GERAD
Groupe d'études et de recherche en analyse des décisions

LOGISTICS SYSTEMS

Design and Optimization

Edited by
André Langevin
Diane Riopel

Springer

LOGISTICS SYSTEMS:
DESIGN AND OPTIMIZATION

GERAD 25th Anniversary Series

- **Essays and Surveys in Global Optimization**
  Charles Audet, Pierre Hansen, and Gilles Savard, editors
- **Graph Theory and Combinatorial Optimization**
  David Avis, Alain Hertz, and Odile Marcotte, editors
- **Numerical Methods in Finance**
  Hatem Ben-Ameur and Michèle Breton, editors
- **Analysis, Control and Optimization of Complex Dynamic Systems**
  El-Kébir Boukas and Roland Malhamé, editors
- **Column Generation**
  Guy Desaulniers, Jacques Desrosiers, and Marius M. Solomon, editors
- **Statistical Modeling and Analysis for Complex Data Problems**
  Pierre Duchesne and Bruno Rémillard, editors
- **Performance Evaluation and Planning Methods for the Next Generation Internet**
  André Girard, Brunilde Sansò, and Félisa Vázquez-Abad, editors
- **Dynamic Games: Theory and Applications**
  Alain Haurie and Georges Zaccour, editors
- **Logistics Systems: Design and Optimization**
  André Langevin and Diane Riopel, editors
- **Energy and Environment**
  Richard Loulou, Jean-Philippe Waaub, and Georges Zaccour, editors

LOGISTICS SYSTEMS: DESIGN AND OPTIMIZATION

Edited by
ANDRÉ LANGEVIN
GERAD and École Polytechnique de Montréal

DIANE RIOPEL
GERAD and École Polytechnique de Montréal

Springer

André Langevin
GERAD & École Polytechnique de Montréal
Montréal, Canada

Diane Riopel
GERAD & École Polytechnique deMontréal
Montréal, Canada

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

ISBN 0-387-24971-0 e-ISBN 0-387-24977-X Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11053095

springeronline.com

Foreword

GERAD celebrates this year its 25th anniversary. The Center was created in 1980 by a small group of professors and researchers of HEC Montréal, McGill University and of the École Polytechnique de Montréal. GERAD's activities achieved sufficient scope to justify its conversion in June 1988 into a Joint Research Centre of HEC Montréal, the École Polytechnique de Montréal and McGill University. In 1996, the Université du Québec à Montréal joined these three institutions. GERAD has fifty members (professors), more than twenty research associates and post doctoral students and more than two hundreds master and Ph.D. students.

GERAD is a multi-university center and a vital forum for the development of operations research. Its mission is defined around the following four complementarily objectives:

- The original and expert contribution to all research fields in GERAD's area of expertise;
- The dissemination of research results in the best scientific outlets as well as in the society in general;
- The training of graduate students and post doctoral researchers;
- The contribution to the economic community by solving important problems and providing transferable tools.

GERAD's research thrusts and fields of expertise are as follows:

- Development of mathematical analysis tools and techniques to solve the complex problems that arise in management sciences and engineering;
- Development of algorithms to resolve such problems efficiently;
- Application of these techniques and tools to problems posed in related disciplines, such as statistics, financial engineering, game theory and artificial intelligence;
- Application of advanced tools to optimization and planning of large technical and economic systems, such as energy systems, transportation/communication networks, and production systems;
- Integration of scientific findings into software, expert systems and decision-support systems that can be used by industry.

One of the marking events of the celebrations of the 25th anniversary of GERAD is the publication of ten volumes covering most of the Center's research areas of expertise. The list follows: **Essays and Surveys in Global Optimization**, edited by C. Audet, P. Hansen and G. Savard; **Graph Theory and Combinatorial Optimization**, edited by D. Avis, A. Hertz and O. Marcotte; **Numerical Methods in Finance**, edited by H. Ben-Ameur and M. Breton; **Analysis, Control and Optimization of Complex Dynamic Systems**, edited by E.K. Boukas and R. Malhamé; **Column Generation**, edited by G. Desaulniers, J. Desrosiers and M.M. Solomon; **Statistical Modeling and Analysis for Complex Data Problems**, edited by P. Duchesne and B. Rémillard; **Performance Evaluation and Planning Methods for the Next Generation Internet**, edited by A. Girard, B. Sansò and F. Vázquez-Abad; **Dynamic Games: Theory and Applications**, edited by A. Haurie and G. Zaccour; **Logistics Systems: Design and Optimization**, edited by A. Langevin and D. Riopel; **Energy and Environment**, edited by R. Loulou, J.-P. Waaub and G. Zaccour.

I would like to express my gratitude to the Editors of the ten volumes, to the authors who accepted with great enthusiasm to submit their work and to the reviewers for their benevolent work and timely response. I would also like to thank Mrs. Nicole Paradis, Francine Benoît and Louise Letendre and Mr. André Montpetit for their excellent editing work.

The GERAD group has earned its reputation as a worldwide leader in its field. This is certainly due to the enthusiasm and motivation of GERAD's researchers and students, but also to the funding and the infrastructures available. I would like to seize the opportunity to thank the organizations that, from the beginning, believed in the potential and the value of GERAD and have supported it over the years. These are HEC Montréal, École Polytechnique de Montréal, McGill University, Université du Québec à Montréal and, of course, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds québécois de la recherche sur la nature et les technologies (FQRNT).

Georges Zaccour
Director of GERAD

Avant-propos

Le Groupe d'études et de recherche en analyse des décisions (GERAD) fête cette année son vingt-cinquième anniversaire. Fondé en 1980 par une poignée de professeurs et chercheurs de HEC Montréal engagés dans des recherches en équipe avec des collègues de l'Université McGill et de l'École Polytechnique de Montréal, le Centre comporte maintenant une cinquantaine de membres, plus d'une vingtaine de professionnels de recherche et stagiaires post-doctoraux et plus de 200 étudiants des cycles supérieurs. Les activités du GERAD ont pris suffisamment d'ampleur pour justifier en juin 1988 sa transformation en un Centre de recherche conjoint de HEC Montréal, de l'École Polytechnique de Montréal et de l'Université McGill. En 1996, l'Université du Québec à Montréal s'est jointe à ces institutions pour parrainer le GERAD.

Le GERAD est un regroupement de chercheurs autour de la discipline de la recherche opérationnelle. Sa mission s'articule autour des objectifs complémentaires suivants :

- la contribution originale et experte dans tous les axes de recherche de ses champs de compétence ;
- la diffusion des résultats dans les plus grandes revues du domaine ainsi qu'auprès des différents publics qui forment l'environnement du Centre ;
- la formation d'étudiants des cycles supérieurs et de stagiaires post-doctoraux ;
- la contribution à la communauté économique à travers la résolution de problèmes et le développement de coffres d'outils transférables.

Les principaux axes de recherche du GERAD, en allant du plus théorique au plus appliqué, sont les suivants :

- le développement d'outils et de techniques d'analyse mathématiques de la recherche opérationnelle pour la résolution de problèmes complexes qui se posent dans les sciences de la gestion et du génie ;
- la confection d'algorithmes permettant la résolution efficace de ces problèmes ;
- l'application de ces outils à des problèmes posés dans des disciplines connexes à la recherche opérationnelle telles que la statistique, l'ingénierie financière, la théorie des jeux et l'intelligence artificielle ;
- l'application de ces outils à l'optimisation et à la planification de grands systèmes technico-économiques comme les systèmes énergétiques, les réseaux de télécommunication et de transport, la logistique et la distributique dans les industries manufacturières et de service ;

- l'intégration des résultats scientifiques dans des logiciels, des systèmes experts et dans des systèmes d'aide à la décision transférables à l'industrie.

Le fait marquant des célébrations du 25^{e} du GERAD est la publication de dix volumes couvrant les champs d'expertise du Centre. La liste suit : **Essays and Surveys in Global Optimization**, édité par C. Audet, P. Hansen et G. Savard ; **Graph Theory and Combinatorial Optimization**, édité par D. Avis, A. Hertz et O. Marcotte ; **Numerical Methods in Finance**, édité par H. Ben-Ameur et M. Breton ; **Analysis, Control and Optimization of Complex Dynamic Systems**, édité par E.K. Boukas et R. Malhamé ; **Column Generation**, édité par G. Desaulniers, J. Desrosiers et M.M. Solomon ; **Statistical Modeling and Analysis for Complex Data Problems**, édité par P. Duchesne et B. Rémillard ; **Performance Evaluation and Planning Methods for the Next Generation Internet**, édité par A. Girard, B. Sansò et F. Vázquez-Abad ; **Dynamic Games : Theory and Applications**, édité par A. Haurie et G. Zaccour ; **Logistics Systems : Design and Optimization**, édité par A. Langevin et D. Riopel ; **Energy and Environment**, édité par R. Loulou, J.-P. Waaub et G. Zaccour.

Je voudrais remercier très sincèrement les éditeurs de ces volumes, les nombreux auteurs qui ont très volontiers répondu à l'invitation des éditeurs à soumettre leurs travaux, et les évaluateurs pour leur bénévolat et ponctualité. Je voudrais aussi remercier Mmes Nicole Paradis, Francine Benoît et Louise Letendre ainsi que M. André Montpetit pour leur travail expert d'édition.

La place de premier plan qu'occupe le GERAD sur l'échiquier mondial est certes due à la passion qui anime ses chercheurs et ses étudiants, mais aussi au financement et à l'infrastructure disponibles. Je voudrais profiter de cette occasion pour remercier les organisations qui ont cru dès le départ au potentiel et la valeur du GERAD et nous ont soutenus durant ces années. Il s'agit de HEC Montréal, l'École Polytechnique de Montréal, l'Université McGill, l'Université du Québec à Montréal et, bien sûr, le Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG) et le Fonds québécois de la recherche sur la nature et les technologies (FQRNT).

Georges Zaccour
Directeur du GERAD

Contents

Contributing Authors

Rosemary T. Berger
Lehigh University, USA
rtb3@lehigh.edu

James H. Bookbinder
University of Waterloo, Canada
jbookbinder@uwaterloo.ca

Nathalie Bostel
Université de Nantes, France
nathalie.bostel@iutsn.univ-nantes.fr

Hanen Bouchriha
Université Laval, Canada
Hanen.Bouchriha@forac.ulaval.ca

James F. Campbell
University of Missouri — St. Louis, USA
campbell@umsl.edu

Jean-François Cordeau
HEC Montréal and GERAD, Canada
cordeau@crt.umontreal.ca

Gilles Cormier
Université de Moncton, Canada
cormieg@UMoncton.ca

Sophie D'Amours
Université Laval, Canada
Sophie.Damours@forac.ulaval.ca

Mark S. Daskin
Northwestern University, USA
m-daskin@northwestern.edu

Pierre Dejax
École des Mines de Nantes, France
Pierre.Dejax@emn.fr

Moshe Dror
University of Arizona, USA
mdror@bpa.arizona.edu

Michel Gendreau
Université de Montréal, Canada
michelg@crt.umontreal.ca

Alain Hertz
École Polytechnique de Montréal and
GERAD, Canada
alain.hertz@gerad.ca

James K. Higginson
University of Windsor, Canada
jhiggin@uwindsor.ca

Kap Hwan Kim
Pusan National University
kapkim@pusan.ac.kr

André Langevin
École Polytechnique de Montréal and
GERAD, Canada
andré.langevin@polymtl.ca

Gilbert Laporte
HEC Montréal and GERAD, Canada
gilbert@crt.umontreal.ca

Zhiqiang Lu
École des Mines de Nantes, France
zhiqiang.lu@emn.fr

Nathalie Marcoux
École Polytechnique de Montréal, Canada
nathalie.marcoux@polymtl.ca

Benoit Montreuil
Université Laval, Canada
benoit.Montreuil@centor.ulaval.ca

Alain Martel
Université Laval, Canada
alain.martel@centor.ulaval.ca

Diane Riopel
École Polytechnique de Montréal and
GERAD, Canada
diane.riopel@polymtl.ca

Nafee Rizk
Université Laval, Canada
rizknaf@centor.ulaval.ca

Lawrence V. Snyder
Lehigh University, USA
lvs2@lehigh.edu

Jean-Sylvain Sormany
HEC Montréal, Canada
sormjs@crt.umontreal.ca

Preface

Logistics is an integral part of our every day life. Today it influences more than ever a large number of human and economic activities. The term logistics, which comes from the French word "logis" meaning dwelling, originally designated the art of organizing the transportation, resupplying, and housing of the troops of an army (that of Napoleon). From the 1960s on, the term logistics has been used in the business field to refer to the means and methods related to the physical organization of a company, and specially the flow of materials before, during, and after production. Logistics includes what is now known as supply chain management. Logistics also includes service activities.

In a context of global competition, the optimization of logistics systems is inescapable. This book falls within this perspective and presents twelve chapters that well illustrate the variety and the complexity of logistics activities. The chapters were written on invitation by recognized researchers and constitute either a summary of a particular topic, or an outline of an emerging field of logistics. The first chapter, by Riopel, Langevin, and Campbell, proposes a reference framework and allows placing the context accordingly of each of the other chapters. It classifies logistics decisions and highlights the relevant linkages among them. The intricacy of these linkages demonstrates how thoroughly the decisions are interrelated and highlights the complexity of managing logistics activities. All the other chapters focus on quantitative methods for the design and optimization of logistics systems.

In Chapter 2, Daskin, Snyder, and Berger outline the importance of facility location decisions in supply chain design. They summarize more recent research aimed at expanding facility location decisions to various supply chain contexts. Higginson and Bookbinder in the following chapter analyze logistics operations in distribution centers. They highlight the specific functions of a distribution center in comparison to those of a classical warehouse. The design and operation of a warehouse entail many challenging decision problems. Cormier presents in Chapter 4 a taxonomy of warehousing decision models and an overview of representative operations research models and solution methods for efficient warehousing.

The next chapter, by Marcoux, Riopel, and Langevin, presents a survey of operations research models and methods for facilities layout and handling system design. The focus is on the applicability of those models and methods to real-life problems. Bostel, Dejax, and Lu review in Chapter 6 applications, case studies, models and techniques proposed for facility location, inventory management, and the transportation and production planning of reverse logistics systems. They consider both cases of separate and integrated handling of original products and return flows throughout the logistics network. In the following chapter, Kim classifies and reviews models and methods for various operations in port container terminals. Considering the large amount of investment needed and the costly time spent by vessels at the terminals, it is important to improve the productivity of the handling activities.

In Chapter 8, Campbell reviews operations research models for strategic design of road transport networks, including network configuration and terminal location. This includes networks for less-than-truckload or truckload transporters, and postal carriers that serve many origins and destinations in large geographic regions. This chapter analyses several shipping strategies. COrdeau, Gendreau, Hertz, Laporte, and Sormany review in Chapter 9 some of the best metaheuristics proposed in recent years for the vehicle routing problem. These are based on local search, population search, and learning mechanisms.

In Chapter 10, Dror, through the description of the practices of a propane distribution company, analyses inventory routing problems and summarizes the literature on that topic. These problems include a family of hard problems of considerable practical significance. Martel, Rizk, D'Amours, and Bouchriha examine next the short-term production, transportation, and inventory planning problems encountered in the fine-paper industry. After placing the problems in the context of a general supply chain planning system, a comprehensive synchronized production-distribution model is gradually developed. The last chapter, by Montreuil, analyses the impact of customer centricity and personalization as well as collaboration and agility of network stakeholders on the operational optimization modeling of demand and supply chains. This chapter deals with the demand and supply chain of manufacturers of high-value products such as vehicles, computers and equipment, which are sold to consumers in a large geographical region through a network of dealers. It introduces a comprehensive operations planning optimization model applicable in such a context. It then demonstrates its application specificities as a function of the characteristics of the demand and supply chain.

This book well illustrates the diversity of logistics. We are aware that its contents do not cover all the richness of the scientific community's contribution. Our choices necessarily omitted a number of relevant topics, but we are convinced that the reader can acquire in a condensed way the knowledge of several important areas of logistics. We hope that this book will be useful both to researchers and to practitioners. We would like to sincerely thank each of the authors for the quality of their contribution. We express our gratitude towards the GERAD personnel and more particularly towards Ms. Nicole Paradis and Francine Benoît, and Mr. André Montpetit of the CRM for their important contribution to the editing of this book.

André Langevin and Diane Riopel
École Polytechnique and GERAD

Préface

La logistique fait partie intégrante de notre quotidien. Elle influence aujourd'hui plus que jamais un grand nombre d'activités humaines et économiques. Le terme logistique, qui vient du mot 'logis', est apparu à l'origine pour désigner l'art de combiner tous les moyens de transport, de ravitaillement et de logement des troupes d'une armée (celle de Napoléon). À partir des années 1960, le terme logistique a été utilisé dans le domaine des affaires pour désigner l'ensemble des moyens et des méthodes concernant l'organisation physique d'une entreprise et spécialement les flux de matières avant, pendant et après une production. La logistique englobe ce que plusieurs appellent maintenant la chaîne logistique ou la chaîne d'approvisionnement. La logistique touche aussi les entreprises de service.

Dans un contexte de compétition planétaire, que l'on parle de globalisation ou de mondialisation, l'optimisation des systèmes logistiques est incontournable. Ce livre s'inscrit dans cette optique et présente douze chapitres qui illustrent bien la grande diversité et la complexité des activités logistiques. Les chapitres ont été écrits, sur invitation, par des chercheurs reconnus et constituent soit une synthèse d'un domaine particulier, soit une présentation d'un champ en émergence de la logistique. Le premier chapitre, écrit par Riopel, Langevin et Campbell, propose un cadre de référence et permet de situer chacun des chapitres. Il classifie les décisions logistiques et met en relief les liens entre celles-ci. La complexité de ces liens montre à quel point ces décisions sont interreliées et la difficulté de gérer l'ensemble des activités logistiques. Tous les autres chapitres du livre focalisent sur les méthodes quantitatives pour la conception et l'optimisation des systèmes logistiques.

Au chapitre 2, Daskin, Snyder et Berger relèvent l'importance des décisions de localisation des installations dans la conception de chaînes logistiques. Ils résument les recherches récentes sur l'extension des modèles de localisation à divers contextes de chaînes logistiques. Higginson et Bookbinder analysent dans le chapitre suivant les opérations logistiques des centres de distribution. Ils comparent les fonctions d'un centre de distribution à celles d'un entrepôt classique. La conception et la conduite des opérations d'un entrepôt sont la source de plusieurs problèmes décisionnels difficiles. Cormier présente au chapitre 4 une taxonomie des modèles de décision en entreposage et un survol des modèles et méthodes les plus représentatifs.

Au chapitre suivant, Marcoux, Riopel et Langevin présentent une synthèse des modèles et des méthodes de recherche opérationnelle pour la conception d'implantations et de systèmes de manutention. Cette synthèse est centrée sur l'applicabilité de ces modèles et méthodes en entreprise. Bostel, Dejax et Lu passent en revue au chapitre 6 les applications, cas d'entreprise, modèles et techniques qui ont été proposées pour la localisation d'installations, la gestion des stocks, le transport et la planification de la production pour les systèmes de logistique inverse. Ils prennent en compte les mouvements de façon séparée ou intégrée des flux de produits originaux et des flux de retours dans tout le réseau logistique. Au chapitre suivant, Kim classifie et passe en revue les modèles et les méthodes pour diverses opérations dans les installations portuaires pour conteneurs. Compte tenu des investissements requis et des coûts reliés au temps à quai des bateaux, il est primordial d'optimiser la productivité des opérations de manutention des conteneurs.

Au chapitre 8, Campbell passe en revue les modèles de recherche opérationnelle pour la conception stratégique de réseau de transport routier, incluant la configuration des réseaux et la localisation des installations. Ceci inclut les réseaux de transporteurs à charges partielles ou de chargements complets, ou transporteurs de courrier postal qui desservent plusieurs origines et plusieurs destinations sur de larges territoires. Ce chapitre analyse plusieurs stratégies de transport. Cordeau, Gendreau, Hertz, Laporte, et Sormany présentent au chapitre 9 une synthèse des meilleures métaheuristiques élaborées ces dernières années pour le problème de tournées de véhicules. Ces métaheuristiques sont basées sur les méthodes de recherche locale, de recherche sur populations et sur des mécanismes d'apprentissage.

Au chapitre 10, Dror, au travers de la description des pratiques d'une compagnie de distribution de propane, analyse les problèmes combinés de tournées de véhicules et de gestion des stocks et présente une synthèse de la littérature sur ces problèmes. Ces problèmes comportent une famille de problèmes difficiles, d'une grande importance pratique. Martel, Rizk, D'Amours et Bouchriha examinent ensuite des problèmes de production, transport et gestion des stocks à court terme dans l'industrie de production de papiers fins. Après avoir situé le problème dans le contexte d'un système général de chaîne logistique, ils développent graduellement un modèle complet de production et distribution synchronisées. Le dernier chapitre, par Montreuil, analyse l'impact du centrage client et de la personnalisation ainsi que de la collaboration et de l'agilité des partenaires d'un réseau, sur la modélisation de l'optimisation des opérations d'une chaîne de demande et d'approvisionnement. Le chapitre traite de chaînes de demande et d'approvisionnement de manufacturiers de produits de grande valeur, comme des véhicules, des ordinateurs et équipements informatiques, vendus au moyen d'un large réseau de concessionnaires à des consommateurs géographiquement dispersés. Il introduit un modèle générique d'optimisation de la planification des opérations dans un tel contexte et en démontre les spécifications d'application en fonction des caractéristiques de la chaîne de demande et d'approvisionnement.

Ce livre illustre bien la diversité du domaine de la logistique. Nous sommes conscients que son contenu pourra paraître bien peu pour appréhender la richesse des travaux effectués par la communauté scientifique. Nos choix ont forcément omis un certain nombre de sujets pertinents. Mais nous sommes convaincus que les lecteurs pourront acquérir de façon condensée les connaissances sur plusieurs domaines importants de la logistique. Nous espérons que ce livre sera utile tant aux chercheurs qu'aux praticiens. Nous voulons remercier chaleureusement chacun des auteurs pour la qualité de leur contribution. Nous exprimons notre gratitude envers le personnel du GERAD, et particulièrement Mmes Nicole Paradis et Francine Benoît, et M. André Montpetit du CRM pour leur apport important à l'édition de ce livre.

André Langevin et Diane Riopel
École Polytechnique et GERAD

Chapter 1

THE NETWORK OF LOGISTICS DECISIONS

Diane Riopel
André Langevin
James F. Campbell

Abstract This chapter provides a framework for business logistics decision-making by classifying logistics decisions and highlighting the relevant linkages among them. We focus on the precedence relationships among logistics decisions and on how each decision influences and is influenced by other decisions. We also identify the key information required for making various logistics decisions. The core of our framework is a three-part decision hierarchy consisting of a strategic planning level, a network level and an operations level for 48 fundamental logistics decisions. The intricacy of the linkages between the various decisions demonstrates how thoroughly the decisions are interrelated and highlights the complexity of managing logistics activities.

1. Introduction

Effective logistics management requires good decision making in a wide variety of areas. Because the scope of logistics is so broad across both the functional areas of an organization and the temporal span of control, and because of the inherent inter-relationships between logistics decisions, logistics decision-makers must contend with a daunting array of issues and concerns. In this chapter we seek to provide a coherent framework for business logistics decision-making by identifying logistics decisions and highlighting the relevant linkages between them. Our focus is on the precedence relationships among logistics decisions, and on how each decision influences, and is influenced by, other decisions. We also identify the key information required for making various logistics decisions.

In this chapter we adopt the Council of Logistics Management definition of Logistics (2003): "Logistics is that part of the supply chain process that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services, and related information between the point of origin and the point of consumption in order to meet customers' requirements." Our goal is to delineate the precedence network of logistics decisions to help organizations better understand the interrelationships between these decisions. The issues to address are usually distributed across several departments or services, and hence across several groups of personnel. The framework we provide emphasizes the multiple links and the complexity of the resulting decision network. This is aimed at helping managers improve the efficiency, agility, and coherence of their logistics systems. This work is in keeping with the contributions to establish and manage fully integrated supply chains.

The field of business logistics has evolved substantially over the past several decades. (See for example Miyazaki et al., 1999; Langley, 1986; Kent and Flint, 1997). In the 1960s, business logistics primarily concerned two groups of functions, materials management and distribution. As Bowersox indicates regarding the founding of the National Council of Physical Distribution Management (NCPDM, now CLM) in 1963: "We were beginning to pioneer educational courses in physical distribution in those days, and nobody was integrating the functions of transportation, warehousing, and inventory to study and discuss how they worked together." (CLM web site, http://www.clm1.org/aboutUs/aboutUs_History.asp 2003)

The 1970s brought an increasing focus on the interdependence of these functions (Heskett et al., 1973; Heskett, 1977; Bowersox, 1978), and logistics "expanded" in the following years to include a more integrated perspective (Hutchinson, 1987; Ballou, 1992; Blanchard, 1992; Langford, 1995). As a reflection of this change in focus, the NCPDM changed its name to the Council of Logistics Management (CLM) in the mid-1980s. Other contributors to the evolution of logistics in organizations in the 1980s were improvements in information technologies and communications, the emergence of third party firms offering varied logistic services, and new techniques such as DRP (Distribution Resource Planning) and JIT (Just-in-Time) (The Logistics Handbook, 1994).

Since the 1990s logisticians have given increased attention to integrating the activities of all the supply chain. Global operations and customer service have become key themes (Bowersox and Closs, 1996; Coyle et al., 2003; Ganeshan et al., 1998; Kasilingam, 1999; Gattorna,1998; Stock, 2001).

The evolution of logistics has entailed an increasingly comprehensive and global vision of logistics, and a corresponding expanding scope for logistics decision-making. The decision environment has become more complex, with new management strategies and business models (e.g., JIT and e-commerce), global markets and sourcing, new information technologies and communications, a renewed focus on customer satisfaction (e.g., 24-hour service), new transport service options (e.g., overnight delivery), and increasing environmental awareness (e.g., recycling), etc. Although the logistics decision environment changes as new services, technologies, markets, and operations arise, the fundamental logistics decisions still must be made (for example, "What mode of transportation should be used?").

Many authors have classified logistics activities into different functions, and most basic logistics or supply chain management textbooks include some form of categorization for logistics activities or decisions (see for example, Ballou, 2004; Bowersox et al., 2002; Chopra and Meindl, 2004; Coyle et al., 2003; Johnson et al., 1999; Simchi-Levi et al., 2003; Stock and Lambert, 2001). The preceding works generally enumerate the logistic functions, and indicate that many of the decisions are interdependent and should be made concurrently. Models for solving various problems (facility location, vehicle routing, inventory management) are often presented in detail, but the higher level view detailing the precedence relationships among all decisions is lacking. Our goal in this chapter is to provide the network of logistics decisions to clearly delineate the precedence relations. From such a network we can then examine the relative positioning of various logistics decisions to assess their influence on other decisions. In order to build a comprehensive network, we have consulted the logistics textbooks listed in Table 1.1.

The chapter is organized as follows: Section 2 presents, for each level of the hierarchy, the relevant logistics decisions. Section 3 discusses the linkages among the decision by depicting graphically their interrelationships. A conclusion follows.

2. Logistics decision

As indicated in the previous section, the activities of logistics can be divided and classified in several different ways. Many of the differences in the various classifications occur with the activities that span the interfaces between the different functional parts of an organization, such as those activities spanning logistics and production, marketing and/or finance. Although authors have adopted different approaches in defining the basic logistics activities, and have developed different frameworks for

Table 1.1. List of textbooks

1993	American Telephone and Telegraph Company	Design's impact on logistics
1999	Anupindi, R., Chopra, S., Deshmukh, S.D., Van Mieghem, J.A. and Zemel, E.	Managing business process flows
1998 3rd edition	Arnold, J.R.T.	Introduction to materials management
2004 5th edition	Ballou, R.H.	Business logistics/Supply chain management
2001	Bauer, M.J., Poirier, C.C., Lapide, L. and Bermudez, J.	e-Business: The strategic impact on supply chain and logistics
1976	Bender, P.	Design and operation of customer service systems
2004 6th edition	Blanchard, B.S.	Logistics engineering and management
2002	Bloomberg, D.J. Lemay S. and Hanna J.B.	Logistics
2000	Bovet, D. and Martha, J.	Value nets: Breaking the supply chain to unlock hidden profits
1978 2nd edition	Bowersox, D.J.	Logistical management
1996	Bowersox, D.J. and Closs, D.J.	Logistical management: The integrated supply chain process
2002	Bowersox, D.J., Closs, D.J. and Cooper, M.B.	Supply chain logistics management
1992	Bowersox, D.J., Daugherty, P.J., Dröge, C.L., Germain, R.N. and Rogers, D.S.	Logistical excellence: It's not business as usual
1999	Boyson, S., Corsi, T.M., Dresner, M.E. and Harrington, L.H.	Logistics and the extended enterprise: Benchmarks and best practices for the manufacturing professional
1997	Bramel, J. and Simchi-Levi, D.	The logic of logistics: Theory, algorithms, and applications for logistics management
1990	Brunet, H. and Le Denn, Y.	La démarche logistique

Table 1.1 (continued)

2004 2nd edition	Chopra, S. and Meindl, P.	Supply chain management Strategy, planning, and operation
1983	Colin, J., Mathe, H. and Tixier, D.	La logistique au service de l'entreprise
1997	Copacino, W.C.	Supply chain management: The basics and beyond
2003 7th edition	Coyle, J.J., Bardi, E.J., Langley Jr., C.J.	The management of business logistics: A supply chain perspective
1998	Dornier, P.-P., Ernst, R., Fender, M. and Kouvelis, P.	Global operations and logistics: Text and cases
1997	Eymery, P.	La logistique de l'entreprise
1992	Fawcett, P., Mcleish, R. and Ogden, I.	Logistics management
2000	Fernández-Ranada, M., Gurrola-Gal, F.X. and López-Tello, E.	3C: A proven alternative to MRPII for optimizing supply chain performance
2001	Fleischmann, M.	Quantitative models for reverse logistics
1992 2nd edition	Francis, R.L., McGinnis, L.F. Jr. and White, J.A.	Facility layout and location: An analytical approach
2001	Fredendall, L.D. and Hill, E.	Basics of supply chain management
1998	Gattorna, J., editor	Strategic supply chain alignment: Best practice in supply chain management
1990	Gattorna, J., Trost, G. and Kerr, A., editors	The Gower handbook of logistics and distribution management
2003	Giard, V.	Gestion de la production et des flux
2001	Gourdin, K.N.	Global logistics management: A competitive advantage for the new millennium
1993	Graves, S.C., Rinnooy Kan, A.H.G. and Zipkin, P.H.	Handbook in Operations Research and Management Science. Volume 4 — Logistics of production and inventory
1999	Handfield, R.B. and Nichols, E.L. Jr.	Introduction to supply chain management
1973 2nd edition	Heskett, J.L., Glaskowsky, N.A. and Ivie, R.M.	Business logistics
1987	Hutchinson, N.E.	An integrated approach to logistics management

Table 1.1 (continued)

1999 7th edition	Johnson, J.C., Wood, D.F., Wardlow, D.L. and Murphy, P.R. Jr.	Contemporary logistics
1998	Kasilingam, R.G.	Logistics and transportation: Design and planning
1978	Kearney, A.T.	Measuring productivity in physical distribution
1998	Lambert, D.M., Stock, J.R. and Ellram, L.M.	Fundamentals of logistics management
1976	Lambillotte, D.	La fonction logistique dans l'entreprise
1995	Langford, J.W.	Logistics: Principles and applications
2003	Lawrence, F.B., Jennings, D.F. and Reynolds, B.E.	eDistribution
2005	Lawrence, F.B., Jennings, D.F. and Reynolds, B.E.	ERP in distribution
1993 10th edition	Leenders, M.R. and Fearon, H.E.	Purchasing and materials management
1999	Lowson, B., King, R. and Hunter, A.	Quick response: Managing the supply chain to meet consumer demand
2000	Lynch, C.F.	Logistics Outsourcing: A management guide
2002	Monczka, R., Trent, R. and Handfield, R.	Puchasing and supply chain management
2004 8th edition	Murphy Jr., P.R. and Wood, D.F.	Contemporay logistics
1973 2nd edition	Muther, R.	Systematic layout planning
2001	Pimor, Y.	Logistique: Techniques et mise en œuvre
1993	Pons, J. and Chevalier, P.	La logistique intégrée
2000 2nd edition	Ptak, C.A. and Schragenheim, E.	ERP: Tools, techniques, and applications for integrating the supply chain
2002	ReVelle, J.B., editor	Manufacturing handbook of best practices: An innovation, productivity, and quality focus
1994	Robeson, J.F., Copacino, W.C. and Howe, R.E., editors	The logistics handbook
2003	Seifert, D.	Collaborative planning, forecasting, and replenishment: How to create a supply chain advantage

Table 1.1 (continued)

2001	Shapiro, J.F.	Modeling the supply chain
2003 2nd edition	Simchi-Levi, D. and Kaminsky, P., Simchi-Levi, E.	Designing and managing the supply chain: Concepts, strategies and case studies
1973	Smykay, E.W.	Physical distribution management
1997	Southern, R.N.	Transportation and logistics basics: A handbook for transportation and logistics, professionals and students
2004	Stevenson, W.J. and Hojati, M.	Operations management
1998	Stock, J.R.	Development and implementation of reverse logistics programs
2001 4th edition	Stock, J.R and Lambert, D.M.	Strategic logistics management
1999	Tayur, S., Ganeshan, R. and Magazine, M., editors	Quantitative models for supply chain management
1997 2nd edition	Tilanus, B.	Information systems in logistics and transportation
2005	Wisner J.D. Leong, G.K. and Tan, K.-C.	Principles of supply chain management: A balanced approach
1992	Womack, J.P., Jones, D.T. and Roos, D.	Le système qui va changer le monde

presenting and organizing the various logistics activities, they all address the same fundamental logistics decisions. These logistics decisions range from long-term strategic decisions involving customer service levels and network design, to short-term tactical or operational decisions, such as daily routing of vehicles. This section delineates the different logistics decisions required in each activity, and indicates linkages between these decisions. Our focus is specifically on the logistics decisions, rather than the logistics activities, and we attempt to indicate clearly the interdependence of decisions, as well as the additional information required as input for these decisions.

Logistics decisions may be divided or grouped in several dimensions based on various criteria. The common grouping into strategic, tactical and operational levels (as in Ballou, 2004) may be based on one or more of the following criteria associated with the decisions: the time frame, the resource requirements, or the level of managerial responsibility. These criteria are generally inter-related -- for example, strategic decisions usually are made at high level in the organization and address long-term issues with significant resource implications, and these are made at a high level in the organization While in reality the range of decisions may be better viewed as a continuum on all dimensions (time frame, resource requirements, and managerial responsibility), for ease of exposition and presentation these decisions are usually separated into distinct categories.

The core of our framework for this presentation is a three-part decision hierarchy consisting of a Strategic Planning level, a Network level and an Operations level. Table 1.2 lists the decision categories within each level of the hierarchy. (Alternate classification systems and hierarchies are possible, but the underlying decisions and inter-relationships between individual decisions would not change.)

The remainder of this section is divided into subsections for the Strategic Planning level decisions, Network Design level decisions, and Operations level decisions. Within each of the subsections we detail the individual decisions, and for each of these decisions we indicate the inputs needed in the form of any previous decisions, and the other information required. For example, the carrier selection decision ("Which transportation carrier(s) should be used?") requires a previous decision on the types of carriers to be used (for example, public vs. private trucks) and additional information on available carriers, and on the organization's performance objectives. Each organization will not need to make every decision that we discuss; some organizations may contract or outsource large portions of their logistics activities, or the nature of the products and business may preclude certain decisions. Thus, our subsequent dis-

Table 1.2. Logistics decision categories

Strategic Planning level
Network level – Physical Facility (PF) Network – Communication and Information (C&I) Network
Operations level – Demand Forecasting – Inventory Management – Production – Procurement and Supply Management – Transportation – Product Packaging – Material Handling – Warehousing – Order Processing

cussions are not focussed on one particular firm or industry, but are meant to provide general coverage of logistics decision-making. To help identify the discussion for each decision, we use an italic font for the decision name throughout this section.

All the decisions described in this section, along with their immediate predecessors and the additional information required are summarized in a table in the Appendix to assist the reader. To keep the table manageable, the additional information listed for each decision is only that information not included as input for a previous decision. It should be understood that each decision may depend on a cumulative collection of previous decisions and associated additional information.

