrated in Fig. 6 by introducing time scales $t = 1, \ldots, T_c$, $t = 1, \ldots, T_1$ and $s = 1, \ldots, k_1$ at the interfaces of bottom-up and bottom-down approaches. Finally, the reader should remark the coupling function of waiting expressions $W_1[s,t]$. An extension to the more general 1-warehouse, n-retailer system is straightforward.

B Cycle Based Approximations

In Fig. 7 the cycle based approach used here is illustrated for a one-stage system.

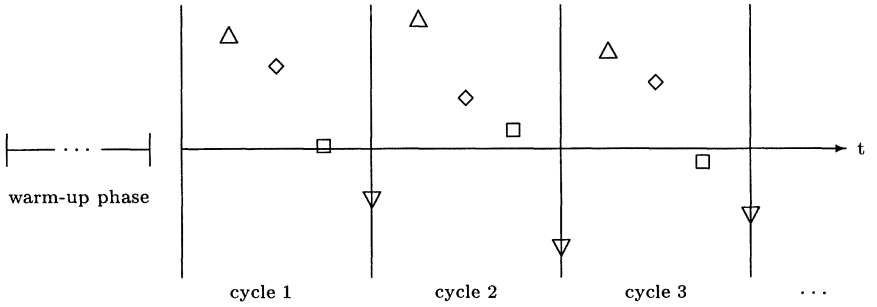


Fig. 7. Cycle based approximation scheme

Note that symbols $\triangle, \diamond, \square$ and $\triangle$ depict four different supporting points, for example net stock at discrete times t. For each supporting point a separate approximation is done. Obviously, after a warm-up phase statistical equilibrium is reached and from that point on the evolution of model quantities can be described by a specific sequence of discrete points in time during a cycle.

C Moment-Based Approximation of a PDF

The method chosen here is a moment aggregation technique based on fundamental information about iid stochastic quantities underlying different stochastic processes (delivery process, ordering process, demand process, for example). Define X_i to be iid random variables underlying separate stochastic processes A_l, $l = 1, \ldots, m$, each. By assumption, the pdf $f_i(x_i)$ and the first two moments of X_i are known.

To begin with, assume that A_l results from a linear combination of X_i, i. e. $A_l = a_l + \sum_{i=1}^{l} b_i X_i$ where $a_l = 0$ and $b_i = 1$ for all $i, l = 1, \ldots, m$. Obviously, all stochastic processes can be put down to the sum of random variables

X_i. In general, the pdf $f_l(a_l)$ of the sum variable A_l results from the l-fold convolution of pdfs $f_i(x_i)$. But, in many cases, no closed-form convolutions exist and, hence, pdf $f_l(a_l)$ must be approximated.

The following procedure is suggested: From the assumption of independence it follows that statistical moments of each stochastic process can be easily derived. Let us denote by μ_{l1}, μ_{l2} the first two non-central moments of A_l. Assume a new variable A_t to be the sum of single stochastic processes A_l, $l = 1, \ldots, m$, at a specific time t:

$$A_t = \sum_{l=1}^{m} l(A_l, t) \cdot A_l \tag{36}$$

with $l(A_l, t)$ as the indicator function for stochastic process A_l which has value 1 if A_l has positive probability to occur at time t and 0 otherwise. Again, from the assumption of independence of processes A_l, it is possible to derive at least two cumulative moments $\mathrm{E}(A_t)$ and $\mathrm{VAR}(A_t)$. Now, fit an appropriate pdf $\tilde{f}_{A_t}$ on cumulative moments $\mathrm{E}(A_t)$ and $\mathrm{VAR}(A_t)$. This yields an approximate expression for the exact pdf.

List of Contributors

Luca **Bertazzi**, University of Brescia, Department of Quantitative Methods, C. da S. Chiara 48b, 25122 Brescia, Italy

Joseph D. **Blackburn**, Vanderbilt University, Owen Graduate School of Management, Nashville, TN 37203, USA

Roman **Boutellier**, Universität St. Gallen, Institut für Technologiemanagement, Unterstrasse 22, 9000 St. Gallen, Switzerland

Arno **Bruns**, Universität St. Gallen, Institut für Unternehmensforschung, Bodanstrasse 6, 9000 St. Gallen, Switzerland

Charles J. **Corbett**, Anderson Graduate School of Management, Box 951481, UCLA, Los Angeles, CA 90095-1481, USA

Joachim R. **Daduna**, Fachhochschule für Wirtschaft Berlin, Badensche Straße 50-51, 10825 Berlin, Germany

Rommert **Dekker**, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Erik B. **Diks**, Eindhoven University of Technology, Department of Mathematics and Computing Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Karel **van Donselaar**, Eindhoven University of Technology, School of Technology Management, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Bernhard **Fleischmann**, Universität Augsburg, Lehrstuhl für Produktion und Logistik, Universitätsstraße 16, 86135 Augsburg, Germany

Lorike **Hagdorn - van der Meijden**, Erasmus University Rotterdam, Rotterdam School of Management, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Claudine **Henaux**, Université Catholique de Louvain, Institut d'Administration et de Gestion, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium

Marcel J. **Kleijn**, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Andreas **Klose**, Universität St. Gallen, Institut für Unternehmensforschung, Bodanstrasse 6, 9000 St. Gallen, Switzerland

Rochus A. **Kobler**, Universität St. Gallen, Institut für Technologiemanagement, Unterstrasse 22, 9000 St. Gallen, Switzerland

A. G. Ton **de Kok**, Eindhoven University of Technology, School of Technology Management, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

René **de Koster**, Erasmus University Rotterdam, Rotterdam School of Management, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Stefan **Kraus**, Universität Augsburg, Institut für Produktion und Logistik, Universitätsstraße 16, 86135 Augsburg, Germany

Erwin A. **van der Laan**, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Sander **de Leeuw**, Massachusetts Institute of Technology, Center for Technology, Policy and Industrial Development, Cambridge, MA 02138, USA

J. Robert **van der Meer**, Erasmus University Rotterdam, Rotterdam School of Management, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Jo A. E. E. **van Nunen**, Erasmus University Rotterdam, Rotterdam School of Management, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Edo **van der Poort**, University of Groningen, Department of Econometrics, P.O. Box 800, 9700 AV Groningen, The Netherlands

Kees J. **Roodbergen**, Erasmus University Rotterdam, Rotterdam School of Management, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Marc **Salomon**, Katholieke Universiteit Brabant, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Pierre **Semal**, Université Catholique de Louvain, Institut d'Administration et de Gestion, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium

M. Grazia **Speranza**, University of Brescia, Department of Quantitative Methods, C. da S. Chiara 48b, 25122 Brescia, Italy

Paul **Stähly**, Universität St. Gallen, Institut für Unternehmensforschung, Bodanstrasse 6, 9000 St. Gallen, Switzerland

Petra **Stumpf**, Universität Augsburg, Institut für Produktion und Logistik, Universitätsstraße 16, 86135 Augsburg, Germany

Ulrich **Tüshaus**, Universität der Bundeswehr Hamburg, 22039 Hamburg, Germany

Christoph **Wahl**, Universität St. Gallen, Institut für Unternehmensforschung, Bodanstrasse 6, 9000 St. Gallen, Switzerland

Luk N. **van Wassenhove**, INSEAD, 77305 Fontainebleau Cedex, France

Stefan **Wittmann**, Universität St. Gallen, Institut für Unternehmensforschung, Bodanstrasse 6, 9000 St. Gallen, Switzerland

Helmut **Wlcek**, Universität Augsburg, Institut für Produktion und Logistik, Universitätsstraße 16, 86135 Augsburg, Germany