2.1 Strategic Planning level decisions

The Strategic Planning level includes high-level logistics decisions of a strategic nature. These types of decisions are likely to span functional areas beyond logistics. The key logistics decisions at this level concern performance objectives and the degree of vertical integration and outsourcing. One fundamental strategic decision is the *definition of customer service* and the associated metrics. This includes identifying the elements of customer service that are most important and most relevant for logistics, and defining exactly what will be measured and how it will be measured. This decision requires knowledge of the organizational mission and strategies, customer expectations, the competitive environment, financial resource availability and the existing logistics system (both the physical system and the information and communication system). Because financial resource availability and knowledge of the existing logistics system are relevant for (nearly) every logistics decision, we will not discuss them in each subsection.

A subsequent decision related to defining customer service is setting the *customer service objectives.* This involves developing performance standards using the previously defined customer service elements and metrics, as well as the previously mentioned additional information.

Other fundamental strategic level decisions concern the *degree of vertical integration and outsourcing* within the supply chain. Decisions on vertical integration include the nature of the integration, the direction (forward towards customers and/or backward towards suppliers), and the extent of integration (for example, which activities, parts or components should be included). Decisions related to outsourcing determine which functions should be outsourced (for example, transportation, distribution, warehousing, order processing, or fulfillment) and the extent and nature of outsourcing agreements. These decisions may rely on the previously defined customer service objectives, the availability of financial, human, material and equipment resources (including production and distribution capabilities), and the additional information needed for the definition of customer service.

There are a variety of additional strategic level decisions that affect logistics, such as determining the organization's overall economic objectives and strategy, determining the range of products and services offered, determining the geographic scope (regional, national, multinational or global) of production, distribution, and marketing, and determining the marketing and information management objectives and strategy (including electronic commerce). However, because the scope of these strategic decisions extends considerably well beyond logistics, they are not included here.

2.2 Network level decisions

Logistics decisions at the Network level are divided into two groups corresponding to the physical facility network and to the communication and information network. These are generally long-range structural decisions and they often involve considerable expenditures. Because the cost of each decision alternative is used as an input in the decision, cost is not included explicitly in the table. Note that in the physical facility network and in the communication and information network, decisions may address both forward and reverse flows.

A key network decision for the physical facility (PF) network is determining the *PF network strategy.* This specifies the overall organization or structure of the network (for example, the degree of hierarchy and number of echelons, and the degree of centralization/decentralization), and depends on the previous decisions regarding customer service objectives and the degree of vertical integration and outsourcing (at the strategic planning level), along with additional information about existing and potential suppliers, customers and markets.

Once the network strategy is determined, the *physical facility network design* must be determined. Several key decisions that concern the facilities are as follows: the type and number of facilities (for example, warehouses, terminals, distribution centers), the size and location of each facility, the activities and services provided from each facility, and whether to use new or existing facilities. Additional decisions address the linkages between facilities. These are all inter-dependent decisions that can not be made in isolation. These decisions rely on a variety of additional information used in the network strategy, customer service and vertical integration and outsourcing decisions (the existing logistics system, the competitive environment, resource availability and constraints, etc.), along with information on capability and availability of labor and support services, availability of sites and transportation, government incentives, community attitudes, environmental and zoning regulations, utilities, and taxes.

Decisions in the communication and information (C&I) network address the creation and maintenance of an effective system for communication and sharing of information throughout the supply chain. Similar to design of the physical facility network, design of the communication and information network relies on a *C&I network strategy* to define the network organization and structure. C&I network strategy decisions include the degree of centralization in information management and information processing, (for example, centralized vs. distributed), the locus of applications development (centralized in-house, distributed in-house, rental, purchase, etc.), the degree of systems integration, including the use of enterprise resource planning (ERP) systems, and the role of e-commerce. Other important C&I network strategy decisions concern the degree of standardization for the hardware, software, operating system, development environment, vendors, etc. These C&I network strategy decisions depend on the previous decisions regarding the customer service objectives, the degree of vertical integration and outsourcing and the physical facility network strategy, along with additional information on the existing C&I systems of the organization, and the existing and potential suppliers and customers.

The *design of the C&I network* requires a host of decisions concerning network architecture and capacities (decisions at nodes regarding the capture, maintenance, storage, and analysis of data and information, and decisions concerning information flows between nodes, and between functional groups at the same physical location). Other issues are the extent of information technology to be used (for example, manual paper filing systems, simple digital files, or relational databases), and hardware, software and vendor selection decisions. These decisions depend on the previous C&I network strategy and network design decisions, as well as a variety of additional information, including telecommunications regulations. Most of the C&I network design issues are beyond the scope of this chapter. See Bayles (2000), Edwards et al. (2001), Lewis and Talalayevsky (1997), Bowersox and Daugherty (1995), Nickles et al. (1998), Tilanus (1997) for details on logistics information systems.

2.3 Operations level decisions

Operations level decisions involve shorter time spans and smaller scopes than the Network level and Strategic Planning level decisions. We have divided these decisions into nine groups corresponding to fundamental logistics activities as follows: demand forecasting, inventory management, production, procurement and supply management, transportation, product packaging, material handling, warehousing, and order processing. Our primary interest is to identify these decisions and the linkages between them.

Demand forecasting. Short and long term demand forecasting are important activities that provide a basis for much logistics planning. The fundamental forecasting decisions are the *magnitude, timing and location(s) of future demand.* For existing products and markets these may be routine decisions made with the support of quantitative models. For new products and/or new markets, and for longer time horizons, more qualitative methodologies may be appropriate. These decisions are made primarily with information on historical sales, demand projections (e.g., population growth) and current/future environmental and economic outlooks, and marketing strategies.

Inventory management. Inventory management has a central role in logistics since many inventory decisions rely upon, and affect, other logistics decisions. The *inventory management strategy* (degree of centralization, push vs. pull, etc.) depends on the customer service objectives and on the availability of appropriate data via the C&I network, as well as on the fundamental nature of the products (for example, value or risk) and of the demand (patterns, dependent vs. independent, etc.). The *relative importance of inventory* items depends on the previous decisions regarding suppliers, which can influence the nature of the items themselves, as well as on the item values and historical sales data. The *methods for controlling inventories* (quantitative methods such as EOQ, kanban, etc.) depend in turn on the relative importance of items, as well as on their nature and the nature of the demand. The *desired inventory levels* are driven by desired customer service levels, the magnitude of future demand, and the supplier selection, along with the characteristics of the production process and the delays in replenishing stocks. Finally, the *safety stock* decision depends on the previous decision regarding desired inventory levels, and on the item value and delays in replenishment.

Production. Several production decisions play an important role in logistics. *Product routing* determines where work is to be completed, and this depends on the characteristics of the products and the production equipment/personnel (such as capability and performance of equipment and personnel). The *layout of production facilities* depends on the previously determined customer service objectives, the activities and services provided from each facility (part of the network design), and the product routing, as well as the production equipment/personnel characteristics (such as size and weight). The *master production schedule* is a production plan for each product, usually derived from a higher level aggregate production plan. This depends on current levels of inventory and the capacities available for production and inventory (in the physical facility network). The master production schedule drives detailed *production scheduling* that also depends on product routing and facility layout.

Procurement and supply management. Fundamental logistics decisions in procurement and supply management involve the acquisition of raw materials, parts, components, products, supplies, equipment, etc. For each product or component to be procured, there is a *procurement type* decision of how best to acquire it (for example, by purchasing or subcontracting). This depends on the customer service objectives, PF network design, the costs, resource availability (capital, personnel, facilities and equipment), the availability of products on the market, and the nature and magnitude of risks involved. For each product, component, or raw material procured, the *specifications of goods* must be determined from the range of choices. For all goods and services procured, whether purchased or subcontracted, *suppliers* must be selected and a relationship established. This depends on the previous decisions regarding the PF network design, the inventory management strategy, and the specifications of the purchased goods, as well as a range of information on suppliers' performance and capabilities (quality, reliability, dependability, etc.), the characteristics of the products, organizational purchasing policies, and transportation options. At a more detailed level the *order intervals and order quantities* must be determined. These are interrelated and depend on the master production schedule and the suppliers selected, as well as opportunities for discounts (for example, from purchasing in larger quantities). Finally, a series of *quality control* decisions (what type of quality control program, what will be measured, where, by whom, etc.) are required to ensure that procured materials are satisfactory. These decisions depend on the suppliers and characteristics of the products.

Transportation. There are eight fundamental transportation decisions involving both inbound and outbound movements. The *transportation mode* decision depends on the previous decisions concerning customer service objectives, the existing network of facilities and transportation links, and the master production schedule. Additional information on available transportation modes, product characteristics, and standards and regulations is required. Once the appropriate transportation mode(s) are identified, the types of carriers (for example public vs. private motor carriers), and the carriers themselves must be selected. The *carrier type* decision depends on the previous decisions regarding transportation modes and production scheduling, along with historical sales data, information on carriers (capabilities, performance, costs, etc.), product characteristics, and standards and regulations. The *carrier* selection decision then depends on the carrier type decision, as well as detailed information on the carriers' performance and capabilities. The *degree of consolidation* is chosen to exploit economies of scale and optimize the total relevant costs. It depends on the previous decisions regarding the physical facility network, the order intervals and quantities, and the types of carriers, as well as on the characteristics of the product(s), and the customers' demands and locations.

The *transportation fleet mix* decision determines the mixture of vehicles (possibly from different modes) comprising the fleet. This depends on the previous decisions concerning the degree of consolidation, the types of carriers, and the demand forecasts. It also depends on the characteristics of the product(s), and a wide range of data on the available transportation fleet options. The *assignment of customers to vehicles, vehicle routing and scheduling*, and *vehicle load plans* are all closely related. These decisions determine when and where each vehicle goes, and what it carries. This depends on the previously defined physical facility network, the transportation fleet mix, and the product packaging, as well as the product characteristics, access to receiving/shipping docks, and the customers' location, demand and time windows. The load plans should allow secure transportation and efficient handling of all the products.

Product packaging. Product packaging is an area that traditionally receives less attention in logistics than areas such as transportation and inventory management. Yet packaging can greatly influence both the economic and environmental aspects of logistics. The *level of protection* to be provided by the packaging depends on the previous decisions regarding transportation modes, the types of material handling equipment and the desired inventory levels. This also requires information on the product characteristics, including product value, environmental conditions (weather, compatibility with other products, etc.), relevant standards and regulations, product characteristics, and the duration of storage. The *information to be provided* about the product by the packaging also needs to be determined. This includes information for the consumer (identification, instructions, warnings, etc.), as well as information for the logistics providers (for example, "This End Up"). This depends on the characteristics of the product, the needs of the customer (often as determined by Marketing), and applicable regulations (for example, regarding hazardous or toxic materials). The *media* used to communicate the information provided with the packaging (labels applied to the package; printing directly on the packaging, radio-frequency tags, etc.) depends on the information to be provided and the options for communication. The *type of packaging* and the *packaging design* need to be determined based on the level of protection required, and the information to be provided, with additional information on the product characteristics, packaging material options (including environmental consequences), customer desires, and possibilities for reuse and recycling of packaging materials.

Material handling. Material handling is concerned with the loading and unloading of vehicles as well as the movement of goods inside the facilities. A fundamental decision is the size of the *unit loads*. This depends on the previous decisions regarding packaging design, production scheduling, order intervals and quantities, and the inventory management strategy, along with additional information on the characteristics of the objects to handle, characteristics of the production equipment and personnel, and customer needs. The *type of material handling equipment* to use (for example, hand trucks, forklifts, or conveyers) depends on the unit loads, order picking procedures, layout of production facilities and warehouses, and the available types of material handling equipment. The type of handling equipment, along with the characteristics of the production equipment and personnel, and information on available material handling equipment, helps determine the *fleet mix* for the material handling equipment. The *material handling fleet control* depends on the inventory management strategy, the production scheduling, the material handling fleet mix, and the order picking procedures.

Warehousing. Warehousing decisions address issues at the storage facilities. The *mission and functions* of warehousing (for example, long-term storage versus cross-docking) depend on the customer service objectives, the nature of demand, and the characteristics of the products. The possible use of third-party logistics services has to be evaluated. *Warehouse layout* depends on the mission and function of the warehouse, the inventory management strategy, desired inventory levels, packaging design and the types of material handling equipment, along with additional information to ensure the safety of employees (including applicable regulations). *Stock location* decisions determine where in the warehouse and according to what storage policy each item is stored. This depends on the previous decisions regarding the customer service objectives, the relative importance of inventory items, and the warehouse layout, along with the characteristics of the products being placed in storage. The *design of receiving and shipping areas* of the warehouse deserves special attention and depends on the transportation modes utilized, the material handling fleet mix, the unit loads, the mission and functions of the warehouse and the packaging design. Additional information on the characteristics of the products being handled at the docks, loading and unloading times, and the safety of employees is also used. Finally, the *safety systems* for warehouse operations must be determined. This depends on the mission and functions of the warehouse, the warehouse layout, including shipping and receiving dock design, and the product characteristics.

Order processing. As several of the order processing decisions involve acquisition and transmission of order information, they are tightly linked to the communication and information system discussed earlier. The *order entry* and *order transmission procedures* depend on the communications and information (C&I) network strategy and network design, along with additional information on the customer demands, the range of products and the capabilities of the relevant personnel. *Order-picking procedures* depend on previous decisions regarding the C&I network strategy, stock locations, unit loads, material handling equipment, and packaging design, along with additional information on the customer demands and range of products. The *order follow-up* decisions include activities after an order is placed to ensure it is successfully fulfilled. This depends on the PF network design, the C&I network design, and the customer demands.

3. Sequences and relationships

The previous section described 48 fundamental logistics decisions at the Strategic Planning, Network, and Operations levels. For each decision we indicated the preceding logistics decisions, as well as additional information required to make the decision. In this section we combine the precedence information from all these decisions to depict graphically the inter-relationships of the logistics decisions. Figure 1.1 shows the 48 decisions with directed arcs indicating the precedence relationships. The numbers in the figure refer to the decisions as listed in Table 1.3. Dashed outlines are drawn around groups of decisions to define the 12 categories as in Table 1.2 and the Appendix. This figure includes 98 arcs linking the various decisions, and the directions of the arcs shows how each decision is influenced by other "upstream" decisions, and in turn, influences various "downstream" decisions.

This figure clearly depicts the complex interrelationships among the logistics decisions. Figure 1.1 also shows that eight cycles of decisions exist involving nine decisions (31, 34, 35, 36, 37, 38, 41, 42, and 47) from four categories: Materials Handling, Product Packaging, Warehousing and Order Processing. The decision cycles are shown in Figure 1.2. Cycles of decisions imply interdependence, and a need for concurrent decision-making.

Table 1.4 summarizes some information regarding the linkages between decisions. The first two columns of Table 1.4 provide the decision category and the number of the decision. The third and fourth columns are the number of decisions *immediately* following (i.e., number of arcs out of), and number of decisions *immediately* preceding (i.e., number of arcs into), each decision, respectively. The fifth and sixth columns are the total number of decisions "downstream" (i.e., number of decisions following), and the total number of decisions "upstream" (i.e., number of decisions preceding) for each decision, respectively. The seventh and eighth columns provide the length of the longest acyclic paths "downstream" and "upstream" from each decision, respectively. The horizontal lines in Table 1.4 separate the different decision categories (as identified in Table 1.2 and the Appendix). The numbers for the nine diferent decisions involved in cycles are shown in bold, and note that they have identical numbers of upstream and downstream decisions.

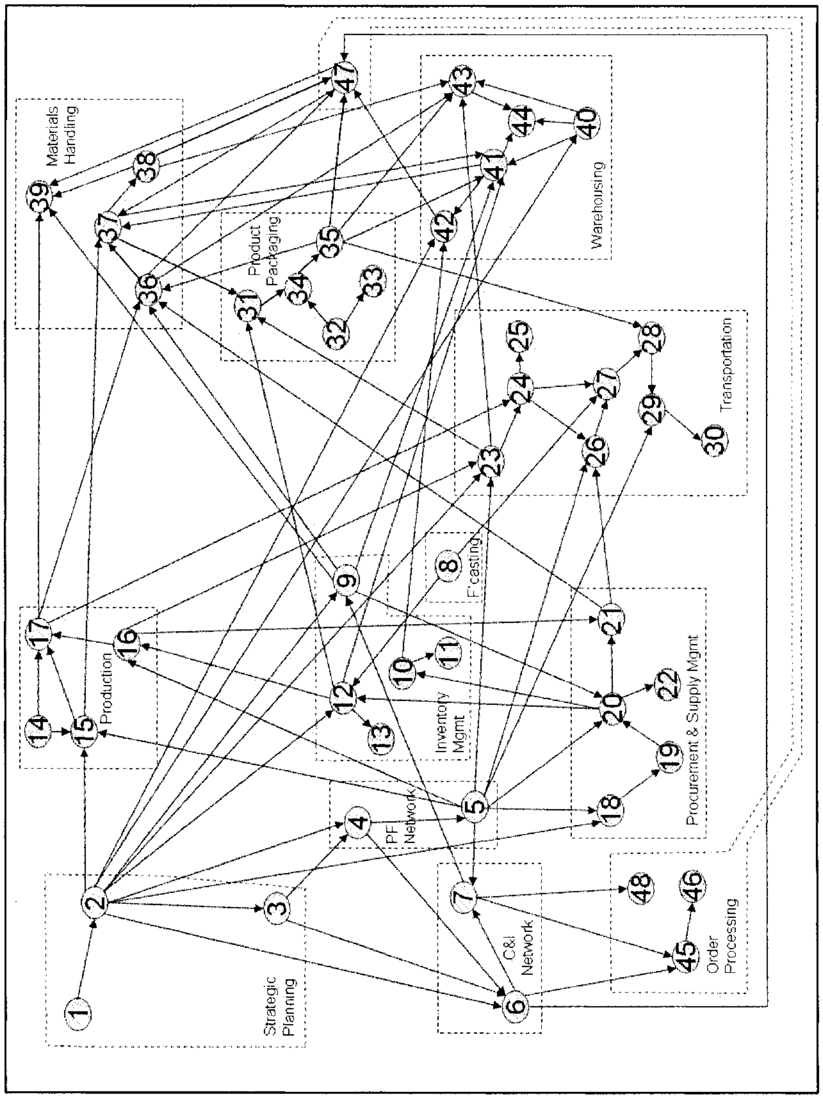


Figure 1.1. Precedence relationships between the decisions

Table 1.3. The 48 logistics decisions

Strategic Planning level
1. Definition of customer service
2. Customer service objectives
3. Degree of vertical integration and outsourcing

Physical Facility (PF) Network
4. PF network strategy
5. PF network design

Communication and Information (C&I) Network
6. C&I network strategy

Inventory Management
7. C&I network design

Demand Forecasting
8. Forecasts of demand magnitude, timing, and locations
9. Inventory management strategy
10. Relative importance of inventory
11. Control methods
12. Desired inventory level
13. Safety stock

Production
14. Product routing
15. Facilities layout
16. Master production schedule
17. Production scheduling

Procurement and Supply Management
18. Procurement type
19. Specifications of goods procured
20. Suppliers
21. Order intervals and quantities
22. Quality control

Transportation
23. Transportation modes
24. Types of carriers
25. Carriers
26. Degree of consolidation
27. Transportation fleet mix
28. Assignment of customers to vehicles
29. Vehicle routing and scheduling
30. Vehicle load plans

Product Packaging
31. Level of protection needed
32. Information to be provided with the
product
33. Information media
34. Type of packaging
35. Packaging design

Material Handling
36. Unit loads
37. Types of material handling equipment
38. Material handling fleet mix
39. Material handling fleet control

Warehousing
40. Warehousing mission and functions
41. Warehouse layout
42. Stock location
43. Receiving/shipping dock design
44. Safety systems

Order Processing
45. Order entry procedures
46. Order transmission means
47. Order picking procedures
48. Order follow-up procedures

As expected, the Strategic Planning level decisions (decisions 1–3) and Network Design decisions (decisions 4–7) have a great influence on subsequent Operations level decisions, as shown by the large values in columns 5 and 7 of Table 1.4. The "most influential decision" is 1 (Definition of customer service), which has 43 downstream decisions and a longest downstream path that involves 20 other decisions (1-2-3-4-5-7-9-20-12-23-31-34-35-41-42-47-37-38-43-44). The only logistics decisions not downstream from decision 1 are: Demand forecasts (8), Product

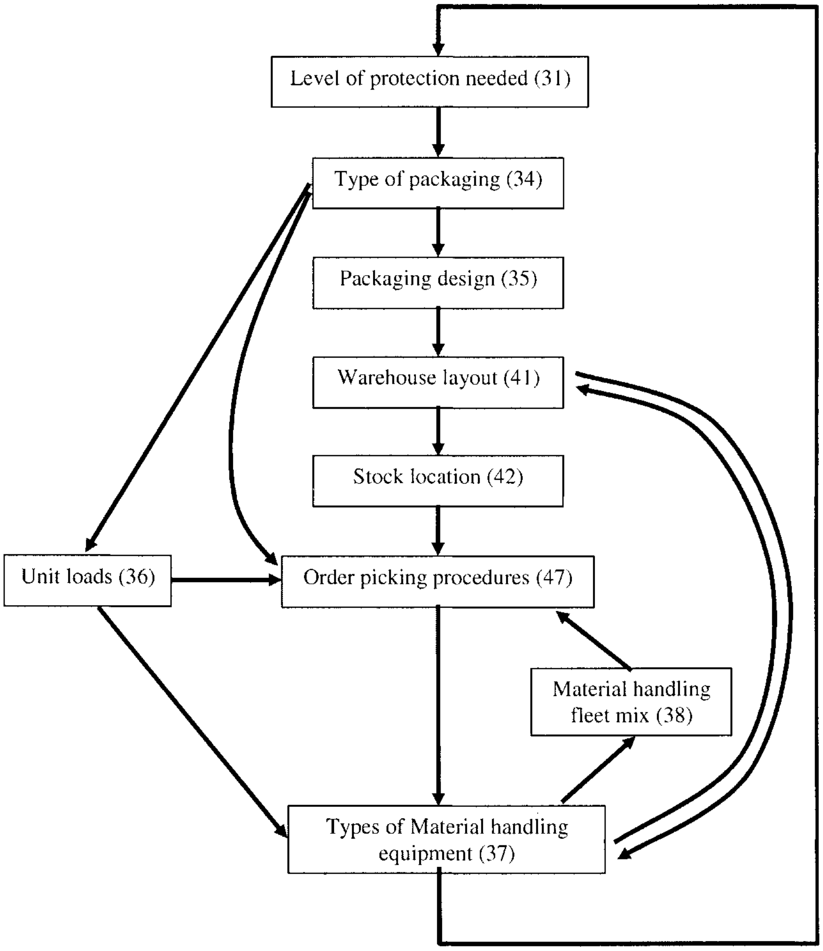

Level of protection needed (31)
Type of packaging (34)
Packaging design (35)
Warehouse layout (41)
Stock location (42)
Unit loads (36)
Order picking procedures (47)
Material handling fleet mix (38)
Types of Material handling equipment (37)

Figure 1.2. Precedence relationships between the decisions

routing (14), Information to be provided with the product (32), and Information media (33). Four decisions have no preceding logistics decisions: (1, 8, 14 and 32) and ten decisions have no subsequent logistics decisions (11, 13, 22, 25, 30, 33, 39, 44, 46, 48).

In general, decisions with a greater number of downstream decisions have a wider influence, and decisions with a greater number of upstream decisions have more decisions influencing them. Nearly all decisions in the Demand Forecasting, Inventory Management, Production, and Procurement and Supply Management categories (decisions 8 – 22) come near the beginning of a decision sequence (the number of downstream

Table 1.4. Decision summary

Category	**Decision**	**Number Out**	**Number In**	**Number Downstream**	**Number Upstream**	**Max Path Downstream**	**Max Path Upstream**
Strategic Planning	1	1	0	43	0	20	0
	2	10	1	42	1	19	1
	3	2	1	40	2	18	2
Physical Fac. Network	4	2	2	39	3	17	3
	5	8	1	37	4	16	4
C&I Network	6	3	3	34	4	16	4
	7	3	2	33	6	15	5
Forecasting	8	2	0	25	0	13	0
Inventory Management	9	4	2	29	7	14	6
	10	2	1	16	11	9	8
	11	0	1	0	12	0	9
	12	4	3	24	12	12	8
	13	0	1	0	13	0	9
Production	14	2	0	21	0	11	0
	15	2	3	20	6	10	5
	16	3	2	22	13	11	9
	17	3	3	19	16	9	10
Procurement and Supply Management	18	1	2	30	5	15	5
	19	1	1	29	6	14	6
	20	4	3	28	10	13	7
	21	2	2	17	14	9	10
	22	0	1	0	11	0	8
Transportation	23	3	3	19	14	10	10
	24	3	2	6	18	5	11
	25	0	1	0	19	0	12
	26	1	3	4	20	4	12

Table 1.4 (continued)

Category	**Decision**	**Number Out**	**Number In**	**Number Downstream**	**Number Upstream**	**Max Path Downstream**	**Max Path Upstream**
	27	1	3	3	21	3	13
	28	1	2	2	34	2	17
	29	1	2	1	35	1	18
	30	0	1	0	36	0	19
Product Packaging	**31**	1	3	14	30	9	14
	32	2	0	16	0	9	0
	33	0	1	0	1	0	1
	34	1	2	14	30	8	15
	35	5	1	14	30	7	16
Material Handling	**36**	3	4	14	30	8	15
	37	3	4	14	30	7	17
	38	3	1	14	30	2	18
	39	0	4	0	31	0	19
Warehousing	40	3	1	15	2	10	2
	41	3	5	14	30	9	17
	42	1	3	14	30	8	18
	43	1	5	1	31	1	19
	44	0	3	0	32	0	20
Order Processing	45	1	2	1	7	1	6
	46	0	1	0	8	0	7
	47	2	5	14	30	7	18
	48	0	1	0	7	0	6

decisions exceeds the number of upstream decisions, and the longest downstream paths exceed the longest upstream paths). This is especially true for Demand forecasting (8), Inventory management strategy (9), Product routing (14), Facilities layout (15), Procurement type (18), and Specification of goods procured (19). All of the decisions in the other Operations level categories (Transportation, Product Packaging, Material Handling, Warehousing and Order Processing (decisions 23–48) come towards the middle or end of a decision sequence, with the exception of Information to be provided (32), and Warehousing mission and functions (40). Decisions with a rather limited upstream and downstream influence include: Information media (33), and decisions in the Order Processing category (45, 46, and 48).

Care must be taken when examining the number of upstream and downstream decisions, and the longest acyclic path lengths, in Table 1.4 due to the cycles of decisions. Because of these cycles, the same decision may appear both upstream and downstream. Furthermore, the relative importance of the decisions is not reflected in our analysis, and clearly some decisions have greater impacts than others.

The information on number of upstream and downstream decisions, and the longest upstream and downstream paths can be used to help assess the relative influence of the decisions, and to distinguish between immediate influences and more remote, transitive influences. Decisions with many immediate successors (larger values of "Number Out") will tend to have their influence felt more rapidly and broadly. Decisions with many immediate predecessors (larger values of "Number In") require integrating direct inputs from many different, and often varied, decisions. For example, consider decisions 37 (Types of material handling equipment) and 38 (Material handling fleet mix). These have very similar numbers in Table 1.4, except for "Number In" and Max Path Downstream." Even though both decision 37 and 38 have 30 different upstream decisions, decision 37 has 4 immediate upstream decisions (Number In=4), while decision 38 has only 1 immediate upstream decision (decision 37!). Even though both decision 37 and 38 have 14 downstream decisions, the influence of decision 38 extends only a maximum of 2 decisions downstream (Max Path Downstream=2), while the influence of decision 37 extends a maximum of 7 decisions downstream. Thus, decision 38, relative to decision 37, has lesser immediate influence from upstream, and its own influence is diffused rather rapidly downstream.

The information on longest path lengths can also help address issues of responsiveness and agility in a logistics systems. Changing a decision for which the maximum downstream path is long may require considerable time for the influence to work its way through the chain of subsequent sequential decisions. For example, decision 15 (Facilities layout) has 15 downstream decisions, but 10 of these need to be made in sequence because of the precedence relations. Thus, organizations should pay special attention to those decisions on these long paths to ensure they can respond and react quickly to changes.

Figure 1.3 helps visualize the linkages between logistics decision categories by grouping the individual decisions as in the Appendix and accumulating the precedence relations (number of arcs) between the different decision categories. The numbers on each arc in Figure 1.3 represent the number of arcs between the *categories* from Figure 1.1. Note that cycles among categories in Figure 1.3 may not reflect cycles among decisions (in Figure 1.1), since each category aggregates several decisions, which may not be linked. For example, in Figure 1.1 there are arcs from decision 9 to 20, 20 to 10, and 20 to 12. These three arcs (in Figure 1.1) produce the two arcs in Figure 1.3 linking Inventory Management and Procurement and Supply Management. However, this does not represent a cycle of decisions, since decisions 9, 10, 12 and 20 are not connected in any cycles.

Table 1.5 provides a summary of the linkages between the decision categories. The first two columns provide the name of the category, and the number of decisions in each category. The third and fourth columns provide the number of *decisions* that are immediately following ("Number of Arcs Out") and immediately preceding ("Number of Arcs In") each category, respectively. The fifth column provides the number of linkages between decisions within a category ("Number of Internal Arcs"). Note the concentration in the Transportation and Warehousing categories. The sixth and seventh columns show the number of other categories that are downstream and upstream, respectively.

This sixth column of Table 1.5 (and Figure 1.3) clearly shows the broad influence of the Strategic Planning and Network Level decisions. The last two columns of Table 1.5 suggest a division of the 12 decision categories into three groups based on their influence, as shown with the horizontal lines in Table 1.5. The Strategic Planning and Network level decisions, along with Demand forecasting form the first group. These categories have strong influence (downstream linkages) on all the operations level decision categories. The second group includes the next three Operations level decision categories (Inventory Management, Production, and Procurement and Supply Management). These categories have a central position, with a strong downstream influence (on eight of the Operations level decision categories — all except Demand forecasting) and links to six upstream decision categories: the four categories in the first group (Demand forecasting, plus the Strategic Planning and

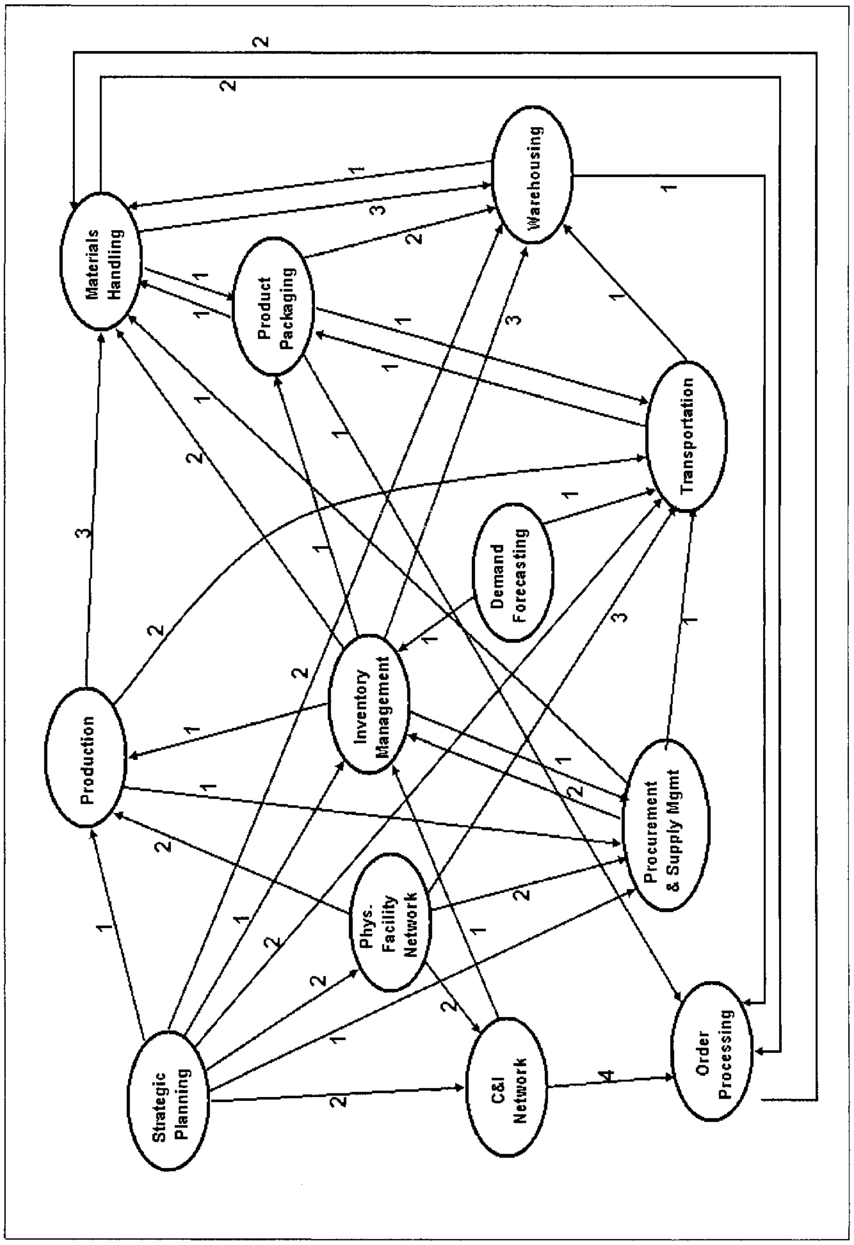


Figure 1.3. Linkages between logistics decision categories

Table 1.5. Decision categories summary

Category	Number of Decisions	Number of Arcs Out	Number of Arcs In	Number of Internal Arcs	Number of Categories Downstream	Number of Categories Upstream
Strategic Planning	3	11	0	2	11	0
Physical Facility Network	2	9	2	1	10	1
C&I Network	2	5	4	1	9	2
Forecasting	1	2	0	0	8	0
Inventory Management	5	8	6	2	7	6
Production	4	6	4	4	7	6
Procurement & Supply Mgmt.	5	4	5	4	7	6
Transportation	8	2	9	8	4	11
Product Packaging	5	5	3	4	4	11
Material Handling	4	6	10	3	4	11
Warehousing	5	2	11	6	4	11
Order Processing	4	2	8	1	4	11

Network categories) and the other two Operations level categories in this group. The last five decision categories (Transportation, Product Packaging, Material Handling, Warehousing, and Order processing) have less downstream influence (only to the other four decision categories in this group), but strong linkages upstream to *all* other decision categories.

4. Conclusion

This chapter presents the network of logistics decisions by focussing on the precedence relationships in logistics decision-making. The intricacy of the linkages demonstrates how thoroughly the decisions are interrelated and highlights the complexity of managing logistics activities. The framework provided here, in the form of a network, tables, and figures, is aimed at helping logistics managers understand and analyze the relationships between the various decisions, which could lead to making better decisions. The Appendix also provides additional information required by each decision, and this should be valuable for planning and optimization of the logistics chain.

The importance of logistics is increasing with the expanding geographic scope of global supply chains. As decision-makers become more dispersed and communications more difficult — both due to the physical distance and the cultural distance, the ability to understand the repercussions of a decision on others may be the key to a successful implementation of global logistics strategies. In the same vein, being able to react quickly to market changes and to design agile supply chains requires a profound knowledge of the relationships between logistics decisions. This article is a step to acquiring such knowledge.

Future research may address how the logistics decision framework in this chapter can be used to provide better information for logistic decision-making. One key is the development of procedures and mechanisms to ensure that the appropriate information is available where it is needed in an accurate and timely fashion. While information technology may offer some solutions, often the organizational issues and changes are more challenging. Another topic for future research is to refine this framework with case studies in a particular organization or industry sector.

Appendix

Decisions	Previous decisions	Additional information
Strategic Planning Level		
1. Definition of customer service		■ Organizational mission and strategy ■ Customer expectations ■ Competitive environment ■ Financial resource availability ■ Existing logistics system
2. Customer service objectives	1. Definition of customer service	
3. Degree of vertical integration and outsourcing	2. Customer service objectives	■ Resource availability (capital, personnel, facilities and equipment)
Physical Facility (PF) Network		
4. PF network strategy	2. Customer service objectives 3. Degree of vertical integration and outsourcing	■ Existing suppliers ■ Existing customers ■ Potential suppliers ■ Potential customers and markets
5. PF network design, including - types of facility - number of each type of facility - size of facility - facility location - activities and services from each facility - utilization of new or existing facilities - links between facilities	4. PF network strategy	■ Capability and availability of labor and support services ■ Availability of appropriate facilities and sites ■ Availability of transportation ■ Government incentives ■ Community attitudes ■ Standards and regulations ■ Utilities ■ Taxes

Decisions	Previous decisions	Additional information
Communication and Information (C&I) Network		
6. C&I network strategy	2. Customer service objectives 3. Degree of vertical integration and outsourcing 4. PF network strategy	■ Existing C&I systems of the organization ■ Existing suppliers ■ Existing customers ■ Potential suppliers ■ Potential customers
Inventory Management		
7. C&I network design, including - network architecture and capacities - hardware selection - software selection - vendor selection - extent of information technology used	5. PF network design 6. C&I network strategy	■ Capability and availability of labor and support services ■ Availability of appropriate facilities and sites ■ Government incentives ■ Community attitudes ■ Standards and regulations
Demand Forcasting		
8. Forecasts of demand magnitude, timing and locations		■ Historical sales data ■ Environmental and economic data ■ Marketing strategies
9. Inventory management strategy	2. Customer service objectives 7. C&I network design	■ Nature of products ■ Nature of demand
10. Relative importance of inventory	20. Suppliers	■ Item value ■ Historical sales data
11. Control methods	10. Relative importance of inventory	■ Nature of products ■ Nature of demand
12. Desired inventory level	2. Customer service objectives 8. Forecasts of demand magnitude, timing and locations 20. Suppliers	■ Production equipment/personnel characteristics ■ Replenishment delay

Decisions	Previous decisions	Additional information
13. Safety stock	12. Desired inventory level	■ Item value ■ Replenishment delay
Production		
14. Product routing		■ Product characteristics ■ Production equipment/personnel characteristics
15. Facilities layout	2. Customer service objectives 5. PF network design 14. Product routing	■ Production equipment/personnel characteristics
16. Master production schedule	5. PF network design 12. Desired inventory level	■ Current inventory levels
17. Production scheduling	14. Product routing 15. Facilities layout 16. Master production schedule	
Procurement and Supply Management		
18. Procurement type	2. Customer service objectives 5. PF network design	■ Cost to make and cost to buy ■ Resource availability (capital, personnel, facilities and equipment) ■ Availability of products ■ Nature and magnitude of risks
19. Specifications of goods procured	18. Procurement type	■ Product design specifications ■ Production equipment/personnel characteristics
20. Suppliers	5. PF network design 9. Inventory management strategy 19. Specifications of goods procured	■ Suppliers performance and capabilities ■ Procurement policies ■ Transportation options
21. Order intervals and quantities	16. Master production schedule 20. Suppliers	■ Discount opportunities
22. Quality control	20. Suppliers	■ Characteristics of products to procure

Decisions	Previous decisions	Additional information
Transportation		
23. Transportation modes	2. Customer service objectives 5. PF network design 16. Master production schedule	■ Transportation options ■ Standards and regulations ■ Product characteristics
24. Types of carriers	17. Production scheduling 23. Transportation modes	■ Historical sales data ■ Carrier options ■ Standards and regulations ■ Product characteristics
25. Carriers	24. Types of carriers	■ Carriers' performance and capabilities
26. Degree of consolidation	5. PF network design 21. Order intervals and quantities 24. Types of carriers	■ Customer locations ■ Product characteristics
27. Transportation fleet mix	8. Forecasts of demand magnitude, timing and locations 24. Types of carriers 26. Degree of consolidation	■ Product characteristics ■ Transport fleet options
28. Assignment of customers to vehicles	27. Transportation fleet mix 35. Packaging design	■ Customer locations ■ Customer demands ■ Product characteristics ■ Access to receiving/shipping docks
29. Vehicle routing and scheduling	5. PF network design 28. Assignment of customers to vehicles	■ Customer locations ■ Customer demands ■ Time windows
30. Vehicle load plans	29. Vehicle routing and scheduling	

Decisions	Previous decisions	Additional information
Product Packaging		
31. Level of protection needed	12. Desired inventory level 23. Transportation modes 37. Types of material handling equipment	■ Product value ■ Environmental conditions ■ Standards and regulations ■ Product characteristics ■ Duration of storage
32. Information to be provided with the product		■ Product characteristics ■ Customer needs ■ Standards and regulations
33. Information media	32. Information to be provided with the product	■ Options for communicating information
34. Type of packaging	31. Level of protection needed 32. Information to be provided with the product	■ Product characteristics ■ Packaging material options ■ Reusing/recycling options
35. Packaging design	34. Type of packaging	■ Product characteristics ■ Customer needs
Material Handling		
36. Unit loads	9. Inventory management strategy 17. Production scheduling 21. Order intervals and quantities 35. Packaging design	■ Characteristics of objects to handle ■ Customer needs ■ Production equipment/personnel characteristics
37. Types of material handling equipment	15. Facilities layout 36. Unit loads 41. Warehouse layout 47. Order picking procedures	■ Material handling options
38. Material handling fleet mix	37. Types of material handling equipment	■ Production equipment/personnel characteristics ■ Material handling equipment performance and capabilities

Decisions	Previous decisions	Additional information
39 Material handling fleet control	9. Inventory management strategy 17. Production scheduling 38. Material handling fleet mix 47. Order picking procedures	
Warehousing		
40 .Warehousing mission and functions	2. Customer service objectives	■ Product characteristics ■ Nature of demand
41. Warehouse layout	9. Inventory management strategy 12. Desired inventory level 35. Packaging design 37. Types of material handling equipment 40. Warehousing mission and functions	■ Safety of employees
42. Stock location	2. Customer service objectives 10. Relative importance of inventory 41. Warehouse layout	■ Product characteristics
43. Receiving/shipping dock design	23. Transportation modes 35. Packaging design 36. Unit loads 38. Material handling fleet mix 40. Warehousing mission and functions	■ Characteristics of received and shipped goods ■ Amount of product to handle at dock ■ Safety of employees
44. Safety systems	40. Warehousing mission and functions 41. Warehouse layout 43. Receiving/shipping dock design	■ Product characteristics
Order Processing		
45. Order entry procedures	6. C&I network strategy 7. C&I network design	■ Customer demands ■ Range of products ■ Capability and availability of labor and support services

Decisions	**Previous decisions**	**Additional information**
46. Order transmission means	45. Order entry procedures	
47. Order picking procedures	6. C&I network strategy 35. Packaging design 36. Unit loads 38. Material handling fleet mix 42. Stock location	■ Customer demands ■ Range of products
48. Order follow-up procedures	5. PF network design 7. C&I network design	■ Customer demands

Acknowledgements This work was supported by the Canadian Natural Sciences and Engineering Council and by the Fonds de recherche sur la nature et les technologies — Québec.

References

American Telephone and Telegraph Company (1993). *Design's Impact on Logistics.* McGraw-Hill.

Anupindi, R., Chopra, S., Deshmukh, S.D., Van Mieghem, J.A., and Zemel, E. (1999). *Managing Business Process Flows.* Prentice Hall.

Arnold, J.R.T. (1998). *Introduction to Materials Management*, 3rd edition. Prentice Hall.

Ballou, R.H. (1992). *Business Logistics Management*, 3rd edition. Prentice-Hall.

Ballou, R.H. (2004). *Business Logistics/Supply Chain Management*, 5th edition. Prentice-Hall.

Bauer, M.J., Poirier, C.C., Lapide, L., and Bermudez, J. (2001). *e-Business: The strategic impact on supply chain and logistics.* Council of Logistics Management.

Bayles, D.L. (2000). *E-Commerce Logistics and Fulfillment: Delivering the Goods*, 1st edition, p. 358. Prentice Hall PTR.

Bender, P. (1976). *Design and Operation of Customer Service Systems.* Amacom.

Blanchard, B.S. (1992). *Logistics Engineering and Management*, 4th edition. Prentice Hall.

Blanchard, B.S. (2004). *Logistics Engineering and Management*, 6th edition. Prentice Hall.

Bloomberg, D.J., Lemay S., and Hanna J.B. (2002). *Logistics.* Prentice Hall.

Bovet, D. and Martha, J. (2000). *Value Nets: Breaking the Supply Chain to Unlock Hidden Profits.* John Wiley & Sons.

Bowersox, D.J. (1978). *Logistical Management*, 2nd edition. Macmillan Publishing Co.

Bowersox, D.J. and Closs, D.J. (1996). *Logistical Management: The Integrated Supply Chain Process*, McGraw-Hill.

Bowersox, D.J., Closs, D.J., and Cooper, M.B. (2002). *Supply Chain Logistics Management*, Mc Graw-Hill.

Bowersox, D.J. and Daugherty, P.J. (1995). Logistics paradigms: The impact of information technology, *Journal of Business Logistics*, 16:65 – 80.

Bowersox, D.J., Daugherty, P.J., Dröge, C.L., Germain, R.N., and Rogers, D.S. (1992). *Logistical Excellence: It's Not Business as Usual.* Digital Press.

Boyson, S., Corsi, T.M., Dresner, M.E., and Harrington, L.H. (1999). *Logistics and the Extended Enterprise: Benchmarks and Best Practices for the Manufacturing Professional.* John Wiley & Sons.

Bramel, J. and Simchi-Levi, D. (1997). *The logic of logistics: Theory, algorithms, and applications for logistics management.* Springer.

Brunet, H. and Le Denn, Y. (1990). *La démarche logistique.* Afnor Gestion.

Chopra, S. and Meindl, P. (2004). *Supply Chain Management. Strategy, planning, and operation*, 2nd edition. Pearson Prentice Hall.

Colin, J., Mathe, H., and Tixier, D. (1983). *La logistique au service de l'entreprise.* Dunod Entreprise.

Copacino, W.C. (1997). *Supply Chain Management: The Basics and Beyond.* St. Lucie Press/APICS Series on Resource Management.

Coyle, J.J., Bardi, E.J. and Langley Jr., C.J. (2003). *The Management of Business Logistics: A Supply Chain Perspective*, 7th edition. South-Western, Thomson Learning.

Dornier, P.-P., Ernst, R., Fender, M. and Kouvelis, P. (1998). *Global Operations and Logistics: Text and Cases.* John Wiley & Sons.

Edwards, P., Peters, M., and Sharman, G. (2001). The effectiveness of information systems in supporting the extended supply chain, *Journal of Business Logistics*, 22:1 – 27.

Eymery, P. (1997). *La logistique de l'entreprise.* Hermes.

Fawcett, P., Mcleish, R., and Ogden, I. (1992). *Logistics Management.* Pitman Publishing.

Fernández-Ranada, M., Gurrola-Gal, F.X., and López-Tello, E. (2000). *3C: A Proven Alternative to MRPII for Optimizing Supply Chain Performance.* St. Lucie Press.

Fleischmann, M. (2001). *Quantitative Models for Reverse Logistics.* Springer-Verlag.

Francis, R.L., McGinnis, L.F., Jr., and White, J.A. (1992). *Facility Layout and Location: An Analytical Approach*, 2nd edition. Prentice-Hall.

Fredendall, L.D. and Hill, E. (2001). *Basics of Supply Chain Management.* St. Lucie Press.

Ganeshan, R., Jack, E., Magazine, M.J. and Stephens, P. (1998). Chapter 27 – A taxonomic review of supply chain management research. In: S. Tayur, R. Ganeshan, and M. Magazine (eds.), *Quantitative Models for Supply Chain Management*, pp. 841-879. Kluwer Academic Publishers.

Gattorna, J. (1998). *Strategic Supply Chain Alignment: Best Practice in Supply Chain Management.* Gower.

Gattorna, J., Trost, G., and Kerr, A., editors (1990). *The Gower Handbook of Logistics and Distribution Management.* Gower.

Giard, V. (2003). *Gestion de la production et des flux.* Economica.

Gourdin, K.N. (2001). *Global Logistics Management: A Competitive Advantage for the New Millennium.* Blackwell.

Graves, S.C., Rinnooy Kan, A.H.G., and Zipkin, P.H. (1993). *Handbook in Operations Research and Management Science. Volume 4 — Logistics of Production and Inventory.* North Holland Publ.

Handfield, R.B. and Nichols, E.L., Jr. (1999). *Introduction to Supply Chain Management.* Prentice Hall.

Heskett, J.L. (1977). Logistics — Essential to strategy. *Harvard Business Review*, November:11.

Heskett, J.L., Glaskowsky, N.A., and Ivie, R.M. (1973). *Business Logistics*, 2nd edition. The Ronald Press Company.

Hutchinson, N.E. (1987). *An Integrated Approach to Logistics Management*, Prentice Hall.

Johnson, J.C., Wood, D.F., Wardlow, D.L., and Murphy, P.R., Jr. (1999). *Contemporary Logistics*, 7th edition. Prentice Hall.

Kasilingam, R.G. (1998). *Logistics and Transportation: Design and Planning.* Kluwer Academic Publishers.

Kearney, A.T. (1978). *Measuring Productivity in Physical Distribution.* National Council of Physical Distribution Management.

Kent, J.L., Jr., and Flint, D.J. (1997). Perspectives on the evolution of logistics thought, *Journal of Business Logistics*, 18:15 – 29.

Lambert, D.M., Stock, J.R., and Ellram, L.M. (1998). *Fundamentals of Logistics Management.* Irwin/McGraw-Hill.

Lambillotte, D. (1976). *La fonction logistique dans l'entreprise.* Dunod Entreprise.

Langford, J.W. (1995). *Logistics: Principles and Applications*, McGraw-Hill.

Langley, C.J., Jr. (1986). The evolution of the logistics concept, *Journal of Business Logistics*, 7:1 – 13.

Lawrence, F.B., Jennings, D.F., and Reynolds, B.E. (2003). *eDistribution.* Thompson South-Western.

Lawrence, F.B., Jennings, D.F., and Reynolds, B.E. (2005). *ERP in Distribution.* Thompson South-Western.

Leenders, M.R. and Fearon, H.E. (1993). *Purchasing and Materials Management*, 10th edition. Irwin.

Lewis, I. and Talalayevsky, A. (1997). Logistics and information technology: A coordination perspective, *Journal of Business Logistics*, 18:141 – 156.

Lowson, B., King, R. and Hunter, A. (1999). *Quick Response: Managing the Supply Chain to meet Consumer Demand.* John Wiley & Sons.

Lynch, C.F. (2000). *Logistics Outsourcing: A Management Guide.* Council of Logistics Management.

Miyazaki, A.D., Phillips, J.K., and Phillips, D.M. (1999). Twenty years of *JBL*: An analysis of published research, *Journal of Business Logistics*, 20:1 – 19.

Monczka, R., Trent, R. and Handfield, R. (2002). *Puchasing and Supply Chain Management.* South-Western.

Murphy Jr., P.R. and Wood, D.F. (2004). *Contemporay Logistics*, 8th edition. Pearson Prentice Hall.

Muther, R. (1973). *Systematic layout planning*, 2nd edition. CBI Publishing Company Inc.

Nickles, T., Mueller, J., and Takacs, T. (1998). Strategy, information technology and the supply chain — managing information technology for success, not just survival. In: J. Gattorna, *Strategic Supply Chain Alignment: Best Practice in Supply Chain Management*, pp. 494 – 508. Gower Publishing.

Pimor, Y. (2001). *Logistique: techniques et mise en œuvre.* Dunod.

Pons, J. and Chevalier, P. (1993). *La logistique intégrée.* Éditions Hermes.

Ptak, C.A. and Schragenheim, E. (2000). *ERP: Tools, Techniques, and Applications for Integrating the Supply Chain*, 2nd edition. St. Lucie Press.

ReVelle, J.B., editor (2002). *Manufacturing handbook of best practices: An innovation, productivity, and quality focus.* St. Lucie Press.

Robeson, J.F., Copacino, W.C., and Howe, R.E., editors (1994). *The Logistics Handbook.* The Free Press.

Seifert, D. (2003). *Collaborative Planning, Forecasting, and Replenishment: How to Create a Supply Chain Advantage.* Amacom.

Shapiro, J.F. (2001). *Modeling the Supply Chain.* Duxbury Thomson Learning.

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E. (2003). *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies*, 2nd edition. Irwin/McGraw-Hill.

Smykay, E.W. (1973). *Physical Distribution Management.* Macmillan Publishing Co..

Southern, R.N. (1997). *Transportation and Logistics Basics: A Handbook for Transportation and Logistics, Professionals and Students.* Continental Traffic Pub. Co.

Stevenson, W.J. and Hojati, M. (2004). *Operations Management.* McGraw-Hill.

Stock, J.R. (1998). *Development and implementation of reverse logistics programs.* Council of Logistics Management, 1998.

Stock, J.R. (2001). Doctoral research in logistics and logistics-related areas: 1992 – 1998, *Journal of Business Logistics*, 22:125 – 256.

Stock, J.R. and Lambert, D.M. (2001). *Strategic Logistics Management*, 4th edition. McGraw-Hill.

Tayur, S., Ganeshan, R., and Magazine, M., editors (1999). *Quantitative models for supply chain management.* Kluwer Academic Publishers.

Tilanus, B. (1997). *Information Systems in Logistics and Transportation*, 2nd edition. Pergamon Press.

Wisner J.D., Leong, G.K., and Tan, K.-C. (2005). *Principles of Supply Chain Management: A Balanced Approach*, Thomson South-Western.

Womack, J.P., Jones, D.T., and Roos, D. (1992). *Le système qui va changer le monde.* Dunod.

Chapter 2

FACILITY LOCATION IN SUPPLY CHAIN DESIGN

Mark S. Daskin
Lawrence V. Snyder
Rosemary T. Berger

Abstract In this chapter we outline the importance of facility location decisions in supply chain design. We begin with a review of classical models including the traditional fixed charge facility location problem. We then summarize more recent research aimed at expanding the context of facility location decisions to incorporate additional features of a supply chain including LTL vehicle routing, inventory management, robustness, and reliability.

1. Introduction

The efficient and effective movement of goods from raw material sites to processing facilities, component fabrication plants, finished goods assembly plants, distribution centers, retailers and customers is critical in today's competitive environment. Approximately 10% of the gross domestic product is devoted to supply-related activities (Simchi-Levi, Kaminsky, and Simchi-Levi, 2003, p. 5). Within individual industries, the percentage of the cost of a finished delivered item to the final consumer can easily exceed this value. Supply chain management entails not only the movement of goods but also decisions about (1) where to produce, what to produce, and how much to produce at each site, (2) what quantity of goods to hold in inventory at each stage of the process, (3) how to share information among parties in the process and finally, (4) where to locate plants and distribution centers.

Location decisions may be the most critical and most difficult of the decisions needed to realize an efficient supply chain. Transportation and inventory decisions can often be changed on relatively short notice in response to changes in the availability of raw materials, labor costs, component prices, transportation costs, inventory holding costs, exchange rates and tax codes. Information sharing decisions are also relatively flexible and can be altered in response to changes in corporate strategies and alliances. Thus, transportation, inventory, and information sharing decisions can be readily re-optimized in response to changes in the underlying conditions of the supply chain. Decisions about production quantities and locations are, perhaps, less flexible, as many of the costs of production may be fixed in the short term. Labor costs, for example, are often dictated by relatively long-term contracts. Also, plant capacities must often be taken as fixed in the short term. Nevertheless, production quantities can often be altered in the intermediate term in response to changes in material costs and market demands.

Facility location decisions, on the other hand, are often fixed and difficult to change even in the intermediate term. The location of a multibillion-dollar automobile assembly plant cannot be changed as a result of changes in customer demands, transportation costs, or component prices. Modern distribution centers with millions of dollars of material handling equipment are also difficult, if not impossible, to relocate except in the long term. Inefficient locations for production and assembly plants as well as distribution centers will result in excess costs being incurred throughout the lifetime of the facilities, no matter how well the production plans, transportation options, inventory management, and information sharing decisions are optimized in response to changing conditions.

However, the long-term conditions under which production plants and distribution centers will operate is subject to considerable uncertainty at the time these decisions must be made. Transportation costs, inventory carrying costs (which are affected by interest rates and insurance costs), and production costs, for example, are all difficult to predict. Thus, it is critical that planners recognize the inherent uncertainty associated with future conditions when making facility location decisions.

Vehicle routing and inventory decisions are generally secondary to facility location in the sense that facilities are expensive to construct and difficult to modify, while routing and inventory decisions can be modified periodically without difficulty. Nevertheless, it has been shown empirically for both location/routing and location/inventory problems that the facility location decisions that would be made in isolation are different from those that would be made taking into account routing or inventory. Similarly, planners are often reluctant to consider robustness and reliability at design time since disruptions may be only occasional; however, large improvements in reliability and robustness can often be attained with only small increases in the cost of the supply chain network.

In this chapter we review several traditional facility location models, beginning with the classical fixed charge location model. We then show how the model can be extended to incorporate additional facets of the supply chain design problem, including more accurate representations of the delivery process, inventory management decisions, and robustness and reliability considerations.

2. The fixed charge facility location problem

The fixed charge facility location problem is a classical location problem and forms the basis of many of the location models that have been used in supply chain design. The problem can be stated simply as follows. We are given a set of customer locations with known demands and a set of candidate facility locations. If we elect to locate a facility at a candidate site, we incur a known fixed location cost. There is a known unit shipment cost between each candidate site and each customer location. The problem is to find the locations of the facilities and the shipment pattern between the facilities and the customers to minimize the combined facility location and shipment costs subject to a requirement that all customer demands be met.

Specifically, we introduce the following notation:

Inputs and sets.

I: set of customer locations, indexed by i
J: set of candidate facility locations, indexed by j
h_i: demand at customer location $i \in I$
f_j: fixed cost of locating a facility at candidate site $j \in J$
c_{ij}: unit cost of shipping between candidate facility site $j \in J$ and customer location $i \in I$

Decision variables.

$$X_j = \begin{cases} 1, & \text{if we locate at candidate site } j \in J, \\ 0, & \text{if not} \end{cases}$$

$Y_{ij} =$ fraction of the demand at customer location $i \in I$ that is served by a facility at site $j \in J$

With this notation, the fixed charge facility location problem can be formulated as follows (Balinski, 1965):

$$\text{minimize} \quad \sum_{j \in J} f_j X_j + \sum_{j \in J} \sum_{i \in I} h_i c_{ij} Y_{ij} \tag{2.1}$$

$$
\begin{aligned}
\text{subject to} \sum_{j \in J} Y_{ij} &= 1 && \forall i \in I && (2.2)\\
Y_{ij} - X_j &\leq 0 && \forall i \in I; \forall j \in J && (2.3)\\
X_j &\in \{0, 1\} && \forall j \in J && (2.4)\\
Y_{ij} &\geq 0 && \forall i \in I; \forall j \in J && (2.5)
\end{aligned}
$$

The objective function (2.1) minimizes the sum of the fixed facility location costs and the transportation or shipment costs. Constraint (2.2) stipulates that each demand node is fully assigned. Constraint (2.3) states that a demand node cannot be assigned to a facility unless we open that facility. Constraint (2.4) is a standard integrality constraint and constraint (2.5) is a simple non-negativity constraint.

The formulation given above assumes that facilities have unlimited capacity; the problem is sometimes referred to as the uncapacitated fixed charge location problem. It is well known that at least one optimal solution to this problem involves assigning all of the demand at each customer location $i \in I$ fully to the nearest open facility site $j \in J$. In other words, the assignment variables, Y_{ij}, will naturally take on integer values in the solution to this problem. Many firms insist on or strongly prefer such *single sourcing* solutions as they make the management of the supply chain considerably simpler. Capacitated versions of the fixed charge location problem do not exhibit this property; enforcing single sourcing is significantly more difficult in this case (as discussed below).

A number of solution approaches have been proposed for the uncapacitated fixed charge location problem. Simple heuristics typically begin by constructing a feasible solution by greedily adding or dropping facilities from the solution until no further improvements can be obtained. Maranzana (1964) proposed a neighborhood search improvement algorithm for the closely related P-median problem (Hakimi, 1964, 1965) that exploits the ease in finding optimal solutions to 1-median problem: it partitions the customers by facility and then finds the optimal location within each partition. If any facility changes, the algorithm repartitions the customers and continues until no improvement in the solution can be found. Teitz and Bart (1968) proposed an exchange or "swap" algorithm for the P-median problem that can also be extended to the fixed charge location problem. Hansen and Mladenović (1997) proposed a variable neighborhood search algorithm for the P-median problem that can also be used for the fixed charge location problem. Clearly, improvement heuristics designed for the P-median problem will not perform well for the fixed charge location problem if the starting number of facilities is sub-optimal. One way of resolving this limitation is to apply more sophisticated heuristics to the problem. Al-Sultan and Al-Fawzan (1999) applied tabu search (Glover 1989, 1990; Glover and Laguna, 1997) to the uncapacitated fixed charge location problem. The algorithm was tested successfully on small- to moderate-sized problems.

Erlenkotter (1978) proposed the well-known DUALOC procedure to find optimal solutions to the problem. Galvão (1993) and Daskin (1995) review the use of Lagrangian relaxation algorithms in solving the uncapacitated fixed charge location problem. When embedded in branch and bound, Lagrangian relaxation can be used to solve the fixed charge location problem optimally (Geoffrion, 1974). The reader interested in a more comprehensive review of the uncapacitated fixed charge location problem is referred to either Krarup and Pruzan (1983) or Cornuéjols, Nemhauser, and Wolsey (1990).

One natural extension of the problem is to consider capacitated facilities. If we let b_j be the maximum demand that can be assigned to a facility at candidate site $j \in J$, formulation (2.1) – (2.5) can be extended to incorporate facility capacities by including the following additional constraint:

$$\sum_{i \in I} h_i Y_{ij} - b_j X_j \leq 0 \quad \forall j \in J \tag{2.6}$$

Constraint (2.6) limits the total assigned demand at facility $j \in J$ to a maximum of b_j. From the perspective of the integer programming problem, this constraint obviates the need for constraint (2.3) since any solution that satisfies (2.5) and (2.6) will also satisfy (2.3). However, the linear programming relaxation of (2.1) – (2.6) is often tighter if constraint (2.3) is included in the problem.

For fixed values of the facility location variables, X_j, the optimal values of the assignment variables can be found by solving a traditional transportation problem. The embedded transportation problem is most easily recognized if we replace $h_i Y_{ij}$ by Z_{ij}, the *quantity* shipped from distribution center j to customer i. The transportation problem for fixed facility locations is then

$$\text{minimize} \sum_{j \in J} \sum_{i \in I} c_{ij} Z_{ij} \tag{2.7}$$

$$\text{subject to} \sum_{j \in J} Z_{ij} = h_i \qquad \forall i \in I \tag{2.8}$$

$$\sum_{i \in I} Z_{ij} \leq b_j \hat{X}_j \quad \forall j \in J \tag{2.9}$$

$$Z_{ij} \geq 0 \qquad \forall i \in I; \forall j \in J \tag{2.10}$$

where we denote the fixed (known) values of the location variables by $\hat{X}_j$.

The solution to the transportation problem (2.7) – (2.10) may involve fractional assignments of customers to facilities. This means that the solution to the problem with the addition of constraint (2.6) will not automatically satisfy the single sourcing condition, as does the solution to the uncapacitated fixed charge location problem in the absence of this constraint. To restore the single sourcing condition, we can replace the fractional definition of the assignment variables by a binary one:

$$Y_{ij} = \begin{cases} 1, & \text{if demands at customer site } i \in I \text{ are served by a facility} \\ & \text{at candidate site } j \in J, \\ 0, & \text{if not.} \end{cases}$$

The problem becomes considerably more difficult to solve since there are now far more integer variables. For given facility sites, even if we ignore the requirement that each demand node is served exactly once, the resulting problems become knapsack problems, which can only be solved optimally in pseudo-polynomial time (as opposed to the transportation problem, which can be solved in polynomial time).

Daskin and Jones (1993) observed that, in many practical contexts, the number of customers is significantly greater than the number of distribution centers that will be sited. As such, each customer represents a small fraction of the total capacity of the distribution center to which it is assigned. Also, if the single sourcing requirement is relaxed, the number of multiply sourced customers is less than or equal to the number of distribution centers minus one. Thus, relatively few customers will be multiply sourced in most contexts. They further noted that warehouse capacities, when measured in terms of annual throughput as is commonly done, are rarely known with great precision, as they depend on many factors, including the number of inventory turns at the warehouse. (We return to the issue of inventory turns below when we outline an integrated location/inventory model.) They therefore proposed a procedure for addressing the single sourcing problem that involves (1) ignoring the single sourcing constraint and solving the transportation problem, (2) using duality to find alternate optima to the transportation problem that require fewer customers to be multiply sourced, and (3) allowing small violations of the capacity constraints to identify solutions that satisfy the single sourcing requirement. In a practical context involving a large national retailer with over 300 stores and about a dozen distribution centers, they found that this approach was perfectly satisfactory from a managerial perspective.

In a classic paper, Geoffrion and Graves (1974) extend the traditional fixed charge facility location problem to include shipments from plants to distribution centers and multiple commodities. They introduce the following additional notation:

Inputs and sets.

K: set of plant locations, indexed by k
L: set of commodities, indexed by l
D_{li}: demand for commodity $l \in L$ at customer $i \in I$
S_{lk}: supply of commodity $l \in L$ at plant $k \in K$
$\underline{V}_j$, $\overline{V}_j$: minimum and maximum annual throughput allowed at distribution center $j \in J$
v_j: variable unit cost of throughput at candidate site $j \in J$
c_{lkji}: unit cost of producing and shipping commodity $l \in L$ between plant $k \in K$, candidate facility site $j \in J$ and customer location $i \in I$

Decision variables:.

$$Y_{ij} = \begin{cases} 1, & \text{if demands at customer site } i \in I \text{ are served by a facility} \\ & \text{at candidate site } j \in J, \\ 0, & \text{if not} \end{cases}$$

Z_{lkji} = quantity of commodity $l \in L$ shipped between plant $k \in K$, candidate facility site $j \in J$ and customer location $i \in I$

With this notation, Geoffrion and Graves formulate the following extension of the fixed charge location problem:

$$\text{minimize} \quad \sum_{j \in J} f_j X_j + \sum_{j \in J} v_j \left(\sum_{l \in L} \sum_{i \in I} D_{li} Y_{ij} \right) + \sum_{l \in L} \sum_{k \in K} \sum_{j \in J} \sum_{i \in I} c_{lkji} Z_{lkji} \tag{2.11}$$

$$\text{subject to} \quad \sum_{i \in I} \sum_{j \in J} Z_{lkji} \le S_{lk} \qquad \forall k \in K; \forall l \in L \tag{2.12}$$

$$\sum_{k \in K} Z_{lkji} = D_{li} Y_{ij} \qquad \forall l \in L; \forall j \in J; \forall i \in I \tag{2.13}$$

$$\sum_{j \in J} Y_{ij} = 1 \qquad \forall i \in I \tag{2.14}$$

$$\underline{V}_j X_j \le \sum_{i \in I} \sum_{l \in L} D_{li} Y_{ij} \le \overline{V}_j X_j \qquad \forall j \in J \tag{2.15}$$

$$X_j \in \{0, 1\} \qquad \forall j \in J \tag{2.16}$$

$$Y_{ij} \in \{0, 1\} \qquad \forall i \in I; \forall j \in J \tag{2.17}$$

$$Z_{lkji} \ge 0 \qquad \forall i \in I; \forall j \in J;$$

$$\forall k \in K; \forall l \in L \qquad (2.18)$$

The objective function (2.11) minimizes the sum of the fixed distribution center (DC) location costs, the variable DC costs and the transportation costs from the plants through the DCs to the customers. Constraint (2.12) states that the total amount of commodity $l \in L$ shipped from plant $k \in K$ cannot exceed the capacity of the plant to produce that commodity. Constraint (2.13) says that the amount of commodity $l \in L$ shipped to customer $i \in I$ via DC $j \in J$ must equal the amount of that commodity produced at all plants that is destined for that customer and shipped via that DC. This constraint stipulates that demand must be satisfied at each customer node for each commodity and also serves as a linking constraint between the flow variables (Z_{lkji}) and the assignment variables (Y_{ij}). Constraint (2.14) is the now-familiar single sourcing constraint. Constraint (2.15) imposes lower and upper bounds on the throughput processed at each distribution center that is used. This also serves as a linking constraint (e.g., it replaces constraint (2.3)) between the location variables (X_j) and the customer assignment variables (Y_{ij}). Alternatively, it can be thought of as an extension of the capacity constraint (2.6) above.

In addition to the constraints above, Geoffrion and Graves allow for linear constraints on the location and assignment variables. These can include constraints on the minimum and maximum number of distribution centers to be opened, relationships between the feasible open DCs, more detailed capacity constraints if different commodities use different amounts of a DC's resources, and certain customer service constraints. The authors apply Benders decomposition (Benders, 1962) to the problem after noting that, if the location and assignment variables are fixed, the remaining problem breaks down into $|L|$ transportation problems, one for each commodity.

Geoffrion and Graves highlight eight different forms of analysis that were performed for a large food company using the model, arguing, as do Geoffrion and Powers (1980), that the value of a model such as (2.11)–(2.18) extends far beyond the mere solution of a single instance of the problem to include a range of sensitivity and what-if analyses.

3. Integrated location/routing models

An important limitation of the fixed charge location model, and even the multi-echelon, multi-commodity extension of Geoffrion and Graves, is the assumption that full truckload quantities are shipped from a distribution center to a customer. In many contexts, shipments are made in less-than-truckload (LTL) quantities from a facility to customers along a multiple-stop route. In the case of full truckload quantities, the cost of delivery is independent of the other deliveries made, whereas in the case of LTL quantities, the cost of delivery depends on the other customers on the route and the sequence in which customers are visited. Eilon, Watson-Gandy and Christofides (1971) were among the first to highlight the error introduced by approximating LTL shipments by full truckloads. During the past three decades, a sizeable body of literature has developed on integrated location/routing models.

Integrated location/routing problems combine three components of supply chain design: facility location, customer allocation to facilities and vehicle routing. Many different location/routing problems have been described in the literature, and they tend to be very difficult to solve since they merge two NP-hard problems: facility location and vehicle routing. Laporte (1988) reviews early work on location/routing problems; he summarizes the different types of formulations, solution algorithms and computational results of work published prior to 1988. More recently, Min, Jayaraman, and Srivastava (1998) develop a hierarchical taxonomy and classification scheme that they use to review the existing location/routing literature. They categorize papers in terms of problem characteristics and solution methodology. One means of classification is the number of layers of facilities. Typically, three-layer problems include flows from plants to distribution centers to customers, while two-layer problems focus on flows from distribution centers to customers.

An example of a three-layer location/routing problem is the formulation of Perl (1983) and Perl and Daskin (1985); their model extends the model of Geoffrion and Graves to include multiple stop tours serving the customer nodes but it is limited to a single commodity. Perl defines the following additional notation:

Inputs and sets.

P: set of points $= I \cup J$
d_{ij}: distance between node $i \in P$ and node $j \in P$
v_j: variable cost per unit processed by a facility at candidate facility site $j \in J$
t_j: maximum throughput for a facility at candidate facility site $j \in J$
S: set of supply points (analogous to plants in the Geoffrion and Graves model), indexed by s
c_{sj}: unit cost of shipping from supply point $s \in S$ to candidate facility site $j \in J$
K: set of candidate vehicles, indexed by k
σ_k: capacity of vehicle $k \in K$
τ_k: maximum allowable length of a route served by vehicle $k \in K$
α_k: cost per unit distance for delivery on route $k \in K$

Decision variables.

$$Z_{ijk} = \begin{cases} 1, & \text{if vehicle } k \in K \text{ goes directly from point } i \in P \text{ to point } j \in P, \\ 0, & \text{if not} \end{cases}$$

$W_{sj} =$ quantity shipped from supply source $s \in S$ to facility site $j \in J$

With this notation (and the notation defined previously), Perl (1983) formulates the following integrated location/routing problem:

$$\text{minimize} \quad \sum_{j \in J} f_j X_j + \sum_{s \in S} \sum_{j \in J} c_{sj} W_{sj} + \sum_{j \in J} v_j \sum_{i \in I} h_i Y_{ij} \sum_{j \in P} \sum_{i \in P} d_{ij} Z_{ijk} \tag{2.19}$$

$$\text{subject to} \quad \sum_{k \in K} \sum_{j \in P} Z_{ijk} = 1 \qquad \forall i \in I \tag{2.20}$$

$$\sum_{i \in I} h_i \sum_{j \in P} Z_{ijk} \leq \sigma_k \qquad \forall k \in K \tag{2.21}$$

$$\sum_{j \in P} \sum_{i \in P} d_{ij} Z_{ijk} \leq \tau_k \qquad \forall k \in K \tag{2.22}$$

$$\sum_{i \in V} \sum_{j \in \bar{V}} \sum_{k \in K} Z_{ijk} \geq 1 \qquad \forall \text{subsets } V \subset P \text{ such that } J \subset V \tag{2.23}$$

$$\sum_{j \in P} Z_{ijk} - \sum_{j \in P} Z_{jik} = 0 \qquad \forall i \in P; \forall k \in K \tag{2.24}$$

$$\sum_{j \in J} \sum_{i \in I} Z_{ijk} \leq 1 \qquad \forall k \in K \tag{2.25}$$

$$\sum_{s\in S} W_{sj} - \sum_{i\in I} h_i Y_{ij} = 0 \qquad \forall j \in J \qquad (2.26)$$

$$\sum_{s\in S} W_{sj} - t_j X_j \leq 0 \qquad \forall j \in J \qquad (2.27)$$

$$\sum_{m\in P} Z_{imk} + \sum_{h\in P} Z_{jhk} - Y_{ij} \leq 1 \qquad \forall j \in J; \forall i \in I; \forall k \in K \qquad (2.28)$$

$$X_j \in \{0,1\} \qquad \forall j \in J \qquad (2.29)$$

$$Y_{ij} \in \{0,1\} \qquad \forall i \in I; \forall j \in J \qquad (2.30)$$

$$Z_{ijk} \in \{0,1\} \qquad \forall i \in P; \forall j \in P; \forall k \in K \qquad (2.31)$$

$$W_{sj} \geq 0 \qquad \forall s \in S; \forall j \in J \qquad (2.32)$$

The objective function (2.19) minimizes the sum of the fixed facility location costs, the shipment costs from the supply points (plants) to the facilities, the variable facility throughput costs and the routing costs to the customers. Constraint (2.20) requires each customer to be on exactly one route. Constraint (2.21) imposes a capacity restriction for each vehicle, while constraint (2.22) limits the length of each route. Constraint (2.23) requires each route to be connected to a facility. The constraint requires that there be at least one route that goes from any set V (a proper subset of the points P that contains the set of candidate facility sites) to its complement $\bar{V}$, thereby precluding routes that only visit customer nodes. Constraint (2.24) states that any route entering node $i \in P$ also must exit that same node. Constraint (2.25) states that a route can operate out of only one facility. Constraint (2.26) defines the flow into a facility from the supply points in terms of the total demand that is served by the facility. Constraint (2.27) restricts the throughput at each facility to the maximum allowed at that site and links the flow variables and the facility location variables. Thus, if a facility is not opened, there can be no flow through the facility, which in turn (by constraint (2.26)) precludes any customers from being assigned to the facility. Constraint (2.28) states that if route $k \in K$ leaves customer node $i \in I$ and also leaves facility $j \in J$, then customer $i \in I$ must be assigned to facility $j \in J$. This constraint links the vehicle routing variables (Z_{ijk}) and the assignment variables (Y_{ij}). Constraints (2.29)–(2.32) are standard integrality and non-negativity constraints.

Even for small problem instances, the formulation above is a difficult mixed integer linear programming problem. Perl solves the problem using a three-phased heuristic. The first phase finds minimum cost routes. The second phase determines which facilities to open and how to allocate the routes from phase one to the selected facilities. The third phase attempts to improve the solution by moving customers between facilities and re-solving the routing problem with the set of open facilities fixed. The algorithm iterates between the second and third phases until the improvement at any iteration is less than some specified value. Wu, Low, and Bai (2002) propose a similar two-phase heuristic for the problem and test it on problems with up to 150 nodes.

Like the three-layer formulation of Perl, two-layer location/routing formulations (e.g., Laporte, Nobert and Pelletier, 1983; Laporte, Nobert and Arpin, 1986; and Laporte, Nobert and Taillefer, 1988) usually are based on integer linear programming formulations for the vehicle routing problem (VRP). Flow formulations of the VRP often are classified according to the number of indices of the flow variable: $X_{ij} = 1$ if a vehicle uses arc (i, j) or $X_{ijk} = 1$ if vehicle k uses arc (i, j). The size and structure of these formulations make them difficult to solve using standard integer programming or network optimization techniques. Motivated by the successful implementation of exact algorithms for set-partitioning-based routing models, Berger (1997) formulates a two-layer location/routing problem that closely resembles the classical fixed charge facility location problem. Unlike other location/routing problems, she formulates the routes in terms of paths, where a delivery vehicle may not be required to return to the distribution center after the final delivery is made. The model is appropriate in situations where the deliveries are made by a contract carrier or where the commodities to be delivered are perishable. In the latter case, the time to return from the last customer to the distribution center is much less important than the time from the facility to the last customer. Berger defines the following notation:

Inputs and sets.

P_j**:** set of feasible paths from candidate distribution center $j \in J$
c_{jk}**:** cost of serving the path $k \in P_j$
a^j_{ik}**:** 1 if delivery path $k \in P_j$ visits customer $i \in I$; 0 if not

Decision variables.

$$V_{jk} = \begin{cases} 1, & \text{if path } k \in P_j \text{ is operated out of distribution center } j \in J, \\ 0, & \text{if not.} \end{cases}$$

Note that there can be any number of restrictions on the feasible paths in set P_j; in fact, the more restrictive the conditions imposed on P_j are, the smaller the cardinality of P_j is. Restricting the total length of the paths, Berger formulates the following integrated location/routing model:

$$\text{minimize} \quad \sum_{j \in J} f_j X_j + \sum_{j \in J} \sum_{k \in P_j} c_{jk} V_{jk} \tag{2.33}$$

$$\text{subject to} \quad \sum_{j \in J} \sum_{k \in P_j} a^j_{ik} V_{jk} = 1 \qquad \forall i \in I \tag{2.34}$$

$$V_{jk} - X_j \leq 0 \qquad \forall j \in J; \forall k \in P_j \tag{2.35}$$

$$X_j \in \{0, 1\} \qquad \forall j \in J \tag{2.36}$$

$$V_{jk} \in \{0, 1\} \qquad \forall j \in J; \forall k \in P_j \tag{2.37}$$

The objective function (2.33) minimizes the sum of the facility location costs and the vehicle routing costs. Constraint (2.34) requires each demand node to be on one route. Constraint (2.35) states that a route can be assigned only to an open facility. Constraints (2.36) and (2.37) are standard integrality constraints.

Although the similarity between this location/routing model and the classical fixed charge location model (2.1) – (2.5) is striking, this model is much more difficult to solve for two reasons. First, the linear programming relaxation provides a weak lower bound. The linear programming relaxation typically has solutions in which the path variables are assigned very small fractional values and the location variables are assigned fractional variables large enough only to satisfy constraints (2.35). To strengthen the linear programming relaxation significantly, constraints (2.35) can be replaced by the following constraints:

$$\sum_{k \in P_j} a^j_{ik} V_{jk} - X_j \leq 0 \quad \forall i \in I; \forall j \in J \tag{2.38}$$

Consider a customer node $i \in I$ that is served (in part) using routes that emanate from facility $j \in J$. The first term of (2.38) is the sum of all route assignment variables that serve that customer and that are assigned to that facility. (In the linear programming relaxation, these assignment variables may be fractional). Thus, this sum can be thought of as the fraction of demand node $i \in I$ that is served out of facility $j \in J$. The constraint requires the location variable to be no smaller than the largest of these sums for customers assigned (in part) to routes emanating from the facility.

Second, there is an exponential number of feasible paths associated with any candidate facility, so complete enumeration of all possible columns of the problem is prohibitive. Instead, Berger develops a branch-and-price algorithm, which uses column generation to solve the linear programs at each node of the branch-and-bound tree. The pricing problem for the model decomposes into a set of independent resource-constrained shortest path problems.

The development and the use of location/routing models have been more limited than both facility location and vehicle routing models. In our view, the reason is that it is difficult to combine, in a meaningful way, facility location decisions, which typically are strategic and long term, and vehicle routing decisions, which typically are tactical and short term. The literature includes several papers that attempt to accommodate the fact that the set of customers to be served on a route may change daily, while the location of a distribution center may remain fixed for years. One approach is to define a large number of customers and to introduce a probability that each customer will require service on any day. Jaillet (1985, 1988) introduces this concept in the context of the probabilistic traveling salesman problem. Jaillet and Odoni (1988) provide an overview of this work and related probabilistic vehicle routing problems. The idea is extended to location/routing problems in Berman, Jaillet and Simchi-Levi (1995). Including different customer scenarios, however, increases the difficulty of the problem, so this literature tends to locate a single distribution center. In our view, the problem of approximating LTL vehicle tours in facility location problems without incurring the cost of solving an embedded vehicle routing or traveling salesman problem remains an open challenge worthy of additional research.

4. Integrated location/inventory models

The fixed charge location problem ignores the inventory impacts of facility location decisions; it deals only with the tradeoff between facility costs, which increase with the number of facilities located (call it N), and the average travel cost, which decreases approximately as the square root of N. Inventory costs increase approximately as the square root of N. As such, they introduce another force that tends to drive down the optimal number of facilities to locate. Baumol and Wolfe (1958) recognized the contribution of inventory to distribution costs over forty years ago when they stated, "standard inventory analysis suggests that, optimally, important inventory components will vary approximately as the square root of the number of shipments going through the warehouse" (p. 255). If the total number of shipments is fixed, the number through any warehouse is approximately equal to the total divided by N. According to Baumol and Wolfe, the cost at each warehouse is then proportional to the square root of this quantity. When the cost per warehouse is multiplied by N, we see that the total distribution cost varies approximately with the square root of N. This argument treats the cost of holding working or cycle stock; Eppen (1979) argued that safety stock costs also increase as the square root of N (assuming equal variance of demand at each customer and independence of customer demands).

While the contribution of inventory to distribution costs has been recognized for many years, only recently have we been able to solve the non-linear models that result from incorporating inventory decisions in facility location models. Shen (2000) and Shen, Coullard, and Daskin (2003) introduced a location model with risk pooling (LMRP). The model minimizes the sum of fixed facility location costs, direct transportation costs to the customers (which are assumed to be linear in the quantity shipped), working and safety stock inventory costs at the distribution centers and shipment costs from a plant to the distribution center (which may include a fixed cost per shipment). The last two quantities — the inventory costs at the distribution centers and the shipment costs of goods to the distribution centers — depend on the allocation of customers to the distribution centers. Shen introduces the following additional notation:

Inputs and sets.

μ_i, σ_i^2: mean and variance of the demand per unit time at customer $i \in I$

c_{ij}: a term that captures the annualized unit cost of supplying customer $i \in I$ from facility $j \in J$ as well as the variable shipping cost from the supplier to facility $j \in J$

ρ_j: a term that captures the fixed order costs at facility $j \in J$ as well as the fixed transport costs per shipment from the supplier to facility $j \in J$ and the working inventory carrying cost at facility $j \in J$

ω_j: a term that captures the lead time of shipments from the supplier to facility $j \in J$ as well as the safety stock holding cost

With this notation, Shen formulates the LMRP as follows:

$$\text{minimize} \quad \sum_{j\in J} f_j X_j + \sum_{j\in J}\sum_{i\in I} c_{ij}\mu_i Y_{ij} + \sum_{j\in J} \rho_j \sqrt{\sum_{i\in I} \mu_i Y_{ij}} + \sum_{j\in J} \omega_j \sqrt{\sum_{i\in I} \sigma_i^2 Y_{ij}} \qquad (2.39)$$

$$\text{subject to} \quad \sum_{j\in J} Y_{ij} = 1 \qquad \forall i \in I \qquad (2.2)$$

$$Y_{ij} - X_j \le 0 \qquad \forall i \in I; \forall j \in J \qquad (2.3)$$

$$X_j \in \{0,1\} \qquad \forall j \in J \qquad (2.4)$$

$$Y_{ij} \ge 0 \qquad \forall i \in I; \forall j \in J \qquad (2.5)$$

The first term of the objective function (2.39) represents the fixed facility location costs. The second term captures the cost of shipping from the facilities to the customers as well as the variable shipment costs from the supplier to the facilities. The third term represents the working inventory carrying costs which include any fixed (per shipment) costs of shipping from the supplier to the facilities. The final term represents the safety stock costs at the facilities. Note that the objective function is identical to that of the fixed charge location problem (2.1) with the addition of two non-linear terms, the first of which captures economies of scale regarding fixed ordering and shipping costs and the second of which captures the risk pooling associated with safety stocks. Also note that the constraints of the LMRP are identical to those of the fixed charge location problem.

Shen (2000) and Shen, Coullard, and Daskin (2003) recast this model as a set covering problem where the sets contain customers to be served by facility $j \in J$. As in Berger's location/routing model, the number of possible sets is exponentially large. Thus, they propose solving the problem using column generation. The pricing problems are non-linear integer programs, but their structure allows for a low-order polynomial solution algorithm. Shen assumes that the variance of demand is proportional to the mean. If demands are Poisson, this assumption is exact and not an approximation. With this assumption, he is able to collapse the final two terms in the objective function into one term. The resulting pricing problems can then be solved in $O(|I| \log |I|)$ time for each candidate facility and in $O(|J||I| \log |I|)$ time for all candidate facilities at each iteration of the column generation algorithm. Shu, Teo, and Shen (2004) show that the pricing problem with two square root terms (i.e., without assuming that the variance-to-mean ratio is constant for all customers) can be solved in $O(|I|^2 \log |I|)$ time. Daskin, Coullard, and Shen (2002) develop a Lagrangian relaxation algorithm for this model and found it to be slightly faster than the column generation method.

One of the important qualitative findings from Shen's model is that, as inventory costs increase as a percentage of the total cost, the number of facilities located by the LMRP is significantly smaller than the number that would have been sited by the uncapacitated fixed charge location model, which ignores the risk pooling effects of inventory management. Shen and Daskin (2003) extend the model above to account for customer service considerations. As customer service increases in importance, the number of facilities used in the optimal solution grows, eventually approaching and even exceeding the number used in the uncapacitated fixed charge model.

Several joint location/inventory models appeared in the literature prior to Shen's work. Barahona and Jensen (1998) solve a location problem with a fixed cost for stocking a given product at a DC. Erlebacher and Meller (2000) use various heuristic techniques to solve a joint location/inventory problem with a highly non-linear objective function. Teo, Ou, and Goh (2001) present a $\sqrt{2}$-approximation algorithm for the problem of choosing DCs to minimize location and inventory costs, ignoring transportation costs. Nozick and Turnquist (2001a,b) present models that, like Shen's model, incorporate inventory considerations into the fixed charge location problem; however, they assume that inventory costs are linear, rather than concave, and DC-customer allocations are made based only on distance, not inventory.

Ozsen, Daskin, and Coullard (2003) have extended the LMRP to incorporate capacities at the facilities. Capacities are modeled in terms of the maximum (plausible) inventory accumulation during a cycle between order receipts. This model is considerably harder to solve than is its uncapacitated cousin. However, it highlights an important new dimension in supply chain operations that is not captured by the traditional capacitated fixed charge location model. In the traditional model, capacity is typically measured in terms of throughput per unit time. However, this value can change as the number of inventory turns per unit time changes. Thus, the measure of capacity in the traditional model is often suspect. Also, using the traditional model, there are only two ways to deal with capacity constraints as demand increases: build more facilities or reallocate customers to more remote facilities that have excess capacity. In the capacitated version of the LMRP, a third option is available, namely ordering more frequently in smaller quantities. By incorporating this extra dimension of choice, the capacitated LMRP is more likely to reflect actual managerial options than is the traditional fixed charge location model.

To some extent, merging inventory management with facility location decisions suffers from the same conceptual problems as merging vehicle routing with location. Inventory decisions, as argued above, can be revised much more frequently than can facility location decisions. Nevertheless, there are three important reasons for research to continue in the area of integrated inventory/location modeling. First, early results suggest that the location decisions that are made when inventory is considered can be radically different from those that would be made by a procedure that fails to account for inventory. Second, as indicated above, the capacitated LMRP better models actual facility capacities than does the traditional fixed charge location model, as it introduces the option of ordering more often to accommodate increases in demand. Third, we can solve fairly large instances of the integrated location/inventory model outlined above. In particular, the Lagrangian approach can often solve problems with 600 customers and 600 candidate facility sites in a matter of minutes on today's desktop computers.

5. Planning under uncertainty

Long-term strategic decisions like those involving facility locations are always made in an uncertain environment. During the time when design decisions are in effect, costs and demands may change drastically. However, classical facility location models like the fixed charge location problem treat data as though they were known and deterministic, even though ignoring data uncertainty can result in highly sub-optimal solutions. In this section, we discuss approaches to facility location under uncertainty that have appeared in the literature.

Most approaches to decision making under uncertainty fall into one of two categories: stochastic programming or robust optimization. In stochastic programming, the uncertain parameters are described by discrete scenarios, each with a given probability of occurrence; the objective is to minimize the expected cost. In robust optimization, parameters may be described either by discrete scenarios or by continuous ranges; no probability information is known, however, and the objective is typically to minimize the worst-case cost or regret. (The regret of a solution under a given scenario is the difference between the objective function value of the solution under the scenario and the optimal objective function value for that scenario.) Both approaches seek solutions that perform well, though not necessarily optimally, under any realization of the data. We provide a brief overview of the literature on facility location under uncertainty here. For a more comprehensive review, the reader is referred to Owen and Daskin (1998) or Berman and Krass (2002).

Sheppard (1974) was one of the first authors to propose a stochastic approach to facility location. He suggests selecting facility locations to minimize the expected cost, though he does not discuss the issue at length. Weaver and Church (1983) and Mirchandani, Oudjit, and Wong (1985) present a multi-scenario version of the P-median problem. Their model can be translated into the context of the fixed charge location problem as follows. Let S be a set of scenarios. Each scenario $s \in S$ has a probability q_s of occurring and specifies a realization of random demands (h_{is}) and travel costs (c_{ijs}). Location decisions must be made now, before it is known which scenario will occur. However, customers may be assigned to facilities after the scenario is known, so the Y variables are now indexed by a third subscript, s. The objective is to minimize the total expected cost. The stochastic fixed charge location problem is formulated as follows:

$$\text{minimize} \quad \sum_{j \in J} f_j X_j + \sum_{s \in S} \sum_{j \in J} \sum_{i \in I} q_s h_{is} c_{ijs} Y_{ijs} \tag{2.40}$$

$$\text{subject to} \quad \sum_{j \in J} Y_{ijs} = 1 \qquad \forall i \in I; \forall s \in S \tag{2.41}$$

$$Y_{ijs} - X_j \leq 0 \qquad \forall i \in I; \forall j \in J; \forall s \in S \tag{2.42}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{2.43}$$

$$Y_{ijs} \geq 0 \qquad \forall i \in I; \forall j \in J; \forall s \in S \tag{2.44}$$

The objective function (2.40) computes the total fixed cost plus the expected transportation cost. Constraint (2.41) requires each customer to be assigned to a facility in each scenario. Constraint (2.42) requires that facility to be open. Constraints (2.43) and (2.44) are integrality and non-negativity constraints. The key to solving this model and the P-median-based models formulated by Weaver and Church (1983) and Mirchandani, Oudjit, and Wong (1985) is recognizing that the problem can be treated as a deterministic problem with $|I|\,|S|$ customers instead of $|I|$.

Snyder, Daskin, and Teo (2003) consider a stochastic version of the LMRP. Other stochastic facility location models include those of Louveaux (1986), França and Luna (1982), Berman and LeBlanc (1984), Carson and Batta (1990), and Jornsten and Bjorndal (1994).

Robust facility location problems tend to be more difficult computationally than stochastic problems because of their minimax structure. As a result, the literature on robust facility location generally falls into one of two categories: analytical results and polynomial-time algorithms for restricted problems like 1-median problems or P-medians on tree networks (see Chen and Lin, 1998; Burkhard and Dollani, 2001; Vairaktarakis and Kouvelis, 1999; and Averbakh and Berman, 2000) and heuristics for more general problems (Serra, Ratick, and ReVelle, 1996; Serra and Marianov, 1998; and Current, Ratick, and ReVelle, 1997).

Solutions to the stochastic fixed charge problem formulated above may perform well in the long run but poorly in certain scenarios. To address this problem, Snyder and Daskin (2003) combine the stochastic and robust approaches by finding the minimum-expected-cost solution to facility location problems subject to an additional constraint that the relative regret in each scenario is no more than a specified limit. They show empirically that by reducing this limit, one obtains solutions with substantially reduced maximum regret without large increases in expected cost. In other words, there are a number of near-optimal solutions to the fixed charge problem, many of which are much more robust than the true optimal solution.

6. Location models with facility failures

Once a set of facilities has been built, one or more of them may from time to time become unavailable — for example, due to inclement weather, labor actions, natural disasters, or changes in ownership. These facility "failures" may result in excessive transportation costs as customers previously served by these facilities must now be served by more distant ones. In this section, we discuss models for choosing facility locations to minimize fixed and transportation costs while also hedging against failures within the system. We call the ability of a system to perform well even when parts of the system have failed the "reliability" of the system. The goal, then, is to choose facility locations that are both inexpensive and reliable.

The robust facility location models discussed in the previous section hedge against uncertainty in the problem data. By contrast, reliability models hedge against uncertainty in the solution itself. Another way to view the distinction in the context of supply chain design is that robustness is concerned with "demand-side" uncertainty (uncertainty in demands, costs, or other parameters), while reliability is concerned with "supply-side" uncertainty (uncertainty in the availability of plants or distribution centers).

The models discussed in this section are based on the fixed charge location problem; they address the tradeoff between *operating cost* (fixed location costs and day-to-day transportation cost — the classical fixed charge problem objective) and *failure cost* (the transportation cost that results after a facility has failed). The first model considers the *maximum* failure cost that can occur when a single facility fails, while the second model considers the *expected* failure cost given a fixed probability of failure. The strategy behind both formulations is to assign each customer to a *primary* facility (which serves it under normal conditions) and one or more *backup* facilities (which serve it when the primary facility has failed). Note that although we refer to primary and backup *facilities*, "primariness" is a characteristic of assignments, not facilities; that is, a given facility may be a primary facility for one customer and a backup facility for another.

In addition to the notation defined earlier, let

$$Y_{ijk} = \begin{cases} 1, & \text{if facility } j \in J \text{ serves as the primary facility and facility} \\ & k \in J \text{ serves as the secondary facility for customer } i \in I, \\ 0, & \text{if not,} \end{cases}$$

and let V be a desired upper bound on the failure cost that may result if a facility fails. Snyder (2003) formulates the maximum-failure-cost reliability problem as follows:

$$\text{minimize} \quad \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} \sum_{k \in J} h_i c_{ij} Y_{ijk} \tag{2.45}$$

$$\text{subject to} \quad \sum_{j \in J} \sum_{k \in J} Y_{ijk} = 1 \qquad \forall\, i \in I \tag{2.46}$$

$$\sum_{k \in J} Y_{ijk} \leq X_j \qquad \forall\, i \in I; \forall\, j \in J \tag{2.47}$$

$$Y_{ijk} \leq X_k \qquad \forall\, i \in I; \forall\, j \in J; \forall\, k \in J \tag{2.48}$$

$$\sum_{i \in I} \sum_{\substack{k \in J \\ k \neq j}} \sum_{l \in J} h_i c_{ik} Y_{ikl} + \sum_{i \in I} \sum_{k \in J} h_i c_{ik} Y_{ijk} \leq V \qquad \forall\, j \in J \tag{2.49}$$

$$Y_{ijj} = 0 \qquad \forall\, i \in I; \forall\, j \in J \tag{2.50}$$

$$X_j \in \{0, 1\} \qquad \forall\, j \in J \tag{2.51}$$

$$Y_{ijk} \geq 0 \qquad \forall\, i \in I; \forall\, j \in J; \forall\, k \in J \tag{2.52}$$

The objective function (2.45) sums the fixed cost and transportation cost to customers from their primary facilities. (The summation over k is necessary to determine the assignments, but the objective function does not depend on the backup assignments.) Constraint (2.46) requires each customer to be assigned to one primary and one backup facility. Constraints (2.47) and (2.48) prevent a customer from being assigned to a primary or a backup facility, respectively, that has not been opened. (The summation on the left-hand side of (2.47) can be replaced by Y_{ijk} without affecting the IP solution, but doing so considerably weakens the LP bound.) Constraint (2.49) is the reliability constraint and requires the failure cost for facility j to be no greater than V. The first summation computes the cost of serving each customer from its primary facility if its primary facility is not j, while the second summation computes the cost of serving customers assigned to j as their primary facility from their backup facilities. Constraint (2.50) requires a customer's primary facility to be different from its backup facility, and constraints (2.51) and (2.52) are standard integrality and non-negativity constraints. This model can be solved for small instances using an off-the-shelf IP solver, but larger instances must be solved heuristically.

The expected-failure-cost reliability model (Snyder and Daskin, 2004) assumes that multiple facilities may fail simultaneously, each with a given probability q of failing. In this case, a single backup facility is insufficient, since a customer's primary and backup facilities may both fail. Therefore, we define

$$Y_{ijr} = \begin{cases} 1, & \text{if facility } j \in J \text{ serves as the level-}r\text{ facility for} \\ & \text{customer } i \in I, \\ 0, & \text{if not.} \end{cases}$$

A "level-r" assignment is one for which there are r closer facilities that are open. If $r = 0$, this is a primary assignment; otherwise it is a backup assignment. The objective is to minimize a weighted sum of the operating cost (the fixed charge location problem objective) and the expected failure cost, given by

$$\sum_{i \in I} \sum_{j \in J} \sum_{r=0}^{|J|-1} h_i c_{ij} q^r (1-q) Y_{ijr}.$$

Each customer i is served by its level-r facility (call it j) if the r closer facilities have failed (this occurs with probability q^r) and if j itself has not failed (this occurs with probability $1 - q$). The full model is omitted here. This problem can be solved efficiently using Lagrangian relaxation.

Few firms would be willing to choose a facility location solution that is, say, twice as expensive as the optimal solution to the fixed charge problem just to hedge against occasional disruptions to the supply chain. However, Snyder and Daskin (2004) show empirically that it often costs very little to "buy" reliability: like robustness, reliability can be improved substantially with only small increases in cost.

7. Conclusions and directions for future work

Facility locations decisions are critical to the efficient and effective operation of a supply chain. Poorly placed plants and warehouses can result in excessive costs and degraded service no matter how well inventory policies, transportation plans, and information sharing policies are revised, updated, and optimized. At the heart of many supply chain facility location models is the fixed charge location problem. As more facilities are located, the facilities tend to be closer to customers resulting in lower transport costs, but higher facility costs. The fixed charge facility location problem finds the optimal balance between fixed facility costs and transportation costs. Three important extensions of the basic model consider (1) facility capacities and single sourcing requirements, (2) multiple echelons in the supply chain, and (3) multiple products.

The fixed charge location problem, as well as these extensions, assume that shipments from the warehouses or distribution centers to the customers or retailers are made in truckload quantities. In reality, distribution to customers is often performed using less-than-truckload routes that visit multiple customers. This chapter reviewed two different approaches to formulating integrated location/routing models. However, as indicated above, these approaches suffer from the fundamental problem that facility locations are typically determined at a strategic level while vehicle routes are optimized at the operational level. In other words, the set of customers and their demands may change daily resulting in daily route changes, while the facilities are likely to be fixed for years. We believe that additional research is needed to find improved ways of approximating the impact of less-than-truckload deliveries on facility location costs without embedding a vehicle routing problem (designed to serve one realization of customer demands) in the facility location model.

Incorporating inventory decisions in facility location models appears to be critical for supply chain modeling. As early as 1958, researchers recognized that inventory costs would tend to increase with the square root of the number of facilities used. Only recently, however, have nonlinear models that approximate this relationship between inventory costs and location decisions been formulated and solved optimally. While we believe that these models represent an important step forward in location modeling for supply chain problems, considerable additional research is needed. In particular, researchers should attempt to incorporate more sophisticated inventory models, including multi-item inventory models and models that account for inventory accumulation at all echelons of the supply chain. Heuristic approaches to the multi-item problem have recently been proposed by Balcik (2003) and an optimal approach has been suggested by Snyder (2003). The latter model, however, assumes that items are ordered separately, resulting in individual fixed order costs for each commodity purchased.

Finally, since facility location decisions are inherently strategic and long term in nature, supply chain location models must account for the inherent uncertainty surrounding future conditions. We have reviewed a number of scenario-based location models as well as models that account for unreliability in the facilities themselves. This too is an area worthy of considerable additional research. For example, generating scenarios that capture future uncertainty and the relationships between uncertain parameters is one critical area of research. Reliability-based location models for supply chain management are still in their infancy. In fact, it is not immediately clear how to marry reliability modeling approaches and the integrated location/inventory models we reviewed, since the nonlinearities introduced by the inventory terms complicate the computation of failure costs. In this regard, the more general techniques of stochastic programming (Birge and Louveaux, 1997) may ultimately prove fruitful.

References

Al-Sultan, K.S. and Al-Fawzan, M.A. (1999). A tabu search approach to the uncapacitated facility location problem. *Annals of Operations Research*, 86:91–103.

Averbakh, I. and Berman, O. (2000). Minmax regret median location on a network under uncertainty. *INFORMS Journal on Computing*, 12(2):104–110.

Balcik, B. (2003). *Multi-Item Integrated Location/Inventory Problem.* M.S. Thesis, Department of Industrial Engineering, Middle East Technical University.

Balinski, M.L. (1965). Integer programming: Methods, uses, computation. *Management Science*, 12:253–313.

Barahona, F. and Jensen, D. (1998). Plant location with minimum inventory. *Mathematical Programming*, 83:101–111.

Baumol, W.J. and Wolfe, P. (1958). A warehouse-location problem. *Operations Research*, 6:252–263.

Benders, J.F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4:238–252.

Berger, R.T. (1997). *Location-Routing Models for Distribution System Design.* Ph.D. Dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Berman, O. and LeBlanc, B. (1984). Location-relocation of mobile facilities on a stochastic network. *Transportation Science*, 18(4):315–330.

Berman, O., Jaillet, P., and Simchi-Levi, D. (1995). Location-routing problems with uncertainty. In: Z. Drezner (ed.), *Facility Location: A Survey of Applications and Methods*, Springer, New York.

Berman, O. and Krass, D. (2002). Facility location problems with stochastic demands and congestion. In: Z. Drezner and H.W. Hamacher (eds.), *Facility Location: Applications and Theory*, pp. 331–373, Springer, New York.

Birge, J.R. and Louveaux, F. (1997). *Introduction to Stochastic Programming.* Springer, New York.

Burkhard, R.E. and Dollani, H. (2001). Robust location problems with pos/neg weights on a tree. *Networks*, 38(2):102–113.

Carson, Y.M. and Batta, Y. (1990). Locating an ambulance on the Amherst Campus of the State University of New York at Buffalo. *Interfaces*, 20(5):43–49.

Chen, B.T. and Lin, C.S. (1998). Minimax-regret robust 1-median location on a tree. *Networks*, 31:93–103.

Cornuéjols, G., Nemhauser, G.L., and Wolsey, L.A. (1990). The uncapacitated facility location problem. In: P.B. Mirchandani and R.L. Francis (eds.), *Discrete Location Theory*, pp. 119–171, Wiley, New York.

Current, J., Ratick, S., and ReVelle, C. (1997). Dynamic facility location when the total number of facilities is uncertain: A decision analysis approach. *European Journal of Operational Research*, 110(3):597–609.

Daskin, M.S. (1995). *Network and Discrete Location: Models, Algorithms and Applications.* John Wiley and Sons, Inc., New York.

Daskin, M.S., Coullard, C.R., and Shen, Z.-J.M. (2002). An inventory-location model: Formulation, solution algorithm and computational results. *Annals of Operations Research*, 110:83–106.

Daskin, M.S. and Jones, P.C. (1993). A new approach to solving applied location/allocation problems. *Microcomputers in Civil Engineering*, 8:409–421.

Eilon, S., Watson-Gandy, C.D.T., and Christofides, N. (1971). *Distribution Management: Mathematical Modeling and Practical Analysis.* Hafner Publishing Co., NY.

Eppen, G. (1979). Effects of centralization on expected costs in a multi-location newsboy problem. *Management Science*, 25(5):498–501.

Erlebacher, S.J. and Meller, R.D. (2000). The interaction of location and inventory in designing distribution systems. *IIE Transactions*, 32:155–166.

Erlenkotter, D. (1978). A dual-based procedure for uncapacitated facility location. *Operations Research*, 26:992–1009.

França, P.M. and Luna, H.P.L. (1982). Solving stochastic transportation-location problems by generalized Benders decomposition. *Transportation Science*, 16:113–126.

Galvão, R.D. (1993). The use of Lagrangean relaxation in the solution of uncapacitated facility location problems. *Location Science*, 1(1):57–79.

Geoffrion, A.M. (1974). Lagrangian relaxation for integer programming. *Mathematical Programming Study*, 2:82–114.

Geoffrion, A.M. and Graves, G.W. (1974). Multicommodity distribution system design by Benders decomposition. *Management Science*, 20(5):822–844.

Geoffrion, A.M. and Powers, R.F. (1980). Facility location analysis is just the beginning (if you do it right). *Interfaces*, 10(2):22–30.

Glover, F. (1989). Tabu search — Part I. *ORSA Journal on Computing*, 1(3):190–206.

Glover, F. (1990). Tabu search — Part II. *ORSA Journal on Computing*, 2(1):4–32.

Glover, F. and Laguna, M. (1997). *Tabu Search.* Kluwer Academic Publishers, Boston, MA.

Hansen, P. and Mladenović, N. (1997). Variable neighborhood search for the p-median. *Location Science*, 5(4):207–226.

Hakimi, S.L. (1964). Optimum location of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12:450–459.

Hakimi, S.L. (1965). Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, 13:462–475.

Jaillet, P. (1985). The probabilistic traveling salesman problem. Technical Report 185, Operations Research Center, M.I.T., Cambridge, MA.

Jaillet, P. (1988). A priori solution of a traveling salesman problem in which a random subset of the customers are visited. *Operations Research*, 36:929–936.

Jaillet, P. and Odoni, A. (1988). Probabilistic vehicle routing problems. In: B.L. Golden and A.A. Assad (eds.), *Vehicle Routing: Methods and Studies*, pp. 293–318, North-Holland, Amsterdam.

Jornsten, K. and Bjorndal, M. (1994). Dynamic location under uncertainty. *Studies in Regional and Urban Planning*, 3:163–184.

Krarup, J. and Pruzan, P.M. (1983). The simple plant location problem: Survey and synthesis. *European Journal of Operational Research*, 12:36–81.

Laporte, G. (1988). Location routing problems. In: B.L. Golden and A.A. Assad (eds.), *Vehicle Routing: Methods and Studies*, pp. 163–197, North-Holland, Amsterdam.

Laporte, G., Nobert, Y., and Arpin, D. (1986). An exact algorithm for solving a capacitated location-routing problem. *Annals of Operations Research*, 6:293–310.

Laporte, G., Nobert, Y., and Pelletier, J. (1983). Hamiltonian location problems. *European Journal of Operational Research*, 12:82–89.

Laporte G., Nobert, Y., and Taillefer, S. (1988). Solving a family of multi-depot vehicle routing and location-routing problems. *Transportation Science*, 22:161–172.

Louveaux, F.V. (1986). Discrete stochastic location models. *Annals of Operations Research*, 6:23–34.

Maranzana, F.E. (1964). On the location of supply points to minimize transport costs. *Operational Research Quarterly*, 15:261–270.

Min, H., Jayaraman, V., and Srivastava, R. (1998). Combined location-routing problems: A synthesis and future research directions. *European Journal of Operational Research*, 108:1–15.

Mirchandani, P.B., Oudjit, A., and Wong, R.T. (1985). Multidimensional extensions and a nested dual approach for the m-median problem. *European Journal of Operational Research*, 21:121–137.

Nozick, L.K. and Turnquist, M.A. (2001a). Inventory, transportation, service quality and the location of distribution centers. *European Journal of Operational Research*, 129:362–371.

Nozick, L.K. and Turnquist, M.A. (2001b). A two-echelon inventory allocation and distribution center location analysis. *Transportation Research Part E*, 37:421–441.

Owen, S.H. and Daskin, M.S. (1998). Strategic facility location: A review. *European Journal of Operational Research*, 111:423–447.

Ozsen, L., Daskin, M.S., and Coullard, C.R. (2003). Capacitated facility location model with risk pooling. Submitted for publication.

Perl, J. (1983). *A Unified Warehouse Location-Routing Analysis.* Ph.D. Dissertation, Department of Civil Engineering, Northwestern University, Evanston, IL.

Perl, J. and Daskin, M.S. (1985). A warehouse location-routing problem. *Transportation Research*, 19B(5):381–396.

Serra, D. and Marianov, V. (1998). The p-median problem in a changing network: The case of Barcelona. *Location Science*, 6:383–394.

Serra, D., Ratick, S., and ReVelle, C. (1996). The maximum capture problem with uncertainty. *Environment and Planning B*, 23:49–59.

Shen, Z.J. (2000). *Efficient Algorithms for Various Supply Chain Problems.* Ph.D. Dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Shen, Z.-J.M., Coullard, C.R., and Daskin, M.S. (2003). A joint location-inventory model. *Transportation Science*, 37(1):40–55.

Shen, Z.-J.M. and Daskin, M.S. (2003). Tradeoffs between customer service and cost in an integrated supply chain design framework. Submitted to *Manufacturing and Service Operations Management.*

Sheppard, E.S. (1974). A Conceptual Framework for Dynamic Location-Allocation Analysis. *Environment and Planning A*, 6:547–564.

Shu, J., Teo, C.-P., and Shen, Z.-J.M. (2004). Stochastic transportation-inventory network design. To appear in *Operations Research.*

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E. (2003). *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies.* Second Edition, McGraw-Hill/Irwin, Boston, MA.

Snyder, L.V. (2003). *Supply Chain Robustness and Reliability: Models and Algorithms.* Ph.D. Dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208.

Snyder, L.V. and Daskin, M.S. (2003). Stochastic p-robust location problems. Working paper.

Snyder, L.V. and Daskin, M.S. (2004). Reliability models for facility location: The expected failure cost case. Under revision for *Transportation Science.*

Snyder, L.V., Daskin, M.S., and Teo, C.-P. (2003). The stochastic location model with risk pooling. Submitted for publication.

Teitz, M.B. and Bart, P. (1968). Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research*, 16:955–961.

Teo, C.-P., Ou, J., and Goh, M. (2001). Impact on inventory costs with consolidation of distribution centers. *IIE Transactions*, 33(2):99–110.

Vairaktarakis, G.L. and Kouvelis, P. (1999). Incorporation dynamic aspects and uncertainty in 1-median location problems. *Naval Research Logistics*, 46(2):147–168.

Weaver, J.R. and Church, R.L. (1983). Computational procedures for location problems on stochastic networks. *Transportation Science*, 17:168–180.

Wu, T.-H., Low, C., and Bai, J.-W. (2002). Heuristic solutions to multi-depot location-routing problems. *Computers and Operations Research*, 29:1393–1415.

Chapter 3

DISTRIBUTION CENTRES IN SUPPLY CHAIN OPERATIONS

James K. Higginson
James H. Bookbinder

Abstract A *supply chain* consists of all flows and transformations from simple raw materials to purchase of end-items by consumers. Various network nodes perform component fabrication, product assembly or sales. These activities, however, require logistical support, e.g., storage of intermediate or finished goods; consolidation of orders; and transportation. The term, *Distribution Centre* (DC) denotes a supply-chain node that furnishes coordination of that sort.

This chapter highlights seven roles played by a DC. We discuss the measurement of distribution-centre performance, and the information required to manage a DC. These need to be approached differently, depending on the facility's function or role.

1. Introduction

Quinn (2000) has suggested that Transportation is a "forgotten area of supply chain management." That is, analysts have put all their attention into designing the perfect network, and have worried too little about managing the flows of products between nodes. It could be argued that Distribution Centres (DCs) are another forgotten area. A review of supply chain management books published from the late 1990s onward reveals that many do not discuss, nor even include in the index, material on distribution centres or warehouses. Researchers seem to take them for granted, assuming that a DC will be there when needed, offering exactly the services required.

This chapter attempts to fill the gap. A distribution centre can play a number of major roles in a supply chain. Beginning in Section 3, we will examine each of them, and the corresponding issues and decisions required. But let us first consider the "big picture."

2. What is a distribution centre?

Warehouses and DCs are important nodes in a supply network; they perform valuable functions that support the movement of materials. Storing goods (temporarily or longer), processing products, de-aggregating vehicle loads, creating SKU assortments, and assembling shipments are all activities commonly performed in these facilities. (OR applications to warehousing are discussed by Cormier (2005) elsewhere in this volume.)

With the increase in the number and types of services offered by a warehouse, the distinction between it and a distribution centre has become cloudy and ignored by many authors and researchers. A DC is, in fact, a *specific* type of warehouse. Coyle et al. (2003), for example, define a distribution centre to be "a post-production warehouse for finished goods held for distribution." Frazelle (2002) refers to distribution centres as distribution *warehouses* (as does Ballou, 2004), and defines them as facilities that "accumulate and consolidate products from various points of manufacture within a single firm, or from several firms, for combined shipment to common customers."

This chapter adopts the common definition of a DC to be a type of warehouse where the storage of goods is limited or non-existent. As a result, distribution centres focus on product movement and throughput (receiving, putaway, order picking, order assembly, and shipping), and information collection and reporting (throughput and utilization, transportation documentation, loss and damage claim support), rather than storage. Daww (1995) lists several other differences between warehouses and DCs. Two fit the definition we will use in this chapter: "Warehouses store all products; distribution centres hold minimum inventories, and of predominantly high-demand items. Warehouses handle most products in four cycles (receive, store, pick, and ship); DCs handle most products in two (receive and ship)." (Bancroft, 1991, discusses changes required in a facility to move its operations away from a warehouse and towards a distribution centre.) Nonetheless, many of the works cited in this chapter use interchangeably the two terms, warehouse and DC.

Since the 1980s, three supply-chain trends have had a major impact on these facilities:

- reduction in the number of warehouses
- greater emphasis on the *flow* of goods, rather than their storage
- increased outsourcing of warehouse/distribution centre activities.

Early supply-chain initiatives changed the emphasis of logistics operations from productivity improvement to inventory reduction. Delaney (1991), for example, reported a $200 billion decrease in inventory investment in the United States between 1980 and 1989. Others (e.g., Ackerman and Brewer 2001) have noted that the largest saving in logistics costs during the last thirty years has been due to reduced stock. Diminished inventories allow for the closing of facilities, which encourages inventory centralization, closer control of safety stocks, and elimination of obsolete and slow-moving items. This further lessens the need to maintain inventory at so many locations, which changes the role of some facilities from storage to product flow; that is, from warehouse to distribution centre.

The impact on overall inventory of fewer stock-keeping locations has been analysed by a number of researchers, including Caron and Marchet (1996); Bordley et al. (1999); Teo et al. (2001); Kim (2002); and Simchi-Levi et al. (2003). Enhanced communication and transportation have further reduced the need for DCs and warehouses. Ackerman and Brewer (2001) note that many of the distribution centres established in the latter part of the 20th century were aimed at strengthening customer service, but "the substantial improvements in delivery capabilities... have made it possible for some distributors substantially to reduce the number of distribution centres without compromising customer service."

A second factor leading to fewer warehouses and DCs is the outsourcing of logistics activities. During the late 1980s and the 1990s, many North American manufacturers spun off their in-house logistics activities to concentrate on core business operations. Warehousing typically is at or near the top of the list of logistics functions commonly outsourced (e.g., Coyle et al., 2003). This allows third-party logistics providers to consolidate the warehousing/distribution centre operations of several clients in a small number of facilities. Each client benefits from the third party's economies of scale in DC and transportation activities.

It is not clear whether the number of DCs or warehouses required by an organization will continue to decrease. The last few years have seen enhanced demand for warehouse and distribution-centre space, due in part to the greater range of services being carried out in modern DCs, and the shift to smaller customer orders (especially those from e-commerce). Activities traditionally performed in factories (such as packaging and labelling, light assembly, and product localization), and services required by e-business (such as invoicing, billing credit cards, arranging transportation, and handling customer returns), which previously were performed by wholesalers and distributors, are now common in DCs.

According to Planeta (2001), the modern Canadian distribution centre is "usually located within close proximity to highway access, has a ceiling height clearance of 28 feet or more, and contains only a small amount of office space... These facilities are typically designed to have one shipping door per 10,000 square feet of warehouse space and a maximum depth (dock to wall) of 350 feet.... [and] bay sizes in the range of 36 feet by 40 feet, which generally produce the most efficient 'generic' warehouse layout."

The distribution centre of the future will be larger, with a greater emphasis on reducing activity times, again as noted by Planeta (2001):

> Facilities are being built to allow for shipping on two sides, effectively turning them into large cross-docks with significant warehouse capacity. These facilities allow ample room for outside storage of trailer equipment and a higher shipping door ratio. Today's buildings are getting higher, have better shipping capacity, and more efficient mechanical systems.... Unless the product flow rate is extremely high, this type of facility may not be the most efficient for the typical user.

One consideration in determining the feasibility of larger distribution centres is discussed by Footlik (1999). An organization often will operate DCs of different sizes; Lee (1996) has incorporated this fact in a mathematical model for facility location. The distribution centre of the future most likely will be owned and operated by a third party (Ackerman and Brewer, 2001). And although modern DCs tend to be highly automated, many activities remain quite labour-dependent.

Reports indicate that in the future, it will not be uncommon for 50 percent of distribution centre employees to be temporary (Reynolds, 2003). If so, the outbound portion of the supply chain will already be set to follow a "chase strategy," in the sense of aggregate production planning. Thus, the major challenge to warehouses and DCs, both today and tomorrow, will relate to workforce issues such as staffing, training, scheduling, and job design (Ackerman and Brewer, 2001).

Trappey and Ho (2002) present an approach to managing employees in distribution centres. They discuss an information system add-on that assigns pick lists to employees, and goods to trucks, in a DC. The assignment routines, based on simple heuristics, are designed to integrate with the human-resource and order-management modules of an ERP system (See Section 5).

We have seen in this section a number of ways to define a DC. Let us now turn attention to the activities there.

3. Roles of a distribution centre in the supply chain

The article of Min and Melachrinoudis (1999) is concerned with a "hybrid" facility, one that performs both manufacturing and distribution. Attention is mostly on *location*, and use of the Analytic Hierarchy

Process (AHP) to assess various site-selection factors. But it is clear, right from the start, that this facility has a dual function.

The present section examines more precisely the various roles that a DC might take on in a supply chain. Specifically, we discuss the issues and literature related to the distribution centre that may act as a make-bulk/break-bulk consolidation terminal, a cross-dock operation, a transshipment node, an assembly facility, a product fulfilment centre, or a returned goods depot. Our definitions of these roles are misleadingly clear. In reality, a distribution centre often performs several of these simultaneously, as will be seen below.

3.1 The DC as a make-bulk/break-bulk consolidation centre

Breaking bulk and making bulk are traditional functions of a distribution centre. In a break-bulk facility, large incoming loads are deaggregated, often for product mixing and to create consolidated outbound shipments. A make-bulk facility, or consolidation centre, combines small quantities of several products in fewer, larger assortments.

Higginson and Bookbinder (1994) note that a program of freight consolidation involves determining those products to be dispatched together; which customer orders will be combined; and when consolidated orders will be released. Also, who will perform these activities, which specific consolidation techniques will be used, and will these activities be carried out at a DC or elsewhere? Hall (1987) provides a good introduction to the impact of consolidation performed at a terminal. Gray et al. (1992) discuss the design and operation of an order-consolidation warehouse.

Ketzenberg et al. (2002) examine the benefits of breaking bulk in retail operations. They suggest the major advantage is better use of retail space, rather than reduced inventory. Diks and de Kok (1996) discuss the allocation to multiple retailers, of inventory incoming to a single distribution centre. Klincewicz and Rosenwein (1997) present a heuristic, based on set partitioning, to determine the shipments that should be made from a warehouse or distribution centre each day.

Daganzo (1988) addresses the case of many origins shipping to one destination through a single consolidation centre. He develops an algorithm for use when vehicles can haul multiple items, and presents an example applying the concepts discussed to the transport of automobiles. Daganzo (1987) looks at the role of terminals in a network of several origins shipping to a number of destinations. He notes that, with certain assumptions, the benefits of consolidation can be achieved in one-to-many networks (or many-to-one networks) without the use of terminals, by having vehicles make multiple stops. Daganzo (1999) combines much of his earlier work, and is discussed later in this chapter.

Baudin (2001) lists a few activities that a consolidation centre should *not* perform:

- Kitting. To do so, consolidation-centre employees would require up-to-date information about pick lists, bills-of-material, and engineering -change notices. Kitting should be performed in the factory; the consolidation centre should deal only with individual items.
- Quality assurance of incoming products. That would require consolidation-centre employees to be trained in the characteristics of parts and the customer's quality assurance methods and requirements.
- Sorting of empty crates and other shipping materials. This "creates work that otherwise wouldn't need to be done. It is more economical to organize the pickup of empties by item."

A common example of the use of DCs as consolidation centres for the inbound-to-factory movement can be found in the automobile manufacturing supply chain. Here, the consolidation centre is a facility located close to a production plant, that "receives large shipments of components and parts from many suppliers, breaks them down into the smaller quantities that the plant needs, disposes of the supplier's shipping materials, places the parts in the plant's reusable containers, and delivers them either to plant receiving or directly to the point of use" (Baudin, 2001). Thus, the consolidation centre acts like a supplier to the manufacturer, making frequent deliveries of components and relieving the factory from having to accept large, less regular deliveries of inappropriately packaged items. This requires the consolidation centre to hold substantial inventory, while facility management must have the ability to influence suppliers to improve deliveries and reduce costs.

Baudin notes that consolidation centres in the automobile industry often "are operated by separate companies, in which the manufacturer may or may not own equity. A consolidation centre can recruit warehouse personnel for half or even one third of a car assembler's wages. It cannot do all the material handling for the manufacturing plant, but what it does, it can make a profit while saving money for the plant."

3.2 The distribution centre as cross-dock (CD)

Another function of a DC, i.e., the *cross-docking* of a product through a distribution centre, is recognized as one of the basic distribution strategies (e.g., Chopra, 2003; Chopra and Meindl, 2004). "Cross docking refers to a process where the product is received in a facility, occasionally married with product going to the same destination, then shipped at the earliest opportunity, without going into long-term storage." (Napolitano, 2001) Forty-eight hours is the often-quoted time limit for a cross-docked item to remain in the facility (resulting in an annual inventory turnover greater than 100), but time limits ranging from one to three days appear in various sources. Some sorting and product consolidation also may occur before shipping.

There is a fundamental difference between the use of a CD and traditional warehousing. Customer orders can be filled from goods stored at the warehouse, whereas with cross-docking, customer orders are filled from some other facility (such as a manufacturing plant) and just pass through the distribution centre or CD.

Cross-docking is a form of transshipment, the two differing in terms of objectives. The former strategy is *customer*-focussed, and attempts to move a product through a facility as quickly as possible. Transshipment (discussed next in this chapter) is a *carrier* strategy that aims to improve truck utilization, typically by better matching the size of the load to that of the vehicle. Transshipment is not new (after all, less-than-truckload or LTL transportation is dominated by transshipment operations), while it is only in the last two decades that use of a CD has received widespread attention.

Cross-docking produces many benefits, including:

- Elimination of activities associated with storage of products, such as incoming inspection, putaway, storage, pick-location replenishment, and order picking. Doing away with the latter is especially beneficial: Order picking is the most labour-intensive, time-consuming, costly, and error-prone of all activities in a typical warehouse.
- Faster product flow and improved customer service. Having eliminated storage, products move directly from receiving to shipping (or at worst sit in a staging area for short periods of time).
- Reduced product handling. The results are decreased probability of product damage, less wear on material handling equipment, and diminished labour.
- Cuts in inventory. Cross-docking avoids the holding of stock at multiple locations.
- Lower costs due to elimination of the above-mentioned activities; smaller inventories; less investment in racking, floor storage, or other equipment; and encouragement of consolidation of products for the same destination.

There are several disadvantages to cross-docking. The major one is the very complex planning and coordination needed to make it work effectively. Heaver and Chow (2003) note that because of this difficulty, many retailers have not been able to achieve anything close to true cross-docking. A major impediment, they add, is that most manufacturers are not equipped to efficiently create store-order quantities. As well, because cross-docks do not hold inventory, some managers feel uneasy that customer requirements must be satisfied from more distant facilities, rather than from local warehouses that carry stock (Jones, 2001).

In general, the best potential for effective cross-docking is for those SKUs where a sense of urgency exists. Examples are fast-selling products, time-sensitive components, and sale and promotional items. Special orders and goods that are backlogged also should be cross-docked: These often arrive at the CD pre-packaged and labelled for delivery to the consignee, and do not have to be combined with additional items to complete the customer's order (Frazelle, 2002).

Cross-docking can provide greater control over delivery schedules. Use of a CD is thus well suited to the Just-In-Time manufacturing environment (Luton, 2003), and also to the make-to-order environment (Copacino, 1997). Other conditions under which cross-docking should be considered are given in *Modern Materials Handling* (1995). These include SKUs that arrive at the warehouse already labelled or priced; receipt of large numbers of individual items; products whose destination is known when received; and goods for customers who are prepared to receive them immediately.

The major prerequisite for successful cross-docking is a system to ensure the efficient exchange of products between supply chain entities. Emphasis should be given to the scheduling and coordination of shipments inbound and outbound at a given node (Bookbinder and Barkhouse, 1993; Jones, 2001). This requires a timely and accurate flow of information between supply chain members. Such an information system should support advanced shipment notifications (ASN), electronic data interchange (EDI), and automatic identification (auto ID) technologies, such as bar codes and radio-frequency tags.

Frazelle (2002) notes that advance knowledge of inbound goods and their destinations allows the CD "to route the product to the proper outbound vehicle, to schedule inbound loads to match outbound requirements on a daily or even hourly basis, and to better balance the use of receiving resources (dock doors, personnel, staging space, and material handling equipment) and, if necessary, shift time-consuming receipts to off-peak hours."

Other requirements of cross-docking include (e.g., Napolitano, 2001):

- Suppliers who can consistently provide the correct quantity of the right product, at the precise time when needed.
- Capital to sustain a cost-justified CD system and personnel who recognize the importance of moving, not storing, products.

- Adequate space for staging, and appropriate docks and material handling equipment (Jones, 2001).
- Inbound shipments consisting of pallets or cases that contain a single SKU or a set of SKUs going to the same destination, so as to minimize sorting (Frazelle, 2002).

As well as information requirements, the physical design of the CD (Bartholdi and Gue , 2001) must be considered. The ideal cross-dock should be rectangular, long and narrow, with loading docks on each side to smooth the product flow and inhibit product storage (Murphy and Wood, 2004). Although the facility should be as small as possible to minimize travel distances between vehicles (Luton, 2003), the cross-dock staging area must be large enough to allow the direct flow of products between receiving and shipping (Jones (2001)). There also must be a sufficient number of doors to avoid backlogs and delays for carriers. Luton (2003) notes that a conventional warehouse can encourage direct-flow operations by having both shipping and receiving docks on the same face of the building.

3.3 The DC as a transshipment facility

Along with breaking bulk and making bulk, a traditional function of a distribution centre is *transshipment.* This refers to the process of taking an item or shipment out of one vehicle and loading it onto another (Daganzo, 1999). Transshipment may or may not include consolidation or de-consolidation. If no items are added or removed during the transshipment, the process is sometimes referred to as *transloading.* Beuthe and Kreutzberger (2001) provide a detailed discussion of transshipment in logistics networks of various designs.

Transshipment occurs when there is good reason to change transportation modes or vehicle type. Transshipment centres "decouple the linehaul transportation and local delivery operations, enabling us to use larger trucks for linehaul than for delivery; they also increase the number of delivery stops that can be made without violating route length limitations." (Daganzo, 1999). Transshipment can be used as well during the final delivery stage to handle time-of-day constraints at customers, or weight restrictions on truck-delivery routes. Vehicles operating out of a transshipment centre are dedicated to specific links of the supply chain; they can thus be optimally sized and configured for the services and routes they handle. Conversely, transshipment does imply greater cost: Less-direct truck routes are employed, transshipment facilities are required, and terminal operations increase transit time and potential for damage.

Transshipment is the main focus of a hub-and-spoke transportation system, aspects of which have been examined by a number of researchers. Examples include Taylor et al. (1995); Pirkul and Schilling (1998); Bryan and O'Kelly (1999); Cheung and Muralidharan (1999); and Campbell et al. (2002).

Pleschberger and Hitomi (1994) and others have noted the negative impacts (noise, air pollution, ...) of frequent JIT deliveries. In Europe, transshipment centres have thus been suggested as a way to reduce environmental problems created by truck traffic in urban areas. Whiteing et al. (2003) observe, however, that such centres have had problems related to insufficient product volumes, relatively high operating costs, and feelings of loss of control by shippers of the goods. Since the major drawback of transshipment facilities is inadequate throughput, proposals for transshipment centres often *require* carriers to consolidate products for delivery or collection in city centres (or include penalties for not doing so). Many carriers, however, have requested exemption from consolidation, claiming that their products are highly perishable, may contaminate other goods, or need intense levels of security (Whiteing et al., 2003).

Similar environmental concerns were part of the discussion by Taniguchi et al. (1999) of the experience in Japan with a *public logistics terminal.* This is a multi-company DC; it may be viewed as the supply-chain generalization of a public warehouse. Those authors employ queuing theory and nonlinear programming in a model to determine the optimal sizes and locations of public logistics terminals. Traffic congestion and energy-environmental issues were accounted for in an application in the Kyoto – Osaka area.

Bendel (1996) describes transshipment centres as key to a concept called *city logistics.* During the 1990s, carriers in several German cities agreed to divide loads (and revenue) so as to improve efficiency and avoid duplication of travel. These schemes sometimes included, with financial assistance of local government, the establishment of a transshipment centre to handle collections and deliveries for the urban area concerned. Kohler and Straub (1997) discuss a city logistics program in Kassel, whereby five German carriers transship freight to a sixth. The latter delivers to retailers in the city centre. This arrangement improved the vehicle load factors by more than 50 percent. It was found, however, that environmental benefits were partly offset by increases in total operating costs. Short case studies of other city logistics schemes are given in Thompson and Taniguchi (2001). Wider issues, i.e., advanced methods to manage urban freight transport, are considered by Taniguchi et al. (2001) and Crainic et al. (2004).

Lee's (1996) facility location/allocation model recognizes that an organization will use distribution centres with differing capacities. This integer linear programming model implicitly treats all DCs as transshipment centres. Bhaskaran (1992) presents a case study from the automobile manufacturing industry. She develops a heuristic to determine the number and location of transshipment centres; here those centres are *CDs*, with no possibility of storage. The paper discusses a good sequential strategy for adding new transshipment centres, one at a time, as demand grows in the network.

Daganzo (1999) provides a comprehensive mathematical examination of different logistics systems, both with and without transshipment centres. (We remark that his work considers the breaking of bulk as included in a transshipment.) He begins by studying loads moving from an origin to a single destination, through one transshipment centre. Daganzo notes that this problem is similar to the classical model for facility location and sizing, with an additional decision related to vehicle scheduling. Thus, the critical step in design of such a system is to determine ideal locations for transshipment facilities. In this case, when pipeline inventory cost is negligible relative to other logistics costs, he concludes that trucks should be filled to capacity. Hence the largest ones possible should be used, which may require transshipments if truck sizes are restricted in a market area.

Daganzo's examination of many-to-many distribution treats facilities as makebulk/breakbulk consolidation centres. These are multicommodity problems where each origin supplies a unique product. When there are no restrictions on vehicle capacity or route length, logistics costs per delivered item improve as more routes transship at the terminal. Logistics systems with one terminal; multiple terminals having a single transshipment per load; and multiple terminals with more than one transshipment per load are discussed. Daganzo shows how, for multiple-terminal systems, determination of truck routes depends on whether the area around a given terminal ships to, or receives from, only that terminal.

A number of researchers have studied the use of transshipments in the management of inventory and its re-allocation. Such models typically employ the term, "transshipment," differently than the transportation-sense adopted in this section. Instead, *transshipment* is defined as a tactic in multi-location inventory control, whereby products can be transferred laterally between stocking-points, as demand requires. (Bertrand and Bookbinder, 1998, term this a *redistribution.*) Thus, contrary to our definition of DC, it is assumed that facilities do hold inventory. Publications in this area include Evers (1996, 2001); Herer et al. (2002); Hong-Minh et al. (2000); and Tagaras and Vlachos (2002).

3.4 The distribution centre as an assembly facility

Having discussed the inventory-transportation interfaces of a DC, let us now consider linkages closer to manufacturing. It is well known that delaying item-differentiation, packaging, and labelling until later stages of the supply chain can improve product allocation. The often-cited case of Hewlett Packard's European distribution centre is a good example of using a DC for minor product assembly (see, for example, Kopczak and Lee, 1994; Simchi-Levi et al., 2003). Prior to moving assembly activities to that facility, HP's DeskJet printer was manufactured in Vancouver, Washington, and shipped by water to the European DC. The latter facility suffered from inaccurate forecasts, serious inventory problems, and poor customer service. HP redesigned the DeskJet so that a single generic model (allowing easy customization) could be produced in Vancouver, then assembled-to-order in one of six ways at the European distribution centre. Simchi-Levi et al. (2003) offer a mathematical illustration of the resulting savings in inventory cost. This hinges on the decreased standard deviation of demand, hence lower safety stock overall, due to generic redesign.

Just as important are the human issues related to HP's decision. Assembly responsibilities were initially resisted by DC employees, who saw their role to be in *distribution*, not manufacturing. As well, the DCs were reluctant to give up some inventory, in light of expectations of high customer service.

A major advantage of using a distribution centre for final assembly activities is "product localization"; that is, the ability to configure an item in a given market area to better reflect the needs and characteristics of that market. Switching to a strategy of performing final assembly at a DC will also change the relative value of an SKU at different stages in the supply chain. Some financial benefits may result. For example, labour often costs less at the distribution centres than in factories. If goods must cross international borders before reaching the DC for final assembly, tariff duties may be lower on the unfinished product than on a finished item.

3.5 The DC as product-fulfilment centre

Let us now consider facilities with stronger links to the end-customer. The term *fulfilment centre* has been used to describe a DC or warehouse whose major function is to respond to product orders from the final consumer, by shipping those items directly there. Usually, customers will have placed those orders via an electronic medium such as the World Wide Web.

Product fulfilment centres differ from traditional warehouses and DCs in a number of ways (Ackerman and Brewer, 2001):

- Because the fulfilment-centre operator deals directly with consumers, customer-service requirements demand greater importance.
- The size of a typical order handled by a product fulfilment centre is smaller, but the number of orders is larger.
- Most or all orders are received electronically (as noted already).
- Fulfilment centres typically must receive customer payments, often by major credit card; some also create customer invoices and handle banking for their clients.
- A large amount of time is spent in dealing with *returns* from customers.
- Computerized information systems and task automation are increasingly critical, and the transportation function (especially residential delivery) is more complex.

Because the role of product-fulfilment centre interacts with several others that the DC may play, there is considerable potential here for further research.

3.6 The distribution centre as depot for returned goods

Although reverse distribution is analysed in greater detail elsewhere in this book, it is useful to briefly examine the role of DCs in the handling of returned items.

Many of the distribution-centre functions discussed previously in this chapter (including consolidation and light assembly) come together in dealing with product returns. The reverse distribution channel typically is more complex than the forward flow. The main objective in many reverse distribution systems is to minimize costs, while quickly getting the returned product back into the forward distribution channel. At the same time, a major management concern in reverse distribution is to avoid the inadvertent mixing of SKUs in the return channel with those in the forward direction. As a result, firms such as Sears Roebuck, Hudson's Bay, Target and K-Mart have outsourced their reverse distribution channel to third parties who operate DCs dedicated to materials returned.

The handling of such items is very labour intensive. All returned products must be inspected, then separated into those that can be repaired or repackaged at the returned-goods depot; others which need go back to the supplier; those that will be sent elsewhere (e.g., donated to charity or sold in a secondary market); and some that must be destroyed or recycled for scrap. Conversely, an organization with a private fleet, and which chooses to manage its own reverse distribution channel, can improve vehicle and driver utilization if returned items are transported on inbound trips back from other facilities.

3.7 The DC in miscellaneous other roles

A distribution centre often performs more than one function simultaneously. We have mentioned transshipment and consolidation (Whiteing et al., 2003); break bulk and light assembly (Kopczak and Lee, 1994); and returned-goods processing at outbound consolidation facilities. In conjunction with material flows, a DC may also act as a depot for trucks or drivers, where the fleet is domiciled or maintained, or where drivers switch vehicles to avoid violating personnel schedules or legal or workforce constraints. Ross and Droge (2002) present an example of this role in the petroleum industry.

Coordination of inbound and outbound vehicles for product distribution has been discussed by several authors (e.g., Daganzo, 1999); restrictions on tour length due to driver issues are common in vehicle routing formulations. Nevertheless, research typically treats the questions of where vehicles or drivers rest as secondary to product decisions.

A distribution centre also can offer customer support. Designing, providing and scheduling services such as installation and repair require operational decisions quite different from those faced by DCs dealing only in goods movement. Similarly, particular SKUs (e.g., repair parts or hazardous items) should be held centrally or in specialized locations. Some distribution centres will thus be assigned these functions.

Lastly, a DC can offer space for retail sales to final customers, i.e., can act as a factory-outlet store. As well as providing a way to dispose of excess, discontinued, returned or slightly soiled items, manufacturers and distributors can retain control over their products while earning the higher revenues associated with retailing (e.g., Berman, 1996).

4. Measuring distribution-centre performance

Section 3 described how a DC can play multiple roles, singly or in combination. We now turn attention to *evaluation* of those activities.

The measurement of performance of an organization's logistics function or its supply chain is addressed in many works (see, for example, Ross et al., 1999; Keebler, 2001; Ballou, 2004). However, performance assessment is not usually discussed explicitly for a *DC*. Fortunately, a number of the measures used in evaluating performance of traditional warehouses are applicable to distribution centres. Metrics for a DC would thus typically benchmark current actual performance against results achieved in the past, output of comparable operations elsewhere in the company, or achievements by other organizations or best performers and industry standards. This comparison is straightforward. A DC carries out a large number of activities, highly repetitive and easily monitored; that encourages quantitative measures.

In fact, a few methods for evaluating performance of the distribution centre as part of a supply chain have been developed. In addition to benchmarking (above), one has available the analysis of cycle time and integrative-evaluation approaches, such as "balanced scorecard" models and *SCOR*, the Supply Chain Operations Reference model.

As in a warehouse, the per-unit and total costs remain critical measures of DC performance (Higginson, 1993). Daganzo (1999) covers the mathematical modelling of distribution centre costs, discussing the charges for inventory holding, transportation and material handling. Another important indication of the viability of a DC is *throughput*; that is, the total amount (weight, dollar-value, etc.) of goods that pass through the facility during a stated period of time. Inventory turnover and the similar shipments-to-inventory ratio also are employed. Performance measures commonly used in distribution centres include total cost per case, or per pallet, or per employee hour; labour utilization percent; fixed cost per square metre; and the time between receipt and dispatch of an order. Additional metrics for DC productivity are listed in Schary (1984, p. 102). Again, some of these measures assume that the facility carries stock.

Frazelle (2003) states, "The most critical quality indicators for distribution centre operations are inventory accuracy (percentage of inventory storage locations without discrepancies), picking accuracy (percentage of lines picked without errors), shipping accuracy (percentage of lines shipped without errors), and warehouse damage percentage (percentage of dollar-value of damages per dollar-value of items shipped)." Clearly most of these standards relate not to product movement, but rather to SKU storage and picking. Those warehouse-type functions ignore the *time-based* element in a DC.

Yang (2000) identifies, through computer simulation, the major policies and environmental factors that affect the performance of a single-warehouse multiple-retailer distribution system. He remarks that the operating environment (e.g., few vs. many stores; low or high variability of demand) often has a greater effect on performance than does choosing the appropriate system or policy (vehicle routing algorithm; periodic vs. continuous inventory review). Ackerman and Brewer (2001) note that one of the most important measures of distribution centre performance is the perception of customers who work with, or receive deliveries from, the DC. They add that, "In a distribution centre where customer service has top priority, the warehouse management system is judged by its capacity to provide service that is superior to its competition."

Kuo et al. (1999) examined performance measurement in six categories (finance, operations, quality, safety, personnel, and customer satisfaction) for five DCs (technically, warehouses). A cross-case comparison showed that the facilities used fairly similar objective measures for the first four categories, including cost per unit, percentage of errors, and number of employee accidents. However, for all five DCs, evaluation of service to clients was limited to customer feedback.

Less traditional methods for evaluating distribution centre performance have been suggested. Noh and Jeon (1999) employ several methodologies, including AHP and data envelopment analysis (DEA), to compare relative efficiencies for the DCs of a Korean telecommunications company. Ross and Droge (2002) present a benchmarking model, also using DEA. To evaluate a set of 100 DCs in the petroleum business, Ross and Droge optimise an objective related to the aggregate efficiency ratio. Their DEA model has three inputs: Fleet size; labour (average no. years experience of personnel assigned to DC); and mean order-throughput time. Outputs are (transformed) sales volumes of each of four products. The resulting efficient frontier gives the top-performing DCs in any time period.

The model of Ross and Droge (2002) appears useful in evaluating distribution centres whose role (among others) is that of vehicle depot. It could be argued, however, that order-throughput time is an *output*, not an input, and that the marketing mix (beyond the DC's control) has a major effect on sales volume. But Ross and Droge do point a way to evaluate the distribution centres of a given supply chain or of competing chains.

5. Information requirements to manage a DC

An additional input that most distribution centres take for granted is *information*, available in the proper format. A DC must be efficient in the retrieval and transfer of data because of today's greater size of facilities, faster product flow, and increased importance of coordinating inbound and outbound shipments. This section focuses on two common computerized information systems - Enterprise Resource Planning

(ERP) and Warehouse Management Systems (WMS) - employed at distribution centres. A non-technical overview of the evolution of logistics information software is given in Ayers (2001).

An Enterprise Resource Planning system is a computer package that integrates the data of the entire organisation into a single relational or object-oriented database linked to various transaction-processing modules. Such modules typically include applications in distribution and sales, finance and accounting, human resources, inventory control and manufacturing, and purchasing. The functions of warehousing and distribution centre management are typically accessed through one of these modules.

Factors contributing to successful and not-so-successful ERP implementations have been well documented in the literature (e.g., Stratman and Roth, 1999; Nah and Lau, 2001; Umble et al., 2003). It has been recognized by several researchers that many ERP systems unfortunately lack the functionality required to adequately support warehouse/distribution centre planning, and other supply chain processes including transportation. (See, for example, Frazelle, 2002; Handfield and Nichols, 2002; Spiegel, 2003).

ERP systems are designed to integrate, via a "suite" of applications, all of the organization's functions, including warehousing. However, as Frazelle (2002) notes, "Many warehousing systems evolved from applications very far removed from warehousing, including accounting, customer service, general ledger, inventory management, and/or manufacturing. Unfortunately, warehousing is typically an afterthought application for these providers, and the full-suite providers typically have very little expertise in warehousing." As well, ERP systems are "transaction-based;" that is, they are intended to record what the organization *has* done, rather than plan what the organization should do.

This has led to the development of Advanced Planning and Scheduling software (APS), which provide the OR capabilities in optimisation lacking in ERP and MRP packages (e.g., Green, 2001). "Bolt-on systems" is the descriptor given to APS: They extend the functionality of other software by drawing their input data from those packages, including ERP and logistics execution systems such as forecasting, production control, transportation, warehousing and order management (Cauthen, 1999).

Aksoy and Derbez (2003) categorize available software according to OR techniques used and the supply-chain application. Many of these packages have been quite successfully utilised. We remark, however, that some APS designed for logistics planning is purely-executional software which lacks a capability for long term planning. This is perhaps one reason why it was recently reported (Foster, 2003) that many users preferred to purchase single programs to address specific logistics problems, rather than *suites* of supply-chain applications.

Let us turn now to the Warehouse Management System. This denotes computer software that tracks, plans, controls, analyses, and records the flow of product through a warehouse or distribution centre. A WMS (like a Transportation Management System) falls into the category of "logistics execution software." Thus, unlike ERP systems, Warehouse Management Systems are intended as *real-time planning tools.*

Particular functional capabilities are common in a Warehouse Management System (for example, see Frazelle, 2003). Such software:

- automates transaction activities such as verification of product weight and cube, and vendor compliance
- determines product storage locations within a facility
- develops and prints order pick-lists
- prints labels for bar code, storage location, product IDs, etc.
- plans inbound and outbound transportation activities, including container optimization, load planning, and dock and yard management
- performs various activities related to workforce management, such as workload planning and scheduling, labour control, and time standards
- supports electronic communication within the facility (such as via radio frequency) and with supply chain partners (e.g., through EDI and ASN)
- compiles and reports activity information, e.g., detailed summaries of each inbound or outbound movement, item activity profiles, and facility performance measures.

The above is the good news. However, Warehouse Management Systems have several fundamental problems. Frazelle (2002) observes, "... Most WMS vendors have few highly qualified engineers and analysts. Those few are typically assigned to the largest and most prestigious accounts. If you are not included in that list, you may not be satisfied with the capabilities of the engineers and analysts assigned to your project." He goes on to note that, "Less than half of all warehouse management systems yield the performance and practice improvements promised during the justification phase."

Moreover, integrating a Warehouse Management System with an ERP system is considered to be quite difficult, really time-consuming, and very expensive (Cooke, 1998). Thus, organizations wishing to have warehouse/distribution centre management functionality as part of their ERP system will have to be involved in a major integration project, or purchase an ERP system that, although possessing such capabilities, probably has less than is desired. It remains to be seen if recent attempts by some major ERP vendors to ally themselves with providers of WMS will be successful. At the same time, there has been a similar move toward integrating WMS and Transportation Management Systems; Mason et al. (2003) discuss the benefits of doing so.

Lastly, ERP and WMS focus on logistics activities for one organization or for one facility of that organization. Although many ERP systems allow electronic communication with suppliers and customers, neither WMS nor ERP is well suited for linking supply chain members, and even less so for planning and coordinating movements between facilities throughout the chain. Distribution Resource Planning (DRP) systems may provide assistance in these tasks. (See, e.g. Vollmann et al., 2004). The initial DRP systems of the 1980s promised smaller inventories, higher in-stock availability, and reduced transportation and operating costs. Those systems did not often achieve their potential, partly because DRP is most beneficial for multi-echelon distribution networks. These have become less popular as companies reduced the number of stock-keeping locations. But note that the original concept of DRP pertained to a single organization.

Channel-wide DRP systems attempt today to link all facilities across the supply chain, something not found traditionally in logistics software. In the channel-wide case, "each customer distribution centre is established as a stocking location in the manufacturer's DRP system. The manufacturer's DRP system manages replenishment from plants to both its own distribution centres and the customer's distribution centres as if the manufacturer owned the entire network. Given that supply chains from manufacturers to their customers are multi-echelon systems, a channel-wide DRP replenishment system invariably produces superior results" (Copacino, 1997). This is clearly a challenging research area, since it involves simultaneous scheduling for multiple decision-makers whose databases have varying degrees of integration.

6. Summary and conclusions

This chapter has described the functions that a distribution centre can assume in the operation of a supply chain, as well as discussed some considerations in monitoring and controlling the activities of a DC. Our final section comments on the weaknesses of published articles related to those roles. We suggest some areas in which future research could be carried out, in addition to topics proposed above.

As stated previously, most academic literature does not distinguish between warehouses and distribution centres, therefore ignoring any distinctions in activity. The paper by Lee (1996) is a good example.

Whereas the title does indicate that various types of facilities are being modelled, this difference is only in terms of capacity and fixed costs, not roles. (Of course, it could be claimed that contrasting functions *imply* the differing capacities.) Ignoring the diversity of activities at a DC can result in assumptions that may not be accurate, or are not explicitly stated to the reader. A common example is assuming that the distribution centre will store inventory. Conversely, recognizing the variety of services the facility may offer provides researchers with potential areas for study.

In fairness, some publications have emphasized that facilities may function as distribution centres, not as warehouses. In the classical transshipment problem, for example, the assumption that all items that move into the facility must also leave implies that inventory is not being held, hence the facility is acting as a DC or cross-dock. However, as observed in Section 3.3, some of the more recent papers employ the term "transshipment" to mean transfers between stock-keeping locations.

Among the roles of a distribution centre discussed in this chapter, the OR literature has given greatest attention to transshipment. This is due in part to its close relationship to vehicle routing or decisions on the number and mix of vehicles in the fleet. The key issue that should be captured in a mathematical model that includes transshipment or cross-docking through a DC is the synchronization of trucks inbound and outbound. This is mentioned by Daganzo (1999) and others.

An obvious and interesting question then is, "When should a facility be used for storing inventory and when should it be limited to the flow-through roles described in the previous sections?" The location model of Gümüş and Bookbinder (2004) aims to decide, from a set of potential sites, where *CDs* will be opened. Only cross-docks are considered; consolidated-shipment opportunities are thus important here. A location model could, more generally, consider two types of intermediate facilities: One would act as a stock-keeping warehouse, the other as a distribution centre.

Similarly, little research has been done on issues in using a DC for light assembly. A model of product flows in this situation would have to include considerations of *time*. Although a shorter interval might be required for production at the factory, the period between arrival at the DC and customer-delivery will increase. If there is not a greater frequency of shipment from factory to distribution centre, total lead-time will grow. Even with that enhanced frequency, the customer's wait will be of longer duration.

Other questions exist: What are the characteristics of a supply chain for which light assembly is preferably done at an intermediate DC? What is the best layout of such a facility? Mathematical modelling has potential application in the design and configuration of distribution centres for use in product assembly, e-commerce fulfilment and returned-goods collection. These analyses would correspond to those done on the layout of warehouses (e.g., Gray et al., 1992) or cross-docks (Bartholdi and Gue, 2001).

The role of a distribution centre as a depot for returned items has been touched upon by vehicle routing studies that consider both deliveries and backhauls. In practice, the major qualifying factor is the volume of product to be brought back on a route; this is rarely as great as the quantity moving outward. Fernie (2003) mentions the case of a Scottish DC, designed to act as collection point for reusable items picked up from retail stores. The volume of such materials, however, insufficiently utilized the trucks returning to the facility.

When a third party handles returned goods, items moving in the reverse direction typically do not flow through the seller's forward distribution system at all. Instead, they go directly from the point of customer return to the third party's facility. That can simplify any modelling or analysis: Forward and backward product flows, now independent, can be optimised separately.

Distribution-centre performance measurement and information systems have their roots in warehouse management. Some approaches may therefore be sub-optimal for application in DCs. An examination is warranted to determine the true utility of Warehouse Management Systems, or warehouse-based performance measures, in managing the operations of a distribution centre. For example, it has been noted that many early WMS did not adequately handle cross-docking; improved functionality in this area is appearing only now.

Clearly, a DC in a supply chain can assume roles that go well beyond the traditional functions of transshipment and breaking bulk. This recognition provides a number of areas for potential study. That research will encourage better understanding and utilization of these facilities.

Acknowledgements The research leading to the preparation of this chapter was partially funded by NSERC, the Natural Sciences and Engineering Research Council of Canada. The authors are grateful for this support.

References

Ackerman, K.B. and Brewer, A.M. (2001). Warehousing: A key link in the supply chain. In: A.M. Brewer, K.J. Button, and D.A. Hensher (eds.), *Handbook of Logistics and Supply-Chain Management*, pp. 225–237. Pergamon, New York.

Aksoy, Y., and Derbez, A. (2003). 2003 software survey: supply chain management. *OR/MS Today*, 30(3):34 – 41.

Ayers, J.B. (2001). Topography of supply chain applications. In: J.B. Ayers (ed.), *Handbook of Supply Chain Management*, pp. 167 – 178. St. Lucie Press, Boca Raton, FL.

Ballou, R.H. (2004). *Business Logistics/Supply Chain Management*. Pearson, Upper Saddle River, NJ. 5th edition.

Bancroft, T. (1991). Strategic role of the distribution centre: How to turn your warehouse into a DC. *International Journal of Physical Distribution and Logistics Management*, 21(4):45 – 47.

Bartholdi, J.J. and Gue, K.R. (2001). The best shape for a cross dock. Under review.

Baudin, M. (2001). Consolidation centers in the lean supply chain. In: J.B. Ayers (ed), *Handbook of Supply Chain Management*, pp. 375 – 383. St. Lucie Press, Boca Raton, FL.

Bendel, H.J. (1996). City logistics. *Logistics Europe*, pp. 16 – 23, February.

Berman, B. (1996). *Marketing Channels*. John Wiley & Sons, New York

Bertrand, L.P. and Bookbinder, J.H. (1998). Stock redistribution in two-echelon logistics systems. *Journal of the Operational Research Society*, 49:966 – 975.

Beuthe, M. and Kreutzberger, E. (2001). Consolidation and trans-shipment. In: A.M. Brewer, K.J. Button, and D.A. Hensher (eds.), *Handbook of Logistics and Supply-Chain Management*, pp. 239 – 252. Pergamon, New York.

Bhaskaran, S. (1992). Identification of transshipment center locations. *European Journal of Operational Research*, 63:141 – 150.

Bookbinder, J.H. and Barkhouse, C.I. (1993). An information system for simultaneous consolidation of inbound and outbound shipments. *Transportation Journal*, 32(4):5 – 20.

Bordley, B., Beltramo, M., and Blumenfeld, D. (1999). Consolidating distribution centres can reduce lost sales. *International Journal of Production Economics*, 58(1):57 – 61.

Bryan, D.L. and O'Kelly, M.E. (1999). Hub-and-spoke networks in air transportation: An analytical review. *Journal of Regional Science*, 39:275 – 295.

Campbell, J.F., Ernst, A.T., and Krishnamoorthy, M. (2002). Hub location problems. In: Z. Drezner and H.W. Hamacher (eds.), *Facility Location: Applications and Theory*. Springer-Verlag, Heidelberg.

Caron, F., and Marchet, G. (1996). The impact of inventory centralization/decentralization on safety stock for two-echelon systems. *Journal of Business Logistics*, 17(1):233 – 257.

Cauthen, R. (1999). APS technology: Powering supply chain management. *Enterprise Systems Journal*, 14(9):41 – 45.

Cheung, R.K. and Muralidharan, B. (1999). Impact of dynamic decision making on hub-and-spoke freight transportation networks. *Annals of Operations Research*, 87:49 – 71.

Chopra, S. (2003). Designing the distribution network in a supply chain. *Transportation Research E*, 39(2):123 – 140.

Chopra, S. and Meindl, P. (2004). *Supply Chain Management: Strategy, Planning, and Operation*. 2nd edition, Upper Saddle River, NJ: Pearson.

Cooke, J.A. (1998). Crossing the great software divide. *Logistics Management*, pp. 72 – 74, June.

Copacino, W.C. (1997). *Supply Chain Management: The Basics and Beyond*. St. Lucie Press, Boca Raton.

Cormier, G. (2005). Operational research methods for efficient warehousing. In: A. Langevin and D. Riopel (eds.), *Logistics Systems: Design and Optimization*, Kluwer, Norwell, MA.

Coyle, J.J., Bardi, E.J., and Langley, C.J., Jr. (2003). *The Management of Business Logistics: A Supply Chain Perspective*, 7th edition. South-Western, Mason, OH.

Crainic, T.G., Ricciardi, N., and Storchi, G. (2004). Advanced freight transportation systems for congested urban areas. Forthcoming in *Transportation Research C.*

Daganzo, C.F. (1987). The break-bulk role of terminals in many-to-many logistic networks. *Operations Research*, 35:543 – 555.

Daganzo, C.F. (1988). Shipment composition enhancement at a consolidation center. *Transportation Research B*, 22:103 – 124.

Daganzo, C.F. (1999). *Logistics Systems Analysis*, 3rd edition. Springer-Verlag, Heidelberg.

Daww, R.L. (1995). Reengineer warehousing. *Transportation and Distribution*, 36(1):98 – 102.

Delaney, R.V. (1991). Trends in logistics and U.S. world competitiveness. *Transportation Quarterly*, 45(1):19 – 41.

Diks, E.B. and de Kok, A.G. (1996). Controlling a divergent 2-echelon network with transshipments using the consistent appropriate share rationing policy. *International Journal of Production Economics*, 45(1/3):369 – 379.

Evers, P.T. (1996). The impact of transshipments on safety stock requirements. *Journal of Business Logistics*, 17(1):109 – 133.

Evers, P.T. (2001). Heuristics for assessing emergency transshipments. *European Journal of Operational Research*, 129:311 – 316.

Fernie, J. (2003). Retail logistics. In: D. Waters (ed.), *Global Logistics and Distribution Planning*. pp. 249 – 275, 4th edition. Kogan Page, London.

Footlik, R.B. (1999). Property tax trends: Do 30' high distribution centres make sense? *Journal of Property Tax Management*, 10(3):63 – 71.

Foster, T.A. (2003). Supply chain top 100 software vendors. *Logistics Management*, 42(9):S9.

Frazelle, E.H. (2002). *World-Class Warehousing and Materials Handling.* McGraw-Hill, New York.

Frazelle, E.H. (2003). *Supply Chain Strategy.* McGraw-Hill, New York.

Gray, A.E., Karmarker, U.S., and Seidmann, A. (1992). Design and operation of an order-consolidation warehouse: Models and application. *European Journal of Operational Research*, 58:14 – 36.

Green, F.B. (2001). Managing the unmanageable: Integrating the supply chain with new developments in software. *Supply Chain Management*, 6(5):208 – 211.

Gümüş, M. and Bookbinder, J.H. (2004). Cross-docking and its implications in location-distribution systems. Forthcoming in *Journal of Business Logistics.*

Hall, R.W. (1987). Consolidation strategy: Inventory, vehicles and terminals. *Journal of Business Logistics*, 8(2):57 – 73.

Handfield, R.B. and Nichols, E.L., Jr. (2002). *Supply Chain Redesign.* Financial Times Prentice Hall, Upper Saddle River, NJ.

Heaver, T. and Chow, G. (2003). Logistics strategies for North America. In: D. Waters (ed.), *Global Logistics and Distribution Planning*, pp. 413 – 427, 4th edition.Kogan Page, London.

Herer, Y.T., Tzur, M., and Yucesan, E. (2002). Transshipments: Emerging inventory recourse to achieve supply chain legality. *International Journal of Production Economics*, 80:201 – 212.

Higginson, J.K. (1993). Modelling shipper costs in physical distribution analysis. *Transportation Research A*, 27:113 – 124.

Higginson, J.K. and Bookbinder, J.H. (1994). Policy recommendations for a shipment consolidation program. *Journal of Business Logistics*, 15(1):87 – 112.

Hong-Minh, S.M., Disney, S.M., and Naim, M.M. (2000). The dynamics of emergency transshipment supply chains. *International Journal of Physical Distribution and Logistics Management*, 30(9):788 - 816.

Jones, A. (2001). Cross docking — is it right for you? *Canadian Transportation & Logistics*, 104(9).

Keebler, J.S. (2001). Measuring performance in the supply chain. In: J.T. Mentzer (ed.), *Supply Chain Management*, pp. 411 – 435. Sage Publications, Thousand Oaks, CA.

Ketzenberg, M., Metters, R., and Vargas, V. (2002). Quantifying the benefits of breaking bulk in retail operations. *International Journal of Production Economics*, 80(3):249 – 263.

Kim, J.-S. (2002). On the benefits of inventory-pooling in production-inventory systems. *Manufacturing and Service Operations Management*, 4(1):112 – 116.

Klincewicz, J.G. and Rosenwein, M.B. (1997). Planning and consolidating shipments from a warehouse. *Journal of the Operational Research Society*, 48(3):241 – 246.

Kohler, U. and Straub, S. (1997). City logistics concept for Kassel. *Proceedings of 25th PTRC European Transport Forum: Seminar B – Freight*, pp. 97 – 103. PTRC Education and Research Services, London.

Kopczak, L. and Lee, H. (1994). *Hewlett-Packard: DeskJet Printer Supply Chain (A)*. Stanford University, Department of Industrial Engineering and Engineering Management.

Kuo, C.-H., Dunn, K.D., and Randhawa, S.U. (1999). A case study assessment of performance measurement in distribution centers. *Industrial Management + Data Systems*, 99(2):54 – 63.

Lee, C.Y. (1996). An algorithm for a two-staged distribution system with various types of distribution centers. *INFOR*, 34(2):105 – 117.

Luton, D. (2003). Keep it moving: A cross-docking primer. *Materials Management & Distribution*, 48(5):29.

Mason, S.J., Ribera, P.M., Faris, J.A., and Kirk, R.G. (2003). Integrating the warehousing and transportation functions of the supply chain. *Transportation Research E*, 39(2):141 – 159.

Min, H. and Melachrinoudis, E. (1999). The relocation of a hybrid manufacturing/distribution facility from supply chain perspectives: a case study. *Omega*, 27(1):75 – 85.

Modern Materials Handling. (1995). Receiving is where efficiency starts. *Modern Materials Handling*, 50(5):9.

Murphy, P.R., Jr. and Wood, D.F. (2004). *Contemporary Logistics*, 8th edition., Pearson, Upper Saddle River, NJ.

Nah, F.F., and Lau, J.L. (2001). Critical factors for successful implementation of enterprise systems. *Business Process Management Journal*, 7(3):285 – 296.

Napolitano, M. (2001). *Making the Move to Cross-Docking* pp. 308 – 320. Warehousing Education and Research Council, Oxford, OH.

Noh, S.-J. and Jeon, S.-H. (1999). A relative efficiency assessment model for logistics systems. *Journal of the Korean Operational Research Society*, 24(4):95 – 109.

Pirkul, H. and Schilling, D.A. (1998). An efficient procedure for designing single allocation hub and spoke systems. *Management Science*, 44:S235 – S232.

Planeta, J. (2001). Real estate logistics: understanding the modern Canadian distribution centre. *Canadian Transportation & Logistics*, 104(9):14.

Pleschberger, T.E. and Hitomi, K. (1994). Just-in-time shipments in a truck-traffic-coordination system. *International Journal of Production Economics*, 33:195 – 205.

Quinn, F.J. (2000). Transportation: The forgotten factor. *Logistics Management*, p. 45, September.

Reynolds, T. (2003). Distribution Center Management. *Annual Trends Report.*

Ross, A., and Droge, C. (2002). An integrated benchmarking approach to distribution center performance using DEA modelling. *Journal of Operations Management*, 20(1):19 – 32.

Ross, A., Venkataramanan, M.A., and Ernstberger, K.W. (1999). Reconfiguring the supply network using current performance data. *Decision Sciences*, 29:707 – 728.

Schary, P.B. (1984). *Logistics Decisions: Text and Cases.* The Dryden Press, Chicago.

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E. (2003). *Designing and Managing the Supply Chain.* 2nd edition. McGraw-Hill Irwin, Boston.

Spiegel, R. (2003). ERP vendors muscle into logistics. *Logistics Management*, 6(4):45 – 48.

Stratman, J.K. and Roth, A.V. (1999). Beyond ERP implementation: Critical success factors for North American manufacturing firms. *Supply Chain and Logistics Journal*, 5(1):5.

Tagaras,G., and Vlachos, D. (2002). Effectiveness of stock transshipment under various demand distributions and nonnegligible transshipment times. *Production and Operations Management*, 11:183 – 198.

Taniguchi, E., Noritake, M., Yamada, T., and Izumitani, T. (1999). Optimal size and location planning of public logistics terminals. *Transportation Research E*, 35:207 – 222.

Taniguchi, E., Thompson, R.G., Yamada, T., and van Duin, J.H.R. (2001). *City Logistics: Network Modelling and Intelligent Transportation Systems.* Pergamon, New York.

Taylor, G.D., Harit, S., English, J.R., and Whicker, G. (1995). Hub and spoke networks in truckload trucking: Configuration, testing and operational concerns. *Logistics and Transportation Review*, 31:209 – 237.

Teo, C.P., Ou, J., and Goh, M. (2001). Impact on inventory costs with consolidation of distribution centres. *IIE Transactions*, 33(2):99 – 110.

Thompson, R.G., and Taniguchi, E. (2001). City logistics and freight transport. In: A.M. Brewer, K.J. Button, and D.A. Hensher (eds.), *Handbook of Logistics and Supply-Chain Management*, pp. 393 – 405. Pergamon, New York.

Trappey, A. and Ho, P.-S. (2001). Human resource assignment system for distribution centers. *Industrial Management + Data Systems*, 102(1):64 – 72.

Umble, E.J., Haft, R.R., and Umble, M.M. (2003). Enterprise resource planning: Implementation procedures and critical success factors. *European Journal of Operational Research*, 146:241 – 257.

Vollmann, T.E., Berry, W.L., Whybark, D.C., and Jacobs, F.R. (2004). *Manufacturing Planning and Control Systems*, 5th edition. McGraw Hill-Irwin, New York.

Whiteing, T., Browne, M., and Allen, J. (2003). City logistics: The continuing search for sustainable solutions. In: D. Waters (ed.), *Global Logistics and Distribution Planning*, pp. 308 – 320, 4th edition. Kogan Page, London.

Yang, K.K. (2000). Managing a single warehouse, multiple retailer distribution center. *Journal of Business Logistics*, 21(2):162 – 172.

Chapter 4

OPERATIONAL RESEARCH METHODS FOR EFFICIENT WAREHOUSING

Gilles Cormier

Abstract The design and operation of a warehouse entail many challenging decision problems. We begin by providing definitions as well as qualitative descriptions of two actual warehouses. This will then set the stage for an overview of representative operational research models and solution methods for efficient warehousing. Problems which will be exposed can be classified into three major categories: throughput capacity models, storage capacity models, and warehouse design models. We conclude by identifying future research opportunities.

1. Introduction

Be they associated with grocery distribution, manufacturing or health care, warehouses are ubiquitous and come in almost all shapes and sizes. So it is certainly of considerable practical interest to identify methods for improving their design and operation, and these span the entire spectrum of analytical models (optimization and queuing) and simulation models. Problems which are surveyed here can be classified into three major categories: throughput capacity models, storage capacity models, and warehouse design models. Note that warehouse location models and container terminals (which serve as temporary buffers for inbound and outbound containers) are respectively examined in the chapters entitled "Facility Location in Supply Chain Design" and "Models and Methods for Operations in Port Container Terminals," in this book.

Throughput capacity models are comprised of order picking policies, akin to vehicle routing problems and which can be further subdivided between picking and batching policies, as well as storage assignment policies and dynamic control policies. Storage assignment policies attempt to match incoming product with available storage locations. Objective functions assumed in the study of these policies include the minimization of material handling cost (or equivalently, the maximization of throughput), as well as the minimization of material handling costs plus inventory holding and reordering costs.

Storage capacity models either find the optimal warehouse size or else maximize space utilization. Meantime, questions such as rack orientation, space allocation and overall building configuration are the purview of warehouse design models. Previous surveys on the use of operational research methods in warehouses were conducted by Ashayeri and Gelders (1985), who concluded that the most practical approach to studying the complexities of a warehousing system is to combine analytical and simulation models, and by Cormier and Gunn (1992), who pointed out that, while warehouses are usually part of a larger supply chain, studying the tradeoffs between all the latter's constituents poses both significant modelling and organizational challenges. The most recent such surveys are by Van den Berg and Zijm (1999) and Rouwenhorst et al. (2000).

Figure 4.1 is an attempt to categorize the various warehousing decision models and proposes a second classification based on a strategic, tactical and operational decision framework. Note that strategic decisions have a significant impact on long-term profitability and do not recur frequently, hence justifying the use of sophisticated analytical and simulation models. On the other hand, operational decisions tend to recur on a daily basis, or even more frequently for that matter, so that the main concern is in having algorithms which yield consistently good solutions quickly.

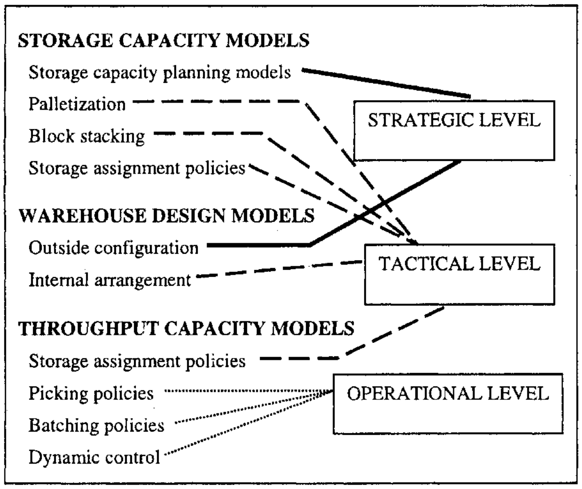


Figure 4.1. A taxonomy of warehousing decision models.

This therefore points the way to the development of efficient heuristics, given that most problems in this category are combinatorial. As for tactical models, they lie in-between strategic and operational models in importance and characterizing an ideal algorithm for them depends on specific circumstances, particularly execution frequency. Some readers might be surprised to find that storage assignment policies appear both under storage capacity models and throughput capacity models. Generally speaking, all storage capacity models exercise some influence over throughput capacity; for instance, think of how far you have to walk to get your groceries in a large grocery store as opposed to a small one.

The remainder of this chapter is organized as follows. In the next section, some terms which frequently appear in the warehousing literature are defined. In order to help the reader better understand the application context, we then describe, in Section 3, actual warehousing operations. This is followed in Section 4 by a presentation of performance evaluation models whose use transcends the three major decision model categories, namely, throughput capacity models, storage capacity models and warehouse design models. These are thereafter reviewed in Sections 5, 6 and 7, respectively. Finally, Section 8 recapitulates this chapter and identifies research gaps.

2. Definitions

Let us begin by defining some terms which are typically encountered in the warehousing literature. An *order* consists of a set of items destined to some customer and which must be retrieved from the warehouse. The *reorder quantity* is the amount of stock received by the warehouse at one time. A *rack* is a set of adjacent storage locations, while an *aisle* is the space in front of the rack where the order picking vehicle travels. The order picking vehicle can take several forms, for instance, a forklift truck, a hand cart, or, in the case of an automated storage and retrieval system (AS/RS), a S/R machine or crane.

Warehouses typically comprise a *reserve storage area*, where product is usually stored on pallets, as well as a *picking area*, where it is more common to place items on shelves or some other form of storage device. As open case stock in the picking area is depleted, new product is transferred from reserve storage to the picking area. Each area serves a specific purpose: in the reserve storage area the main concern is achieving high storage density whereas in the picking area the objective is to maximize picking efficiency.

Once the allocation of stock between the reserve and picking areas has been established, the items must be assigned to specific storage locations. In a *dedicated storage policy*, a set of locations is reserved for each product for the duration of the planning horizon. Furthermore, since the same priority is given to all units of the same product, these units are assigned to contiguous locations. A *shared storage policy*, on the other hand, allows units of different products to successively occupy the same locations. A common example of a shared storage policy is *random storage*, in which products are assigned to storage locations randomly. A popular hybrid approach is the *class-based dedicated storage policy*, which entails assigning products to a class of storage locations based on their class of turnover, while within any given class products are stored randomly.

A *cross-docking warehouse* is one in which incoming items are moved directly from receiving to shipping, thereby avoiding intermediate storage and retrieval. The *pickup and delivery (P/D) point* is the transfer point in and out of the warehouse (this term is most often associated with AS/RS's, and is analogous to a dock). *Single*, *dual* and *multi-command systems* refer to the number of locations which the order picking vehicle can visit between consecutive trips to the P/D point. The term *interleaving* signifies dual command as well, which consists of one leg from the P/D point to the first rack location where a pallet is placed, a second leg from the first location to a second location where a pallet is retrieved, and a final leg back to the P/D point where the retrieved pallet is deposited.

Travel time and distance in a two dimensional warehouse (that is, one in which the order picking vehicle's travel time depends on two axes, as in the case of a rack where neither horizontal nor vertical travel time dominates the other) may either be determined by the rectilinear norm or the Chebyshev norm. Travel time measurement according to the *rectilinear norm* implies that travel between any pair of locations occurs along only one Cartesian axis at a time, while in the case of the *Chebyshev norm* travel occurs in both directions simultaneously, albeit usually at different speeds. Let t_H and t_V denote the travel times from the P/D point to the farthest horizontal and vertical rack locations, respectively, and define the *rack shape factor* as $b = \min\{t_H, t_V\}/\max\{t_H, t_V\}$. The resultant *normalized rack* is called *square-in-time* if $b = 1$, meaning that travel time between the P/D point and the farthest horizontal location is identical to that to the farthest vertical location.

3. Examples of warehousing systems

Two actual warehousing operations are briefly described next and their distinct features underlined. The purpose here is twofold; first, to give the reader some insights into the organizational and technical issues often characterizing warehousing environments, and second, to motivate the modelling approaches reviewed in the remainder of this chapter.

3.1 A warehouse for grocery products

In this section, we outline order picking procedures in a large grocery warehouse. In addition to having a top and a bottom cross-aisle, the warehouse has a central cross-aisle. The order picking vehicle can carry two pallets at once, and since each order exceeds its capacity, the computerized warehouse management system (WMS) partitions each order into several tours. A dedicated storage policy is enforced inside the shelves, while the top of each shelf constitutes the reserve storage area in which a random storage policy is employed. The order picking activities are directed by the WMS under the following assumptions:

(1) Each tour is restricted to a single order.
(2) The central cross-aisle is not used.
(3) Aisles are always traversed in the same direction.
(4) All tours begin with the left-most aisle where items are to be picked, if the direction of travel in that aisle is from the docks to the back of the warehouse. Otherwise, the picker has to first travel empty to the back of the warehouse.

Tours hence follow a serpentine path, from the left side to the right side of the warehouse, with new tours started whenever the vehicle is full. Intuitively, relaxing the above assumptions can only lead to more efficient order picking. Some evidence supporting this conjecture was obtained through the adaptation of Clarke and Wright's (1964) heuristic, which, using data from an actual order, yielded savings (in terms of distance) of 13% compared to the WMS. Moreover, some results obtained by Racine (2000) on the subject of order batching in this situation are given in Section 5.1.

3.2 A warehouse for hardware products

This zoned warehouse, used for storing hardware products, is divided into three major storage areas: i) a storage area where a dedicated storage policy is used and non-conveyable products are kept, for instance, shower units and lawn mowers; ii) a reserve storage area consisting of a number of zones in which a random storage policy is in force and conveyable high volume products are stored; and iii) a forward pick area, operating under a dedicated storage policy and where many of the products from the reserve storage area are also kept for the purpose of open case picking. Note that, given that the random storage policy makes no distinction between zones, a product could be stored in any number of locations and zones.

A *wave* regroups some or all the orders that constitute a truckload. There is often a single wave per truckload, but sometimes two or three. Orders cannot be split between waves, although portions of orders on the same wave are generally picked in different zones. The purpose of waves is mostly to keep the work flowing smoothly at the sortation station, just prior to shipping. On the one hand, a wave cannot be too big since the area around the sortation station only allows for the accumulation of about six orders at a time. On the other hand, a wave cannot be too small as this reduces the probability of each zone having items to pick, resulting in a workload imbalance.

A worker determines the number of waves, and the computer performs the assignment of orders to waves. It should further be noted that each zone picks on a single wave at a time, and that the waves are taken in the same sequence in all zones, although all zones are not necessarily working on the same wave simultaneously. Picking is done mostly during the day and midnight shifts, with re-stocking (put-away) carried out during the evening shift, along with some limited picking. A *pick list* is a document specifying a *work assignment*, i.e., the set of items constituting a *batch* which is a subset of a *wave* to be picked on a single tour of an order picking vehicle. Moreover, each *line* on a pick list specifies the number of units of a certain product demanded on an *order*, the latter of which originates with a particular customer.

Note that an order cannot be split between work assignments in the same zone; conversely, it is usual for several orders to appear on the same work assignment. However, batches in different zones are independent of one another. Furthermore, in each zone, a work assignment is restricted to a certain number of orders and lines owing to the configuration of the order picking vehicles.

4. Performance evaluation models

Travel-time models can be useful for comparing both alternative operating scenarios and warehouse designs. Such a model, derived by Bozer and White (1984), includes several P/D point locations and dwell point strategies, the latter of which involve dynamically positioning the S/R machine when it becomes idle after completing a cycle. Other factors that can be incorporated in travel-time models include, for instance, the acceleration and deceleration of the S/R machine, maximum velocity restrictions, and various travel speeds; see Hwang and Lee (1990) as well as Chang et al. (1995).

Seidmann (1988) developed a travel-time model for the situation in which the number of items to be picked is a random variable, while Elsayed and Unal (1989) obtained an expression to estimate the travel-time as a function of the number of locations to be visited and the physical configuration of the warehouse. Expressions for upper and lower bounds on travel time are also developed. Hwang and Ko (1988) derived travel-time expressions for multi-aisle AS/RS's, assuming that the S/R machine is transferred between adjacent aisles by a "traverser." They also investigate the problem of partitioning the aisles into a number of classes so as to minimize the required number of S/R machines subject to the throughput constraint, each class having a dedicated S/R machine. Kim and Seidmann (1990) show that previously published models are special cases of their own throughput rate expressions.

It is noteworthy that Riaz Khan (1984) appears to be the only author who devotes a paper to the efficiency measurement of warehouse personnel. The proposed model estimates the time required to complete a picking cycle, considering lead time, travel time and non-efficient time. Foley and Frazelle (1991) assume the time required for the picker to retrieve items from containers to be either deterministic or exponentially distributed. Their purpose is to determine the maximum throughput at which a miniload AS/RS can process requests, as a function of such parameters as rack dimensions, S/R machine speed, and so on. They also derive closed-form expressions for the probability distribution function of dual command travel time, the utilization of the picker, and the utilization of the S/R machine. Meanwhile, the expected value of S/R machine travel time for multi-command order picking was derived as a function of the number of addresses and rack area by Guenov and Raeside (1992). Kouvelis and Papanicolaou (1995) present travel-time formulas for a two-class-based storage and retrieval system, for both single and dual command cycles, and obtain the optimal boundary between the two storage areas. A review of travel-time models is provided by Sarker and Babu (1995).

A framework for a dual command cycle travel-time model under class-based storage assignment is described by Pan and Wang (1996). De Koster (1994) presents a method which is based on Jackson network modelling and analysis (Jackson, 1957) for estimating the throughput performance of pick-to-belt order picking systems. Malmborg and Al-Tassan (2000) formulate travel-time models for single and dual command transactions in less than unit load order picking systems. These are applied to predict the operating performance of a reorder point stock management system with respect to item retrieval throughput capacity, physical storage space requirements, inventory service level and system responsiveness. Throughput models for unit-load cross-docking were developed by Gue and Kang (2001). For that purpose, they introduce a new type of queue, called a *staging queue*, which is characterized partly by the fact that, as the server pulls a pallet from the queue, the remaining pallets do not automatically move forward. Other travel-time models were proposed by Chew and Tang (1999) and by Koh et al. (2002), the latter of which considered the crane to be located at the centre of a round storage area.

In many complex situations it is necessary to resort to simulation models in order to quantify performance measures. Assuming a single-aisle dual command AS/RS, Azadivar (1986) constructed a simulation model in order to evaluate system response under various operating policies. An optimization problem is solved which maximizes throughput while respecting upper bounds on maximum queue length and average waiting time, as well as the acceptable risks with which the constraints can be violated. In addition, since warehouses are an integral part of global supply chains, Mason et al. (2003) develop a discrete event simulation model of a multi-product supply chain to assess the total cost reductions that can be achieved through the increased global visibility provided by integrating transportation and warehouse management systems.

5. Throughput capacity models

By some estimates, order picking costs account for about 55% of the recurring costs of operating a warehouse (Tompkins et al., 2003). It is therefore hardly surprising that researchers and companies alike have devoted so much effort toward improving the efficiency of order picking operations, be it at the operational level, in particular routing and batching policies, or at the tactical level, namely, storage assignment policies. A review of these policies will be followed by a brief discussion of dynamic control of warehouses.

5.1 Order picking policies

We hereafter review order picking policies, which are further divided between routing policies, analogous to the TSP, and batching policies.

5.1.1 Routing policies. Kanet and Ramirez (1986) propose a mixed zero-one nonlinear programming formulation for the problem of selecting from alternate picking locations on single command tours so as to minimize a combination of breakdown cost along with fixed and variable picking costs. The variable cost is a function of travel time while the fixed cost depends on such things as pallet loading and unloading times. Whenever the batch of stock retrieved exceeds the quantity required, the excess quantity must either be returned to storage or applied to another requisition, thereby causing the so-called breakdown cost to be incurred.

Graves et al. (1977), Han et al. (1987), Eynan and Rosenblatt (1993) as well as Lee and Schaefer (1996) all analyzed dual command storage and retrieval systems (Han et al.'s paper is discussed in Section 5.3). Considering the operating characteristics of a man-on-board storage and retrieval system, Hwang and Song (1993) present a heuristic procedure for the problem of sequencing a given set of retrieval requests. They also develop expected travel-time models based on a probabilistic analysis for single and dual commands assuming a random storage assignment policy.

The following papers all assume a multi-command system. Given a Chebyshev rack, Bozer (1985) derived an analytical expression for the expected tour length of the band heuristic, as well as the optimal number of bands as a function of the number of picks. The band heuristic divides the rack into a number of horizontal bands, with picking performed following a serpentine path defined by those bands.

Bozer et al. (1986, 1990) compared a number of tour construction heuristics, enhanced by some tour improvement routines such as 2-*opt* and 3-*opt* exchanges. The best were found to be the half-band insertion heuristic and the convex hull heuristic, the former running about 60% faster than the latter. Moreover, the decrease in tour length does not seem to warrant the additional implementation complexity and increased computation time of the *k*-*opt* exchanges. The convex hull heuristic consists of two phases. First, the convex hull of all the points to be visited is determined. If all the points are on the boundary, the tour is optimal; otherwise, the remaining points are inserted individually so as to minimize additional travel time. The half-band insertion heuristic starts by crossing out the middle half area of the rack. Points in the remaining top quarter area are joined sequentially, followed by the same procedure for points in the remaining bottom quarter area. Finally, the points in the crossed-out middle half are inserted using the same procedure as the convex hull heuristic.

Assuming a warehouse consisting of a single block of parallel aisles with crossovers only at the ends of the aisles, Ratliff and Rosenthal (1983) developed a procedure for finding an optimal picking sequence requiring a computational effort linear in the number of aisles. The dynamic programming algorithm proposed is based on graph theory, with the nodes corresponding to the depot along with the top and bottom of every aisle while the arcs connect nodes that can be visited consecutively. This method was extended by De Koster and Van der Poort (1998) to the case where the vehicle can pickup and deposit loads at the head of every aisle. Roodbergen and De Koster (2001a,b) as well as Vaughan and Petersen (1999), show the benefits of using cross-aisles. The latter show that such benefits are a function of the length of the main aisle. Furthermore, the number of cross-aisles that should be provided depends upon the total number of aisles, the number of picks per aisle as well as the ratio between the main aisle length and the cross-aisle width. Indeed, too many cross-aisles can effectively increase tour lengths.

Goetschalckx and Ratliff (1988) show that the optimal traversal picking tour is obtained by finding the shortest path in an acyclic graph. A traversal tour is one in which the vehicle enters at one end of the aisle and exits at the other, picking from both sides simultaneously. They also discuss a procedure for finding the optimal z-pick tour, where each slot is picked in a fixed sequence that remains the same for all orders. Meanwhile, under the assumption of a dedicated storage policy, Hall (1993) compares several routing strategies, namely, the traversal, midpoint, largest gap, and double traversal strategies, on the basis of number of picks and the warehouse's geometry (aisle width and length). With the double traversal strategy the picker enters each aisle from both ends, picking from only one side each time. The midpoint and largest gap strategies, both variants of the return strategy in which each aisle is entered and exited from both ends, differ in the criteria used to determine at which point the picker turns around. That is, in the former, the picker simply turns around at the middle of each aisle, while in the latter, the picker turns around at the point where the gap between successive items is greatest.

For his part, Petersen (1997), assuming a random storage policy, analyzed various routing strategies as well as the effect of the warehouse dimensions and the location of the pickup and delivery point, while the impact of routing strategies and storage assignment policies on warehousing efficiency is reported in Petersen (1999). In addition, Caron et al. (1998) consider jointly storage assignment policies and routing policies, while Hwang et al. (2001) developed travel-time models for traversal and return travel policies, which were then compared with respect to various ABC curves, number of picks, and length to width ratios of the warehouse. Under the assumption that each product may be picked from alternative locations, Daniels et al. (1998) formulate a model for simultaneously assigning inventory to an order and routing. Assuming a given order sequence, Van den Berg (1996) presents an efficient dynamic programming algorithm for the problem of sequencing picks in a set of orders on a single carousel. He then considers the problem in which the orders are not sequenced and simplifies this problem to a rural postman problem on a circle, solving it to optimality. For their part, Bartholdi and Gue (2000) concentrate on labour costs in a cross-docking terminal, while Apte and Viswanathan (2000) give an overview of cross-docking and discuss its various managerial aspects.

Whereas the previous studies focused mainly on distance minimization, Cormier (1987) describes an order picking problem in which the objective is to minimize the total weighted tardiness incurred when items are not delivered to the pickup and delivery point before their respective due-dates. Lee and Kim (1995) consider the problem of scheduling storage and retrieval orders under dual command operations in a unit-load automated storage and retrieval system. The objective is to minimize the weighted sum of earliness and tardiness penalties about a common due-date. Note that some of the methods developed for generic vehicle routing problems may also be applied to warehouses, see for instance the chapter in this book entitled "New Heuristics for the Vehicle Routing Problem."

5.1.2 Batching policies. Bozer (1985) identifies the following batching alternatives: single-order picking, batch picking, and zone picking. Under single-order picking, different orders are never combined on the same trip of the order picking vehicle. By contrast, batch picking relaxes this restriction. In zone picking, each vehicle, or picker, operates within specific geographical boundaries of the warehouse (as in the warehouse in Section 3.2). Pick-to-pack systems are a type of zone picking where items are placed directly in the shipping container and the container is transferred between zones. Batch picking can result in savings over single-order picking whenever items on different orders can be processed together and are located in close proximity in the warehouse. In addition, recall from Section 3.2 the concept of waves, whose purpose is to smooth the workflow by essentially assigning orders to time windows so that only orders within the same wave can be batched together. Meller (1997) proposes an algorithm to assign orders to lanes based on the arrival sequence of items to the sortation system. Significant throughput increases are achieved, with throughput based on the time to sort a complete order pick-wave. Moreover, Gue (2001) seeks to determine the optimal timing of pick-waves to minimize average order cycle time. He also contends that his proposed "percent making cut-off" metric, which establishes a cut-off time for orders to be guaranteed shipping on the next delivery cycle, is better for warehouses using cyclical transportation providers. Almost all of the literature on order batching assumes that several orders are to be combined on the same tour of the order picking vehicle and that orders cannot be split between tours.

Hwang et al. (1988) describe a clustering algorithm for batching in a Chebyshev rack. Elsayed and Unal (1989) compared four different order batching heuristics, under the assumptions that the number of orders is normally distributed and that the number of different items in an order and the quantity of each item are uniformly distributed. Their best algorithm entails first classifying each order as large or small relative to a preset fraction of vehicle capacity. Large orders are then combined in pairs for which savings over single-order picking are computed. The pair yielding the largest savings is kept and the process is repeated until all large orders have been assigned to batches. Small orders are thereafter considered in the same fashion, starting with the one having the largest quantity.

Gibson and Sharp (1992) compared three order batching heuristics for different experimental factors, among others, the distance measure, the order size and the storage assignment policy. They found that the method which outperforms the others consists of starting a new batch with an arbitrary seed order, and then augmenting the batch with other orders by minimizing a certain distance measure, starting new batches as necessary to ensure that vehicle capacity is not exceeded. The key to the efficiency of this algorithm is obviously the accuracy of the distance measure, which is obtained without solving any TSP's. A comparative study of order batching algorithms was carried out by Pan and Liu (1995). De Koster et al. (1999) conducted further comparisons between order batching methods in combination with the routing method (traversal and largest gap) and the warehouse type. They recommend choosing the routing method before the batching method. In narrow-aisle warehouses with pallet racks, the best batching algorithm was found to be one based on Clarke and Wright savings.

By proposing a new distance measure between orders, called minimum additional aisle (MAA), used in the procedure for selecting the order which is added to the seed order, Rosenwein (1996) obtained better results than did Gibson and Sharp, both in terms of distance (33% reduction in number of aisles traversed) and number of tours (12% reduction), but at the expense of more computer time. Gademann et al. (2001) developed an optimal branch and bound method for order batching to minimize the maximum picking time in all zones in a zoned warehouse. Meanwhile, Jewkes et al. (2004), under the assumption of out-and-back routing, consider the concurrent problems of product location, picker home base location, and allocation of products to each picker so that the expected order cycle time is minimized.

Ruben and Jacobs (1999) studied jointly batching and storage assignment policies, the performance measures being person-hours, distance, as well as the utilization of workers and the capacity utilization of the order picking vehicles. They conclude that the "First Fit-Envelope Based Batching" (FF-EBB) method is the most effective, which models the batching problem as a bin-packing problem and uses the gap between the first and last aisles traversed by each order as the distance measure between them. For their part, Elsayed et al. (1993) tackled routing and batching problems in the presence of both earliness and tardiness penalties, for which they propose a priority index methodology. Sequencing and batching of storage and retrieval requests to minimize total tardiness was considered by Elsayed and Lee (1996).

By contrast to the foregoing, the size of individual orders in some situations exceeds the capacity of the vehicle, so that the methods for order batching studied in the literature do not apply. Racine (2000) hence tackled the batching problem arising in Section 3.1. Due to the fact that the savings resulting from combining orders are attributable mainly to the reduction in stopping and item identification times, the coupling model proposed uses as a measure of performance the number of items in common between each pair of orders. The resulting solution was compared with the existing single-order picking method and found to yield reductions of 3% in the number of tours, 15% to 20% in the distance and 6% to 10% in picking time. Somewhat smaller savings were also achieved by using the cross-aisle and by solving the TSP optimally.

5.2 Storage assignment policies

In the dedicated storage assignment case, many studies have been published for the purpose of minimizing average workload, beginning with Heskett (1963), who proposed the cube-per-order index (COI) rule. The COI of an item is defined as the ratio of its total required space to its turnover, while Heskett's algorithm locates the items with the lowest COI closest to the pickup and delivery point. Assuming that the travel independence condition holds, implying that the cost of moving all items is constant and proportional to the distance travelled, several studies have since proved the optimality of this rule under various conditions, e.g., Harmatuck (1976), for single command systems, Malmborg and Krishnakumar (1987), for dual command systems, and Malette and Francis (1972), for a single command multi-dock facility in which all items have the same probability mass function for selection of a dock. Malmborg (1995) extended the COI methodology to the case of zoned warehouses.

Assuming racks of equal sizes and the same space utilization for all methods, Hausman et al. (1976) demonstrated that turnover-based dedicated storage is significantly better than random storage, while Rosenblatt and Eynan (1989) report that a four-class (twelve-class) system provides 90% (99%) of the benefits of full turnover-based storage assignment. However, Goetschalckx (1983) shows that methods such as random storage can in fact increase space utilization. Thonemann and Brandeau (1998) demonstrate that both turnover-based and class-based storage assignment policies in a stochastic environment reduce expected storage and retrieval time compared with the random storage policy. Malmborg (1996) developed a method which can estimate space and retrieval efficiency of random and dedicated storage policies and notes that average retrieval costs may decline with the number of storage slots utilized. Using simulation, Linn and Wysk (1987) concluded that random storage is best for low space utilization, while turnover-based dedicated storage is better at very high space utilization. In the paper by Montulet et al. (1998), mixed integer programming models are presented for the problem of minimizing, over a fixed horizon, the peak load in single command dedicated storage systems.

Wilson (1977), Hodgson and Lowe (1982) along with Malmborg and Deutsch (1988) consider the problem of establishing jointly a dedicated storage policy and an inventory policy. For instance, Wilson's algorithm works by first setting all reorder quantities equal to the economic order quantity (EOQ) and allocating stock by the COI rule. A gradient search procedure is then used to generate a new reorder quantity vector, the COI is reapplied, and so on, until the variation in reorder quantities between successive iterations is very small. Furthermore, Hodgson and Lowe extend the COI rule to the case where the travel independence condition does not hold.

Situations in which items are picked together on multi-command tours were studied by Jarvis and McDowell (1991) as well as Rosenwein (1994), the latter of whom applied clustering analysis. Van Oudheusden et al. (1988), and Van Oudheusden and Zhu (1992), tackled the case in which orders recur according to a known probability. Moon-Kyu (1992) developed a heuristic for the storage assignment problem based on group technology considering both order structure and frequency. Space requirements along with storage location assignment were modelled jointly by Kim (1993), while a class-based storage assignment policy for a carousel system was developed by Ha and Hwang (1994).

Shared storage policies are the subject of studies by Goetschalckx and Ratliff (1990) and Montulet et al. (1997), who show that they yield reductions in both space and travel time compared to dedicated storage. Given the potential of these policies, we outline a heuristic proposed by Montulet et al. (1997) which generally outperforms that by Goetschalckx and Ratliff (1990). In fact, comparisons between the former and the optimal solutions, obtained by means of an exact formulation solved by column generation, reveal that Montulet et al. (1997)'s heuristic systematically finds the optimal solution. It is assumed that all items are identical from the point of view of the storage system and that single command travel is employed. Let $G = (X, A)$ where $X = \{1, 2, \ldots, T\}$ is the set of nodes which correspond to the dates over the planning horizon. Each unit of product to be stored has associated with it an arc in A, called *item arc*, which connects the arrival date node to the departure date node. Arcs connecting each consecutive dates constitute the remaining members of A and have zero weight. Let the weight of each item arc equal $K + DS_{item}$, where K is a constant greater than T and DS_{item} is the duration of stay of a particular item. The algorithm's pseudocode is as follows:

```
Repeat while $A$ contains item arcs:
■ Obtain the longest path (without cycles) from 1 to $T$.
■ Assign the items corresponding to the arcs on this path to the most accessible remaining available locations.
■ Delete from $A$ the item arcs along this path.
```

Finally, the allocation of items to an AS/RS was studied by Hackman and Rosenblatt (1990), the tradeoff to be optimized being the cost of replenishing the items assigned to the AS/RS from their other warehouse locations versus the savings per retrieval request if an item is stored in the AS/RS. A heuristic algorithm is developed based on the relationship between this problem and the knapsack problem.

5.3 Dynamic control of warehouses

This section presents methods for coping with operating environments that vary over time. In order to meet short-term throughput requirements of a fluctuating demand pattern, Jaikumar and Solomon (1990) examine the relocation of pallets which have a high expectancy of retrieval in a future time period closer to the pickup and delivery point. Likewise, Muralidharan et al. (1995) present a shuffling heuristic where random storage is employed for the initial storage assignment, but when the crane is idle, the more frequently accessed product is shifted nearer to the pickup and delivery point while the less frequently accessed product is shifted farther. Along the same lines, Sadik et al. (1996) describe a heuristic for the dynamic reconfiguration of the order picking system.

Linn and Wysk (1990) developed a prototype expert system in which the hierarchical control structure consists of strategic, tactical and process control levels. A multi-pass simulation technique is employed to adapt control policies to real-time system behaviour. Seidman (1988) also used such an artificial intelligence approach, while Knapp and Wang (1992), Lin and Wang (1995), and Hsieh et al. (1998) propose Petri nets for AS/RS operation modelling.

Han et al. (1987) propose the nearest neighbour heuristic for routing in a dual command warehouse, whereby each storage location is matched with the closest available retrieval location. Since the list of retrievals changes over time, a block of retrievals is selected, these retrievals are sequenced following which another block of retrievals is considered. The expected throughput of the nearest neighbour heuristic is shown to be within 8% of the upper bound on throughput for any block sequencing rule. Eben-Chaime (1992) alternatively proposes the dynamic application of the same rule with a resulting increase in performance. Kim et al. (2002) study an order picking problem where the pick location of goods can be selected in near real time, to which they apply an intelligent agent-based model. Egbelu (1991), Egbulu and Wu (1993) along with Hwang and Lim (1993) focused their attention on dwell-point strategies.

Bartholdi et al. (2001) introduce the concept of bucket brigades, which are a way of sharing work on a flow line that results in the spontaneous emergence of balance and consequent high throughput while requiring neither traditional assembly line balancing technology nor any central planning. They report a 35% increase in productivity at the national distribution center of a major chain retailer after the workers began picking orders by bucket brigade (see also Bartholdi and Hackman, 2003).

6. Storage capacity models

The following papers assume demand for space to be given, so that lot sizing is not incorporated in the modelling framework, which is not unreasonable in those instances where the purchasing department functions independently of the warehouse. In fact, using a discounted inventory cost approximation and a linear warehousing cost model, Cormier and Gunn (1996a) showed that such a sequential policy is near-optimal if the products are characterized by a very high purchasing cost relative to the marginal cost of the storage space (expensive jewellery is a good example of this). Goh et al. (2001) extended this framework to allow for a step function of the warehouse space to be acquired. They consider the cases of a single item, multi-items with separable costs together with multi-items with joint inventory replenishment costs.

White and Francis (1971), and Lowe et al. (1979) describe network flow formulations for some single-location, multi-period warehouse leasing problems, with demand specified by a probability mass function. The problem of determining the capacity of a single production facility along with the amount of warehouse space to lease in each of several regions is studied by Jucker et al. (1982).

Integrated models such as that of Cokelez and Burns (1989) are useful for coordinating various inter-related decisions. They present a mixed-integer linear programming model incorporating product mix, transportation, warehouse location and warehouse capacity. Also of considerable interest in a multi-item context, provided that a shared storage policy is employed, is the coordination of inventory cycles, with objective functions such as minimizing maximum storage, e.g., Zoller (1977), and minimizing the sum of ordering and holding costs, plus a cost determined by peak inventory levels (Hall, 1988; Rosenblatt and Rothblum, 1990; Anily, 1991).

Some situations warrant establishing capacity in view of achieving a target service level rather than minimizing costs. For instance, Rosenblatt and Roll (1988) used simulation in order to generate the cumulative distribution of the number of days requiring a certain capacity. Meanwhile, algorithms based on queueing theory proposed by Sung and Han (1992) yield the minimum number of storage spaces for which the blocking probability does not exceed a specified threshold.

If the items to be stored are fairly voluminous and inexpensive, e.g., sheets of polystyrene thermal insulation, then it is more appropriate to optimize inventory costs and space costs simultaneously. Assuming stationary demand, Herron and Hawley (1969) present analytical and graphical procedures for such a situation, while Levy (1974) restricts his attention to a single expansion under non-stationary demand. For the case of arbitrarily increasing product demand, Rao (1976) minimizes the sum of discounted production (comparable to product procurement), carrying, investment, and idle capacity costs, all of them concave. He proposes a discrete-time dynamic programming algorithm which is equivalent to finding the shortest path in an acyclic network.

Cormier and Gunn (1999) tackled a warehouse sizing problem in which product demands vary arbitrarily over a finite planning horizon and the expansion cost consists of fixed and variable components. The state variable and the stages in their proposed dynamic programming formulation correspond to the warehouse size and time periods, respectively. Moreover, under constant product demand, Cormier and Gunn (1996b) demonstrated through both analytical and numerical means that it can be worthwhile to lease space temporarily at the beginning of each inventory cycle. Qualitative research has also been done on the subject of outsourcing logistics services, including warehouses; see for instance Maltz (1992, 1994). The paper by Chen et al. (2001) considers multi-period warehousing contracts under random space demand characterized by a starting space commitment plus a certain number of times at which the commitment can be modified. Furthermore, Colson and Dorigo (2003) developed a database which allows the user to make a multiple criteria selection of a subset of public warehouses fitting as well as possible his or her needs and preferences.

Another subject related to storage capacity is that of the utilization of storage space. The unitization problem has been investigated by Steudal (1979), his objective being to partition a pallet into smaller identical rectangular areas so as to minimize the amount of unused pallet area. This is a special case of the two-dimensional cutting stock problem which allows non-guillotine cuts. Tsai et al. (1988) address the use of linear programming to determine an optimal solution to a similar problem, except that they allow a wide product mix of different box sizes to be loaded on the same pallet. Balasubramanian (1992) provides a survey of research relating to models and solution procedures used in pallet packing, while Dowsland and Herbert (1996) propose two new crossover operators used with genetic algorithms for the same purpose.

Carrying out a case study at a large distribution centre, Carlson and Yao (1996) develop decision rules that reveal that storage capacity could increase by at least 4% through optimum pallet stacking and a further 5 – 7% by standardizing the wooden pallets themselves. Additionally, Abdou and El-Masry (2000) devised a heuristic for random stacking, which entails loading boxes in undefined patterns incorporating load stability and box demand requirements. The container loading problem is formulated by Chen et al. (1996) as a zero-one mixed integer programming model, with consideration of multiple containers, multiple carton sizes, carton orientations, and the overlapping of cartons in a container. See also Scheithauer (1996), and Morabito and Morales (1998).

Block stacking is used for storing large quantities of palletized or boxed products on top of each other in stacks, without racks. Usually, forklift trucks are used to manipulate the pallets one at a time. A storage lane remains unavailable for arriving pallets until its current content has been totally depleted by demand, thereby creating the need to optimize storage lane depth. Marsh (1979) developed a simulation model in order to investigate the effect of alternate lane depths on space utilization, using statistical analysis to determine if they significantly influence performance measures such as primary storage area and lineal aisle frontage. Goetschalckx and Ratliff (1991) describe a dynamic programming algorithm for a single product and integer multiple lane depths, the states and stages of which correspond to the length and number of storage lanes, respectively. Heuristics are described for the case where lane depths are restricted to a finite set due to practical implementation considerations.

7. Warehouse design models

Warehouse design models attempt to optimize such things as the orientation of storage racks, the allocation of space among competing uses, the number of cranes and the overall configuration of the facility. Let us first consider papers that deal with the tactical design questions arising inside warehouses. Bozer (1985) develops performance models for in-the-aisle picking versus end-of-aisle picking, with the objective of minimizing cost subject to throughput and storage space constraints. He also analyzes the tradeoff, in terms of the increase in picking time versus the decrease in replenishment frequency (from reserve area to picking area) as the picking area is increased. With demand described as a Poisson process and an objective function comprising storage cost, runout cost (the cost of a stockout in the active pick zone), and picking cost, Bhaskaran and Malmborg (1990) as well tackled the question of relative sizing between the reserve storage area and the active pick area. Van den Berg et al. (1998) consider a problem where the objective is to determine which replenishments minimize the expected amount of labour during the picking period, the decision taking place prior to the picking period. For their part, Bozer and White (1990) analyzed end-of-aisle picking.

Bassan et al. (1980) compared two alternative shelf arrangements, considering material handling cost, annual cost per unit of storage area, and annual cost per unit length of external walls. The analysis yields the optimum number of storage spaces along a shelf, number of shelves, location of doors, as well as warehouse dimensions. Recognizing the reality that decision makers must often allocate scarce resources, Pliskin and Dori (1982) proposed a method for ranking alternative area assignments subject to the amount of space available. Likewise, Azadivar (1989) looks at allocating scarce floor space between a random access area and a rack storage area. Larson et al. (1997) outline a procedure for warehouse layout consisting of determination of aisle layout and storage zone dimensions, assignment of material to a storage medium, and allocation of floor space.

Now, turning our attention to overall warehouse design, an optimization model for the design of a dual command AS/RS was proposed by Ashayeri et al. (1985). The relative proportions of the various objective function terms and the convexity of the objective function in the number of aisles allow a one-dimensional sequential search over the number of aisles to yield the optimal solution. Park and Webster (1989a) compared alternative warehousing systems through exhaustive enumeration on the basis of the following factors: control procedure, handling equipment movement, storage assignment rule, input and output patterns for product flow, storage rack structure, and the economics of each storage system. In addition, Gray et al. (1992) propose a multi-stage hierarchical decision approach to solve a design model which encompasses warehouse layout, equipment and technology selection, item location, zoning, picker routing, pick list generation and order batching.

Park and Webster (1989b) extended the concept of square-in-time to that of cubic-in-time and subsequently presented an algorithm to design pallet rack storage systems based upon equipment characteristics. Malmborg (1994) proposes an analytical model which incorporates the tradeoffs between handling and storage requirements to support development of layout alternatives. The operating dynamics of factors such as production scheduling and part routing as well as handling and storage parameters are all captured and the author claims that his model provides a higher degree of modelling ease than either simulation or stochastic Petri nets. Yoon and Sharp (1996) characterized the general structure of order picking systems and proposed a design procedure consisting of an input stage, a selection stage and an evaluation stage. Bartholdi and Gue (2004) address the question of the best shape for a cross-docking warehouse and make recommendations based on the size (number of doors) of the facility.

Also quite common in evaluating warehouse designs is the use of computer simulation, given the complexity and stochastic nature of such systems. An hybrid method was developed by Rosenblatt et al. (1993), in which output values are passed back and forth between an analytical optimization model and a simulation model until target values of the performance measures are attained. Taboun and Bhole (1993) developed a simulation model of an AS/RS which they then use to study the effect of four different warehouse configurations, large and small holding pallets, and four sizes of stored items on system performance.

A simulation model developed by Randhawa and Shroff (1995) allows for the evaluation of the effect of layout arrangements and scheduling policies on the performance of unit-load automated storage and retrieval systems. Finally, Eben-Chaime and Pliskin (1997) investigate an integrative model of a warehouse, containing several S/R machines and considering performance measures such as response times, queue lengths, and utilization of the S/R machines. A simulation of the model demonstrates that economic gains are possible as a result of decreasing the number of S/R machines and reducing building space as a consequence of shorter queues.

8. Concluding remarks

Let us now recapitulate some of the observations made in this chapter and identify research gaps that are ripe for future study. While computer simulation can provide valuable insights in comparing alternative designs and operating scenarios, time and budget constraints often preclude its use, particularly in small and medium sized companies. A warehouse simulation package with some standard components would thus be most useful as a rapid modelling tool.

We learned from the case of the grocery warehouse in Section 3.1 that warehouse management systems (WMS) do not always optimize certain elements of the order picking process. Furthermore, as remarked by Barnes (1999), "very seldom does the base price of a WMS comprise the majority of the total system cost." The bottom line is that recurring operating costs should be taken into account when designing a warehouse or selecting a WMS. As well, research has revealed that order batching can reduce order picking costs, but this has to be traded off against the equipment and operating costs of sortation. It thus appears that further research on batching methods is justified.

As for the warehouse discussed in Section 3.2, not a lot of research has been done on zoned warehouses, especially establishing the loading sequence of delivery trucks, determining the number of waves, assigning orders to waves, determining which part of each order to pick in each zone, and assigning pickers to zones. Literature is also quite scant on the dynamic control of warehouses, and tackling this subject together with order picking with due-dates is all the more important in the presence of a just-in-time requirement.

Let us also stress that there is still a lot of room left for studying storage capacity problems, for instance, under the assumption of various underlying inventory policies. And, while cross-docking and order picking by bucket brigades would also benefit from further research, these newer concepts actually remind us of something more fundamental: Think out of the box!

We leave you with two additional references which we believe open up another realm of possibilities for warehousing research. Recognizing the fact that on-line resources are becoming more prevalent, the first is the *ERASMUS Logistica Warehouse Website*, http://www.fbk.eur.nl/OZ/LOGISTICA/, which is very rich in information and moreover contains links to all other important web sites dealing with warehousing research that we have found. The second is a paper by Brockmann (1999), who gives his opinion on which warehousing innovations seem to be the most promising for the future. These are listed below along with our suggestions on how operations research can help.

***Focusing on the customer*:** This dictates, among other things, developing relevant performance measures.

***Consolidation*:** Achieving optimal economies-of-scale is a natural byproduct of certain storage capacity planning models.

***Continuous flow of material and information*:** This provides further justification for the cross-docking concept.

***Information technology*:** Information technology will become more prevalent in warehousing, justified in part by embedded operations research methods that will reduce operating costs.

***Space compression*:** Proliferation of products results in greater space requirements as well as the need for judicious space allocation. Thus, we foresee a need for joint planning of storage capacity and storage assignment policies.

***Time compression*:** This means installing systems and methods that will allow time-based performance measures to be optimized.

***Distribution requirements planning*:** It is imperative that distribution plans adapt to changing customer orders, thereby justifying further research into the dynamic control of warehouses.

***Reverse logistics*:** Returned product has given rise to this field, for which allocating space and controlling labour costs is becoming more and more of an issue.

***Global supply chain optimization*:** This concept, which can only be envisaged through sharing of information among partners, is making it possible for suppliers and their customers to jointly optimize their operations.

***Third-party warehousing*:** Further research into optimal leasing arrangements is suggested.

***Automation*:** The increased availability of various levels of automation gives rise to an important decision problem: determining its optimal level. This requires analyzing the tradeoffs between, among other things, throughput, labour costs, equipment costs and flexibility. Further complicating this type of study is the fact that many of the pertinent factors are not readily quantifiable.

References

Abdou, G. and El-Masry, M. (2000). Three-dimensional random stacking of weakly heterogeneous palletization with demand requirements and stability measures. *International Journal of Production Research*, 38:3149 – 3163.

Anily, S. (1991). Multi-item replenishment and storage problem (MIRSP): Heuristics and bounds. *Operations Research*, 39:233 – 243.

Apte, U.M. and Viswanathan, S. (2000). Effective cross docking for improved distribution efficiencies. *International Journal of Logistics: Research and Applications*, 3:291 – 302.

Ashayeri, J. and Gelders, L.F. (1985). Warehouse design optimization. *European Journal of Operational Research*, 21:285 – 294.

Ashayeri, J., Gelders, L.F., and Van Wassenhove, L.A. (1985). A micro-computer-based optimization model for the design of automated warehouses. *International Journal of Production Research*, 23:825 – 839.

Azadivar, F. (1986). Maximization of the throughput of a computerized automated warehousing system under system constraints. *International Journal of Production Research*, 24:551 – 566.

Azadivar, F. (1989). Optimum allocation of resources between the random access and rack storage spaces in an automated warehousing system. *International Journal of Production Research*, 27:119 – 131.

Balasubramanian, R. (1992). The pallet loading problem: A survey. *International Journal of Production Economics*, 28:217 – 225.

Barnes, C.R. (1999). The hidden costs of a WMS. *IIE Solutions*, January:40 – 44.

Bartholdi III, J.J., Eisenstein, D.D., and Foley, R.D. (2001). Performance of bucket brigades when work is stochastic. *Operations Research*, 49:710 – 719.

Bartholdi III, J.J. and Gue, K.R. (2000). Reducing labor cost in an LTL crossdocking terminal. *Operations Research*, 48:823 – 832.

Bartholdi III, J.J. and Gue, K.R. (2004). The best shape for a crossdock. *Transportation Science*, 38:235 – 244.

Bartholdi III, J.J. and Hackman, S.T. (2003). *Warehouse & Distribution Science*. http://www.warehouse-science.com.

Bassan, Y., Roll, Y., and Rosenblatt, M.J. (1980). Internal layout design of a warehouse. *AIIE Transactions*, 12:317 – 322.

Bhaskaran, K. and Malmborg, C.J. (1990). Economic tradeoffs in sizing warehouse reserve storage area. *Applied Mathematical Modelling*, 14:381 – 385.

Bozer, Y.A. (1985). *Optimizing Throughput Performance in Designing Order Picking Systems*. Unpublished Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, GA.

Bozer, Y.A., Schorn, E.C., and Sharp, G.P. (1986). Analyzing picker-to-part order picking problems. *Proceedings of the 7th International Conference on Automation in Warehousing*, pages 155 – 164, IFS Publications.

Bozer, Y.A., Schorn, E.C., and Sharp, G.P. (1990). Geometric approaches to the Chebyshev traveling salesman problem. *IIE Transactions*, 22:238 – 252.

Bozer, Y.A. and White, J.A. (1984). Travel-time models for automated storage/retrieval systems. *IIE Transactions*, 16:329 – 338.

Bozer, Y.A. and White, J.A. (1990). Design and performance models for end-of-aisle order picking systems. *Management Science*, 36:852 – 866.

Brockmann, T. (1999). 21 warehousing trends in the 21st Century. *IIE Solutions*, July:36 – 40.

Carlson, J.G. and Yao, A.C. (1996). A visually interactive expert system for a distribution center environment. *International Journal of Production Economics*, 45:101 – 109.

Caron, F., Marchet, G., and Perego, A. (1998). Routing policies and COI-based storage policies in picker-to-part systems. *International Journal of Production Research*, 36:713 – 732.

Chang, D.T., Wen, U.P., and Lin, J.T. (1995). The impact of acceleration/deceleration on travel-time models for automated storage/retrieval systems. *IIE Transactions*, 27:108 – 111.

Chen, C.S., Lee, S.M., and Shen, Q.S. (1995). An analytical model for the container loading problem. *European Journal of Operational Research*, 80:68 – 765.

Chen, F.Y., Hum, S.H., and Sun, J. (2001). Analysis of third-party warehousing contracts with commitments. *European Journal of Operational Research*, 131:603 – 610.

Chew, E.P. and Tang, L.C. (1999). Travel time analysis for general item location assignment in a rectangular warehouse. *European Journal of Operational Research*, 112:582 – 597.

Clarke, G. and Wright, J.W. (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research*, 12:568 – 581.

Colson, G. and Dorigo, F. (2003). A public warehouses selection support system. *European Journal of Operational Research*, 153:332 – 349.

Cormier, G. (1987). On the scheduling of order picking operations in a single-aisle automated storage and retrieval system. In: A Kusiak (ed.), *Modern Production Management Systems*, pages 75 – 87, North-Holland, Amsterdam.

Cormier, G. and Gunn, E.A. (1992). A review of warehouse models. *European Journal of Operational Research*, 58:3 – 13.

Cormier, G. and Gunn, E.A. (1996a). Simple models and insights for warehouse sizing. *Journal of the Operational Research Society*, 47:690 – 696.

Cormier, G. and Gunn, E.A. (1996b). On the coordination of warehouse sizing, leasing and inventory policy. *IIE Transactions*, 28:149 – 154.

Cormier, G. and Gunn, E.A. (1999). Modelling and analysis for capacity expansion planning in warehousing. *Journal of the Operational Research Society*, 50:52 – 59.

Daniels, R.L., Rummel, J.L. and Schantz, R. (1998). A model for warehouse order picking. *European Journal of Operational Research*, 105:1 – 17.

De Koster, R. (1994). Performance approximation of pick-to-belt orderpicking systems. *European Journal of Operational Research*, 72:558 – 573.

De Koster, R. and Van der Poort, E. (1998). Routing order pickers in a warehouse: A comparison between optimal and heuristic solutions. *IIE Transactions*, 30:469 – 480.

De Koster, R., Van der Poort, E., and Wolters, M. (1999). Efficient orderbatching methods in warehousing. *International Journal of Production Research*, 37:1479 – 1504.

Dowsland, K.A. and Herbert, E.A. (1996). A family of genetic algorithms for the pallet loading problem. *Annals of Operations Research*, 63:415 – 436.

Eben-Chaime, M. (1992). Operations sequencing in automated warehousing systems. *International Journal of Production Research*, 30:2401 – 2409.

Eben-Chaime, M. and Pliskin, N. (1997). Operations management of multiple machine automatic warehousing systems. *International Journal of Production Economics*, 51:83 – 98.

Egbelu, P.J. (1991). Framework for dynamic positioning of storage/retrieval machines in automated storage/retrieval system. *International Journal of Production Research*, 29:17 – 37.

Egbelu, P.J. and Wu, C.T. (1993). A comparison of dwell point rules in an automated storage/retrieval system. *International Journal of Production Research*, 31:2515 – 2530.

Elsayed, E.A. and Lee, M.-K. (1996). Order processing in automated storage/ retrieval systems with due-dates. *IIE Transactions*, 28:567 – 577.

Elsayed, E.A., Lee, M.-K., Kim, S., and Scherer, E. (1993). Sequencing and batching procedures for minimizing earliness and tardiness penalty of order retrievals. *International Journal of Production Research*, 31:727 – 738.

Elsayed, E.A. and Unal, O.I. (1989). Order batching algorithms and travel-time estimation for automated storage/retrieval systems. *International Journal of Production Research*, 27:1097 – 1114.

Eynan, A. and Rosenblatt, M.J. (1993). An interleaving policy in automated storage/retrieval systems. *International Journal of Production Research*, 31:1 – 18.

Foley, R.D. and Frazelle, E.H. (1991). Analytical results for miniload throughput and the distribution of dual command travel time. *IIE Transactions*, 23:273 – 281.

Gademann, A.J.R.M., Van den Berg, J.P., and Van der Hoff, H.H. (2001). An order batching algorithm for wave picking in a parallel-aisle warehouse. *IIE Transactions*, 33:385 – 398.

Gibson, D.R. and Sharp, G.P. (1992). Order batching procedures. *European Journal of Operational Research*, 58:57 – 67.

Goetschalckx, M. (1983). *Storage and Retrieval Policies for Efficient Order Picking Operations*. Unpublished Ph.D. Thesis. Georgia Institute of Technology, Atlanta, GA.

Goetschalckx, M. and Ratliff, H.D. (1988). Order picking in an aisle. *IIE Transactions*, 20:53 – 62.

Goetschalckx, M. and Ratliff, H.D. (1990). Shared storage policies based on the duration stay of unit loads. *Management Science*, 36:1120 – 1132.

Goetschalckx, M. and Ratliff, H.D. (1991). Optimal lane depths for single and multiple products in block stacking storage systems. *IIE Transactions*, 23:245 – 258.

Goh, M., Jihong, O. and Chung-Piaw, T. (2001). Warehouse sizing to minimize inventory and storage costs. *Naval Research Logistics*, 48:299 – 312.

Graves, S.C., Hausman, W.H. and Schwarz, L.B. (1977). Storage-retrieval interleaving in automatic warehousing systems. *Management Science*, 23:935 – 945.

Gray, A.E., Karmakar, U.S. and Seidmann, A. (1992). Design and operation of an order-consolidation warehouse: Models and application. *European Journal of Operational Research*, 58:14 – 36.

Gue, K.R. (2001). *Timing Picking Waves in a Warehouse*. Working paper, Naval Postgraduate School, Monterey, CA.

Gue, K.R. and Kang, K. (2001). Staging Queues in Material Handling and Transportation Systems. *Proceedings of the 33rd Winter Simulation Conference*, pages 1104 – 1108, IEEE Computer Society.

Guenov, M. and Raeside, R. (1992). Zone shapes in class based storage and multicommand order picking when storage/retrieval machines are used. *European Journal of Operational Research*, 58:37 – 47.

Ha, J.-W. and Hwang, H. (1994). Class-based storage assignment policy in carousel system. *Computers & Industrial Engineering*, 26:489 – 499.

Hackman, S.T. and Rosenblatt, M.J. (1990). Allocating items to an automated storage and retrieval system. *IIE Transactions*, 22:7 – 14.

Hall, N.G. (1988). A multi-item eoq model with inventory cycle balancing. *Naval Research Logistics*, 35:319 – 325.

Hall, R.W. (1993). Distance approximations for routing manual pickers in a warehouse. *IIE Transactions*, 25:76 – 87.

Han, M.-H., McGinnis, L.F., Shieh, J.S., and White, J.A. (1987). On sequencing retrievals in automated storage/retrieval systems. *IIE Transactions*, 19:56 - 66.

Harmatuck, D.J. (1976). A comparison of two approaches to stock location. *The Logistics and Transportation Review*, 12:282 - 284.

Hausman, W.H., Schwarz, L.B., and Graves, S.C. (1976). Optimal storage assignment in automatic warehousing systems. *Management Science*, 22:629 - 638.

Herron, D.P. and Hawley, R.L. (1969). Establishing the optimum inventory size and stocking policy for a warehouse. *AIIE Transactions*, 1:75 - 80.

Heskett, J.L. (1963). Cube-per-order index - a key to warehouse stock location. *Transportation and Distribution Management*, 3:27 - 31.

Hodgson, T.J. and Lowe, T.J. (1982). Production lot sizing with material handling cost considerations. *IIE Transactions*, 14:44 - 51.

Hsieh, S., Hwang, J.-S., and Chou, H.-C. (1998). A Petri-net - based structure for AS/RS operation modelling. *International Journal of Production Research*, 36:3323 - 3346.

Hwang, H., Baek, W., and Lee, M.K. (1988). Clustering algorithms for order picking in an automated storage and retrieved system. *International Journal of Production Research*, 26:189 - 201.

Hwang, H. and Ko, C.S. (1988). A study on multi-aisle system served by a single storage/retrieval machine. *International Journal of Production Research*, 26:1727 - 1737.

Hwang, H. and Lee, S.B. (1990). Travel-time models considering the operating characteristics of the storage and retrieval machine. *International Journal of Production Research*, 28:1779 - 1789.

Hwang, H., Lee, Y.K., Lee, S., and Ko, C.S. (2001). Routing policies in an order picking operation. *Proceedings of the 16th International Conference on Production Research.*

Hwang, H. and Lim, J.M. (1993). Deriving an optimal dwell point of the storage/retrieval machine in an automated storage/retrieval system. *International Journal of Production Research*, 31:2591 - 2602.

Hwang, H. and Song, J.Y. (1993). Sequencing picking operations and travel time models for man-on-board storage and retrieval warehousing system. *International Journal of Production Economics*, 29:75 - 88.

Jackson, J.R. (1957). Networks of waiting lines. *Operations Research*, 5:518 - 521.

Jaikumar, R. and Solomon, M.M. (1990). Dynamic operational policies in an automated warehouse. *IIE Transactions*, 22:370 - 376.

Jarvis, J.M. and McDowell, E.D. (1991). Optimal product layout in an order picking warehouse. *IIE Transactions*, 23:93 - 102.

Jewkes, E., Lee, C., and Vickson, R. (2004). Product location, allocation and server home base location for an order picking line with multiple servers. *Computers & Operations Research*, 31:623 - 636.

Jucker, J.V., Carlson, R.C., and Kropp, D.H. (1982). The simultaneous determination of plant and leased warehouse capacities for a firm facing uncertain demand in several regions. *IIE Transactions*, 14:99 - 108.

Kanet, J.J. and Ramirez, R.G. (1986). Optimal stock picking decisions in automatic storage and retrieval systems. *OMEGA International Journal of Management Science*, 14:234 - 239.

Kim, B.-I., Graves, R.J., Heragu, S.S., and St-Onge, A. (2002). Intelligent agent modeling of an industrial warehousing problem. *IIE Transactions*, 34:601 - 612.

Kim, J. and Seidmann, A. (1990). A framework for the exact evaluation of expected cycle times in automated storage systems with full-turnover item allocation and random service requests. *Computers & Industrial Engineering*, 18:601 – 612.

Kim, K.H. (1993). A joint determination of storage locations and space requirements for correlated items in a miniload automated storage-retrieval system. *International Journal of Production Research*, 31:2649 – 2659.

Knapp, G.M. and Wang, H.-P. (1992). Modeling of automated storage/retrieval systems using Petri nets. *Journal of Manufacturing Systems*, 11:20 – 29.

Koh, S.G., Kim, B.S., and Kim, B.N. (2002). Travel time model for the warehousing system with a tower crane S/R machine. *Computers & Industrial Engineering*, 43:495 – 507.

Kouvelis, P. and Papanicolaou, V. (1995). Expected travel time and optimal boundary formulas for a two-class-based automated storage/retrieval system. *International Journal of Production Research*, 33:2889 – 2905.

Larson, T.N., March, H., and Kusiak, A. (1997). A heuristic approach to warehouse layout with class-based storage. *IIE Transactions*, 29:337 – 348.

Lee, M.-K. and Kim, S.-Y. (1995). Scheduling of storage/retrieval orders under a just-in-time environment. *International Journal of Production Research*, 33:3331 – 3348.

Lee, H.F. and Schaefer, S.K. (1996). Retrieval sequencing for unit-load automated storage and retrieval systems with multiple openings. *International Journal of Production Research*, 34:2943 – 2962.

Levy, J. (1974). The optimal size of a storage facility. *Naval Research Logistics Quarterly*, 21:319 – 326.

Lin, S.-C. and Wang, H.-P.B. (1995). Modelling an automated storage and retrieval system using Petri nets. *International Journal of Production Research*, 33:237 – 260.

Linn, R.J. and Wysk, R.A. (1987). An analysis of control strategies for automated storage and retrieval systems. *INFOR*, 25:66 – 83.

Linn, R.J. and Wysk, R.A. (1990). An expert system framework for automated storage and retrieval system control. *Computers & Industrial Engineering*, 18:37 – 48.

Lowe, T.J., Francis, R.L., and Reinhardt, E.W. (1979). A greedy network flow algorithm for a warehouse leasing problem. *AIIE Transactions*, 11:170 – 182.

Malette, A.J. and Francis, R.L. (1972). Generalized assignment approach to the optimal facility layout. *AIIE Transactions*, 4:144 – 147.

Malmborg, C.J. (1994). A heuristic model for simultaneous storage space allocation and block layout planning. *International Journal of Production Research*, 32:517 – 530.

Malmborg, C.J. (1995). Optimization of cube-per-order index warehouse layouts with zoning constraints. *International Journal of Production Research*, 33:465 – 482.

Malmborg, C.J. (1996). Storage assignment policy tradeoffs. *International Journal of Production Research*, 34:363 – 378.

Malmborg, C.J. and Al-Tassan, K. (2000). An integrated performance model for order-picking systems with randomized storage. *AppliedMathematical Modelling*, 24:95 – 111.

Malmborg, C.J. and Deutsch, S.J. (1988). A stock location model for dual address order picking systems. *IIE Transactions*, 20:44 – 52.

Malmborg, C.J. and Krishnakumar, B. (1987). On the optimality of the cube per order index for warehouses with dual command cycles. *Journal of Material Flow*, 4:169 – 175.

Maltz, A. (1994). Outsourcing the warehousing function: Economic and strategic considerations. *The Logistics and Transportation Review*, 30:245 – 265.

Maltz, A.B. (1992). The relative importance of cost and quality in outsourcing the warehousing function. *Journal of Business Logistics*, 15:45 – 62.

Marsh, W.H. (1979). Elements of block storage design. *International Journal of Production Research*, 4:377 – 394.

Mason, S.J., Ribera, P.M., Farris, J.A., and Kirk, R.G. (2003). Integrating the warehousing and transportation functions of the supply chain. *Transportation Research Part E: Logistics and Transportation Review*, 39:141 – 159.

Meller, R.D. (1997). Optimal order-to-lane assignments in an order accumulation/sortation system. *IIE Transactions*, 29:293 – 301.

Montulet, P., Langevin, A., and Riopel, D. (1997). Le problème de l'optimisation de l'entreposage partagé : méthodes exacte et heuristique. *INFOR*, 35:138 – 153.

Montulet, P., Langevin, A., and Riopel, D. (1998). Minimizing the peak load: An alternate objective for dedicated storage policies. *International Journal of Production Research*, 36:1369 – 1385.

Moon-Kyu, L. (1992). A storage assignment policy in a man-on-board automated storage/retrieval system. *International Journal of Production Research*, 30:2281 – 2292.

Morabito, R. and Morales, S. (1998). A Simple and effective recursive procedure for the manufacturer's pallet loading problem. *Journal of the Operational Research Society*, 49:819 – 828.

Muralidharan, B., Linn, R.J., and Pandit, R. (1995). Shuffling heuristics for the storage location assignment in AS/RS. *International Journal of Production Research*, 33:1661 – 1672.

Pan, C.-H. and Liu, S.-Y. (1995). A comparative study of order batching algorithms. *OMEGA International Journal of Management Science*, 23:691 – 700.

Pan, C.-H. and Wang, C.-H. (1996). A framework for the dual command cycle travel time model in automated warehousing systems. *International Journal of Production Research*, 34:2099 – 2117.

Park, Y.H. and Webster, D.B. (1989a). Modelling of three-dimensional warehouse systems. *International Journal of Production Research*, 27:985 – 1003.

Park, Y.H. and Webster, (1989b). Design of class-based storage racks for minimizing travel time in three-dimensional storage system. *International Journal of Production Research*, 27:1589 – 1601.

Petersen II, C.G. (1997). An evaluation of order picking routeing policies. *International Journal of Operations and Production Management*, 17:1098 – 1111.

Petersen II, C.G. (1999). The impact of routing and storage policies on warehouse efficiency. *International Journal of Operations and Production Management*, 19:1053 – 1064.

Pliskin, J.S. and Dori, D. (1982). Ranking alternative warehouse area assignments: A multiattribute approach. *IIE Transactions*, 14:19 – 26.

Racine, N. (2000). *Optimisation du prélèvement des commandes dans un centre de distribution.* Master's project report, Department of Mathematics and Industrial Engineering, École Polytechnique de Montréal.

Randhawa, S.U. and Shroff, R. (1995). Simulation-based design evaluation of unit load automated storage/retrieval systems. *Computers & Industrial Engineering*, 28:71 – 79.

Rao, M.R. (1976). Optimal capacity expansion with inventory. *Operations Research*, 24:291 – 300.

Ratliff, H.D. and Rosenthal, A.S. (1983). Order picking in a rectangular warehouse: A solvable case of the traveling salesman problem. *Operations Research*, 31:507 – 521.

Riaz Khan, M. (1984). An efficiency measurement model for a computerized warehousing system. *International Journal of Production Research*, 22:443 – 452.

Roodbergen, K.J. and de Koster, R. (2001a). Routing methods for warehouses with multiple cross aisles. *International Journal of Production Research*, 39:1865 – 1883.

Roodbergen, K.J. and de Koster, R. (2001b). Routing order pickers in a warehouse with a middle aisle. *European Journal of Operational Research*, 133:32 – 43.

Rosenblatt, M.J. and Eynan, A. (1989). Deriving the optimal boundaries for class-based automatic storage/retrieval systems. *Management Science*, 35:1519 – 1524.

Rosenblatt, M.J. and Roll, Y. (1988). Warehouse capacity in a stochastic environment. *International Journal of Production Research*, 26:1847 – 1851.

Rosenblatt, M.J., Roll, Y., and Zyser, V. (1993). A combined optimization and simulation approach for designing automated storage/retrieval systems. *IIE Transactions*, 25:40 – 50.

Rosenblatt, M.J. and Rothblum, U.G. (1990). On the Single resource capacity problem for multi-item inventory systems. *Operations Research*, 38:686 – 693.

Rosenwein, M.B. (1994). An application of cluster analysis to the problem of locating items within a warehouse. *IIE Transactions*, 26:101 – 103.

Rosenwein, M.B. (1996). A comparison of heuristics for the problem of batching orders for warehouse selection. *International Journal of Production Research*, 34:657 – 664.

Rouwenhorst, B., Reuter, B., Stockrahm, V., Van Houtum, G.J., Mantel, R.J., and Zijm, W.H.M. (2000). Warehouse design and control: Framework and literature review. *European Journal of Operational Research*, 122:515 – 533.

Ruben, R.A. and Jacobs, F.R. (1999). Batch construction heuristics and storage assignment strategies for walk/ride and pick systems. *Management Science*, 45:575 – 596.

Sadik, M., Landers, T.L. and Taylor, G.D. (1996). An assignment algorithm for dynamic picking systems. *IIE Transactions*, 28:607 – 616.

Sarker, B.R. and Babu, P.S. (1995). Travel time models in automated storage/retrieval systems: A critical review. *International Journal of Production Economics*, 40:173 – 184.

Scheithauer, G. (1996). The G4-heuristic for the pallet loading problem. *Journal of the Operational Research Society*, 47:511 – 522.

Seidmann, A. (1988). Intelligent control schemes for automated storage and retrieval systems. *International Journal of Production Research*, 26:931 – 952.

Steudal, H.J. (1979). Generating pallet loading patterns: A special case of the two-dimensional cutting stock problem. *Management Science*, 25:997 – 1004.

Sung, C.S. and Han, Y.H. (1992). Determination of automated storage/retrieval system size. *Engineering Optimization*, 19:269 – 2862.

Taboun, S.M. and Bhole, S.D. (1993). A simulator for an automated warehousing system. *Computers & Industrial Engineering*, 24:281 – 290.

Thonemann, U.W. and Brandeau, M.L. (1998). Note. Optimal storage assignment policies for automated storage and retrieval systems with stochastic demands. *Management Science*, 44:142 – 148.

Tompkins, J.A. White, Bozer, Y.A., and Tanchoco, J.M.A. (2003). *Facilities Planning.* John Wiley & Sons Inc., New York.

Tsai, R.D., Malstrom, E.M., and Meeks, H.D. (1988). A two-dimensional palletizing procedure for warehouse loading operations. *IIE Transactions*, 20:418 – 425.

Van den Berg, J.P. (1996). Multiple order pick sequencing in a carousel system: A solvable case of the rural postman problem. *Journal of the Operational Research Society*, 47:1504 – 1515.

Van den Berg, J.P., Sharp, G.P., Gademann, A.J.R.M., and Pochet, Y. (1998). Forward-reserve allocation in a warehouse with unit-load replenishments. *European Journal of Operational Research*, 111:98 – 113.

Van den Berg, J.P. and Zijm, W.H.M. (1999). Models for warehouse management: Classification and examples. *International Journal of Production Economics*, 59:519 – 528.

Van Oudheusden, D.J., Tzen, Y.J., and Ko, H. (1988). Improving storage and order picking in a person-on-board AS/R system. *Engineering Costs and Production Economics*, 13:273 – 283.

Van Oudheusden, D.L. and Zhu, W. (1992). Storage layout of AS/RS racks based on recurrent orders. *European Journal of Operational Research*, 58:48 – 56.

Vaughan, T.S. and Petersen, C.G. (1999). The effect of warehouse cross-aisles on order picking efficiency. *International Journal of Production Research*, 37:881 – 897.

White, J.A. and Francis, R.L. (1971). Normative models for some warehouse sizing problems. *AIIE Transactions*, 9:185 – 190.

Wilson, H.C. (1977). Order quantity, product popularity, and the location of stock in a warehouse. *AIIE Transactions*, 9:230 – 237.

Yoon, C.S. and Sharp, G.P. (1996). A structured procedure for analysis and design of order pick systems. *IIE Transactions*, 28:379 – 389.

Zoller, K. (1977). Deterministic multi-item inventory systems with limited capacity. *Management Science*, 24:451 – 455.

Chapter 5

MODELS AND METHODS FOR FACILITIES LAYOUT DESIGN FROM AN APPLICABILITY TO REAL-WORLD PERSPECTIVE

Nathalie Marcoux
Diane Riopel
André Langevin

Abstract This chapter first presents an extensive list of the strategic, tactical, and operational objectives found in the literature for facilities layout and handling system design. Based on these premises, the main objective of this chapter is to present a survey of operations research processors (models and methods) for the macro design problem, the focus being on their *applicability* to real-life problems.

1. Introduction

Facilities layout design — the core element of facilities layout planning — is still an important research issue even if, since the sixties, the evolution of research has greatly improved the tools for the solution of the facilities layout problem. Works prior to the computer era include elements of analysis for facilities layout, defined by pioneers such as Immer (1953), Reed (1961, 1967), Moore (1962), Apple (1963), and Nadler (1967). From the nineteen sixties, computerized techniques for the design or the improvement of a layout have been proposed. CRAFT, CORELAP, ALDEP, and PLANET are among the classical computer-aided techniques. The Systematic Layout Planning (SLP) method of Muther (1973) appears as a milestone for both research and practice.

Since then, there has been an intensification of research which has led to many models of optimization based on operational research tools. Also, techniques such as Graph theory, Expert systems, Simulated annealing, Tabu search, Fuzzy set theory, and Genetic algorithms were used to develop facilities layout design methods.

The main objective of this chapter is to present a survey of models and methods proposed in the literature. Unlike literature surveys associated with the mathematical comparison of models or methods for the facilities layout problem, this chapter focuses on their *applicability* for real-life problems. Three categories of characteristics are included in our analysis: materials, activities, and physical arrangement. Before discussing models and methods for facilities layout, we present a brief historical perspective.

Plant layout was the time-honored expression used by the pioneers in facilities planning. The terms *physical arrangement, efficiency, workforce, materials and machinery* are an integral part of each definition of plant layout. The most complete definition (Moore, 1962) is:

> "Plan of, or the act of planning, an optimum arrangement of industrial facilities, including personnel, operating equipment, storage space, materials-handling equipment, and all other supporting services, along with the design of the best structure to contain these facilities. Good plant layout is fundamental to the operation of an efficient industrial organization."

Moore (1962) emphasizes that the term *optimum* is related to whatever criteria may be chosen to evaluate a plant layout. This term was later on substituted by the word *efficiency*, taken up by both the industrial and research communities and defined as the output-input ratio. Outputs include the finished goods themselves, but also criteria such as cost, flexibility, safety, and so on. Inputs include all the resources needed to produce a product. They are related to *materials*, *equipment for production, handling, storage*, and *workforce*.

The terminological change from plant layout to facilities layout occurred in the early 1970's. However the confusion between the terms *design* and *layout* lasted until the end of that decade. In fact, the objectives of *plant layout* as defined in Apple (1963) are labeled *facilities design* in Apple (1977). Tompkins and White (1984) put an end to this confusion with a facilities planning hierarchy still in use in their latest book, Tompkins et al. (2003). This hierarchy is presented in Section 2 of this chapter.

Facilities layout design — one component of facilities planning — was the main topic of Muther's work (1973). He proposed a systematic methodology for facilities layout design called SLP — Systematic Layout Planning. He takes up several elements of data collection and of analysis introduced by the pioneers and integrates them into his methodology. Following his example, the expression "facilities layout" has finally been taken up by the industrial and scientific communities. Even today, his work is an important reference in the field of facilities layout. The Muther and Hales (1979) definition of facilities layout is equivalent to that of plant layout proposed by Moore (1962).

In the 1990's, the textbooks of Askin and Standridge (1993), Sule (1994), Sheth (1995), and Heragu (1997) are in agreement with Moore's definition. They however developed this definition, either by detailing certain elements, or by adding others. Wrennal (2001) presents a new formulation of the definition of facilities layout which includes that of Moore (1962) and, therefore, that of Muther and Hales (1979). The use of the term *resources* allows the integration of all materials, equipment, and workforce.

> "Facilities are the physical representation of the capacity of an operation. They promote or constrain the efficiency of operations. Facilities layout is the planning, designing, and physical arrangement of processing and support areas within a facility; the goal is to create a design that supports company and operating strategies. From the Latin *facilis*, meaning *easy*, a facility should free operations within it from difficulties or obstacles. A good layout optimizes the use of resources while satisfying other criteria such as quality, control, image, and many other factors."
>
> Wrennal (2001, p. 8.21)

Considering the various definitions — or parts of definitions — quoted in the literature, we propose an updated definition of facilities layout design:

> The physical arrangement in a certain space of all activities (e.g., production, handling, warehousing, and services to production and staff) related to materials, equipment and workforce to allow efficient production according to market specifications.

It should be noted that facilities layout design encompasses a much more complex process than just the physical arrangement of machines, workstations and support services. The definition of facilities layout design allows the answer to the question "What are the overall functions of facilities layout?" The next question is "What should facilities layout consider for efficiently fulfilling its functions?" Answering this question leads to an analysis of the models and the methods proposed in the literature with respect to the degree of realism of the characteristics of the problem studied.

This chapter presents the facilities planning hierarchy, including facilities layout design in Section 2. We present an important list of design objectives. Then Section 3 presents a list of input parameters and variables to be defined a priori and of problem characteristics related to the applicability of the models and methods for real-life problems. Section 4 reviews the models and methods proposed in the literature. A conclusion follows.

2. Facilities planning

In parallel with the definition of facilities layout, the concepts of *Facilities planning* and *Facilities design* must be clarified. As mentioned previously, Tompkins and White (1984) formalize a hierarchy linking those concepts. This hierarchy has been adopted by several other authors and an update of the terminology is presented by Tompkins et al. (2003). (Figure 5.1)

Along with this hierarchy, Muther (1973) presents a time-related framework for facilities planning. This framework consists of 4 phases: *Location*, *Overall layout*, *Detail layouts*, and *Installation.* Other authors such as Philips (1997), Wrennal (2001), and Heragu (1997) have revisited this framework to include other phases: *Needs analysis*, *Operations, and Follow up* respectively. Based on those works, an updated time-related framework is proposed in Figure 5.2. It includes the hierarchy of Tompkins et al. (2003) with a time dimension. Contrary to Figure 5.1, *Layout design* and *Handling system design* are combined in Figure 5.2. Indeed, the design of a layout is strongly influenced by the materials handling network and by personnel movements.

The methodology of Muther (1973), which integrates phases I through VIII, is still in use, e.g., see the books of Sule (1994) and Heragu (1997), and the work of Gómez et al. (2003). The methodologies presented by Wrennal (2001) and Tompkins et al. (2003) — called *affinity analysis* — are in fact a variant of Muther's methodology. In parallel to this methodology, Operations Research (OR) methods have been mainly used for addressing some of the phases: I. Location; III. Macro layout; and V. Detail layouts. For **Layout and handling system design**, phase

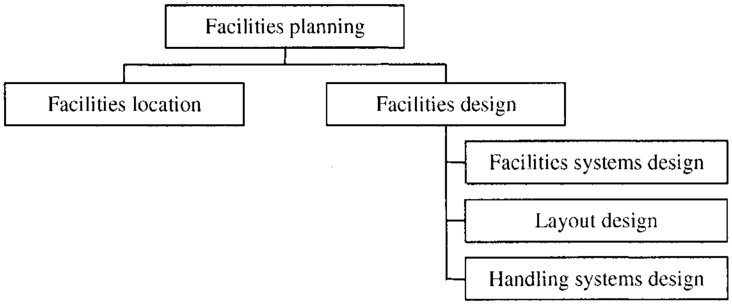


Figure 5.1. The facilities planning hierarchy

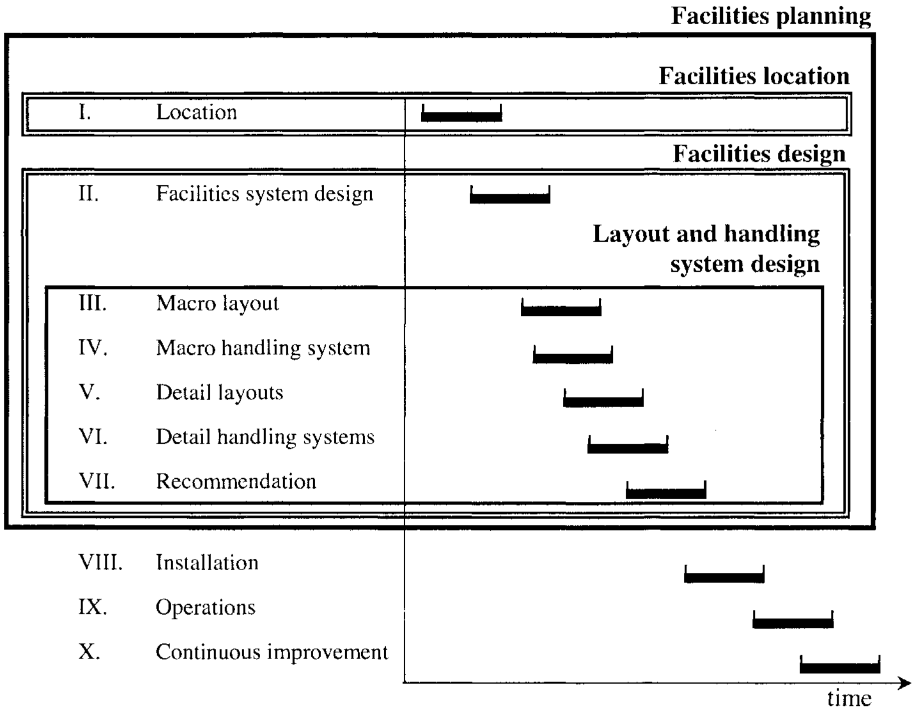


Figure 5.2. Time-related framework for facilities planning

III (Macro layout) is associated with *block layout.* This type of representation is considered classic due to its widespread use in research. Two basic elements of phase IV (Macro handling system) are the materials handling equipment and the aisle network. For the former, most OR models do not consider the selection of materials handling equipment. For the latter, there have been some attempts at integrating phases III (Macro layout) and IV (Macro handling system). This integration, called ***Macro design***, is the focus of this chapter. Finally, from an OR perspective, phases V (Detail layouts) and VI (Detail handling systems) can be associated with the machine layout problem, the layout problem for automated guided vehicle systems, or the layout problem for flexible manufacturing systems. Due to similarities between the macro design and the machine layout problem, we include the latter in our analysis.

Between 1960 and 1980, a number of computerized heuristic methods were developed to help the industrial engineer in the design or the improvement of facilities layouts. At the beginning of the 1980's, new heuristics based on simulation, graph theory, artificial intelligence, and other approaches were proposed. Furthermore, with the development of

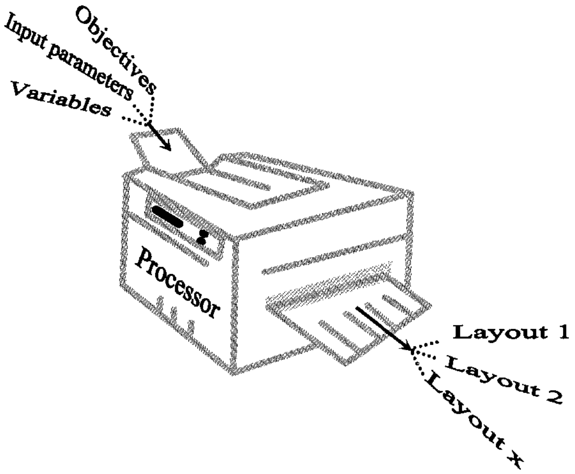


Figure 5.3. Process for macro design

computer technology, optimal approaches based on mathematical programming are widely used.

The process related to macro design, as illustrated in Figure 5.3, has 5 components: the objectives, the input parameters, the variables, the processor, and the layouts. The **processor** represents an operational research model or method for facilities layout design. It uses **input parameters** and **variables** to be quantified, both of which are translated in terms of an **objective** function and constraints for the mathematical representation of the facilities layout problem. The quantification of all the variables leads to the generation of a feasible **layout**. For facilities layout design, even though it is recommended to generate several solutions which will be further evaluated, OR models and methods are usually concerned with the generation of a single final solution.

The generation of layouts requires defining one or several objectives. These objectives could either be translated in terms of an objective function or in terms of layout evaluation criteria. Several authors present a more or less exhaustive list of objectives to consider. For example, Muther (1973) proposes 20 key points to consider in making an evaluation of a facilities layout. More recently, Tompkins et al. (2003) enumerate 35 criteria for evaluating a layout. Table 5.1 presents a list of objectives quoted by various authors. These objectives are classified as strategic, tactical or operational. As detailed in Section 4, only a few of the objectives have been considered in the various OR models and methods for the facilities layout problem. Of these, the most usual objectives (in ***bold italics*** in Table 5.1) are related to: flow (especially of materials), handling, space and equipment use, and capital investment.

3. Input parameters and variables

The input parameters correspond to all the data and information to be gathered a priori and used as input to the processor. For instance, the characteristics of the products to be manufactured and the appropriate type of layout (by products, by processes, cellular, or other) may have a substantial impact on the design of the layout. The variables are the elements whose value must be determined by the processor. Examples of possible variables are: the sizes and shapes of the departments and the building.

In the 1960's, engineering analysis and design methods included elements (input parameters and variables) defined as initial parameters, system characteristics, constraints to satisfy, variables to determine, and others. A comprehensive checklist, presented by Apple (1963, 1977), enumerates 56 elements grouped under eight themes: Materials or product; Moves; Handling methods; Process; Building; Site; Personnel; and Miscellaneous. Since then, and even today, textbooks in facilities layout design make reference to this list or enumerate a subset of input parameters and variables with the addition of certain new elements. The reader will find in the Appendix an updated list. To simplify the presentation, only the input parameters and variables not cited by Apple (1963, 1977) are referenced. In order to link the elements with the OR models and methods for facilities layout design, each element is identified as a variable or as an input parameter. It should be noted that some elements could be considered as variables or as input parameters according to the context (construction or improvement), or according to the processor used.

Of all these elements, only some are considered in the literature for the development of processors. These input parameters and variables can be translated into a set of **problem characteristics** appropriate for these tools, as proposed by Marcoux (1999). These characteristics are used in a comparative analysis of the processors in the following section. Table 5.2 presents the correspondence between the input parameters and variables considered in the literature and the problems' characteristics. They are grouped according to the components of the definition of the facilities layout design previously defined: materials, activities, and physical arrangement. In the context of macro design, the equipment component, used mainly in detail design, is left out. Also, since no processor proposed in the literature includes the workforce com-

Table 5.1. Objectives of facilities layout and handling system design

Strategic objectives	
- Be able to meet forecasted capacity needs (Adaptability and Versatility)	Muther (1973), Wrennal (2001), Tompkins et al. (2003)
- Plan for future expansion (Modularity)	Reed (1961), Muther (1973), Cedarleaf (1994), Sheth (1995), Tompkins et al. (2003)
- Be consistent with company image, appearance, promotional value, public or community relations	Reed (1961), Muther (1973), Wrennal (2001), Tompkins et al. (2003)
- ***Optimize capital investment (initial investment, installation fixed costs, start-up costs, annual operating costs, maintenance costs, return on investment, payback period)***	Moore (1962), Apple (1963), Muther (1973), Sheth (1995), Tompkins et al. (2003)
- Minimize impact on production during the installation period (including training and debugging)	Tompkins et al. (2003)
- Minimize negative effect on environment (including use of energy)	Tompkins et al. (1996)
- Integrate with external elements (other facilities, transportation)	Muther et Haganäs (1969), Tompkins et al. (2003)
Tactical objectives	
- Fit with organization structure	Muther (1973), Tompkins et al. (2003)
- Facilitate supervision, control, and communication	Muther (1973), Apple (1977), Heragu (1997), Tompkins et al. (2003)
- ***Optimize space utilization***	Moore (1962), Apple (1963), Muther (1973), Sheth (1995), Heragu (1997), Tompkins et al. (2003)
- Provide overall simplification, standardization	Moore (1962), Apple (1963)
- Maintain flexibility of arrangement and of operations	Reed (1961), Apple (1963), Muther (1973), Sheth (1995), Tompkins et al. (2003)
- Maximize storage and supporting services	Apple (1977), Muther (1973), Tompkins et al. (2003)
- Optimize use of natural conditions, building or surroundings	Muther (1973)
- Facilitate maintenance and housekeeping	Muther (1973), Apple (1977), Tompkins et al. (2003)
- Consider needs of workers with disabilities	Cedarleaf (1994)
Operational objectives	
- Provide high WIP turnover	Moore (1962), Apple (1963, 1977), Wrennal (2001)

Table 5.1 (continued).

- ***Optimize flow (materials, information, and personnel)***	Muther (1973), Heragu (1997), Apple (1977), Cedarleaf (1994)
- ***Optimize handling (e.g., minimize cost of materials handling)***	Moore (1962), Apple (1963, 1977), Muther (1973), Heragu (1997), Cedarleaf (1994), Tompkins et al. (2003)
- Promote safety and security of materials, equipment and employees	Moore (1962), Apple (1963, 1977), Muther (1973), Sheth (1995), Heragu (1997), Tompkins et al. (2003)
- Provide convenience for workers and promote job satisfaction	Moore (1962), Apple (1963), Muther (1973), Heragu (1997), Tompkins et al. (2003)
- ***Optimize use of equipment***	Muther (1973), Apple (1977), Tompkins et al. (2003)
- Stimulate optimal workforce utilization	Moore (1962), Apple (1963, 1977), Tompkins et al. (2003)

ponent, and since the efficiency component is related to the objectives element of the macro design process, these components are not considered in Table 5.2. There are other characteristics that are not used in our analyses. Table 5.3 lists those characteristics and the reasons for excluding them in the analyses of the following section.

4. The macro design problem

In the literature, numerous processors using OR tools have been proposed for the macro design problem. To establish the degree of performance of their processor, authors use criteria such as CPU time, closeness to the optimum, and materials handling costs. As stated by Levary and Kalchik (1985), an analysis or comparison of the facilities layout processors should not be limited to these criteria. In a particular industry, a generally low number of facilities layout reviews combined with the present computer capacity make the CPU time criterion irrelevant. In addition, the range of CPU times found in the literature (a few minutes to a few hours) is not significant for a real-life problem. The closeness to the optimum criterion is not applicable in industry where the majority of layout projects are layout improvements. The materials handling costs should not be the only criterion for selecting a facilities layout. Other criteria must be considered, such as the department shapes, the type of layout, the qualitative relationships, the flow orientation, and so on.

Keeping in mind the objective of the evaluation of a processor's applicability to a real-life problem, a different perspective of analysis for the macro design problem is proposed. For this purpose, the character-

Table 5.2. Input parameters and variables used in models and methods for the facilities layout problem and the corresponding problem characteristics

	Problem characteristics
Materials Volume of production Frequency Sources Destinations Unit load	- probabilistic nature - evaluation nature - time nature
Activities Equipment required Capacity requirements Possible alternatives Space requirements (size, shape, type, characteristics) Adjacency restrictions	- perimeter sizing - department size restrictions (e.g., aspect ratio, upper and lower bounds) - department location restrictions
Physical arrangement Distance Cross-traffic Location of receiving & shipping General linear flow Type (e.g., products layout, process layout) Equipment or department location Desired location of production services areas Location of utilities and auxiliaries Building size and shape Docks and doors - number, opening, size, location, height Floors - numbers, condition, load capacity, type of flooring, resistance Possible use of mezzanines, balconies, basement, roof Space availability and characteristics Elevators, ramps Loading and unloading facilities Aisle requirements - quantity, type, location, width Aisle congestion First aid facilities Desired location of personnel services areas Supervisory requirements	- aisle network - materials handling equipment selection - building shape - department shape - number of floors

istics identified in Section 3 are used. For each one, complexity levels are defined in Table 5.4.

Several authors have proposed a classification of the models and methods. Kusiak and Heragu (1987) have established four classes: construc-

Table 5.3. Characteristics not used in the analyses of OR processors

Characteristics	Reasons for exclusion
Materials	
Direction of interdepartmental flow	Processors in the macro design literature use undirected flow.
Activities	
Alternative product routings	Processors in the macro design literature consider only one routing per product.
Department orientation restrictions	Using bounds on the width and length of a department, this indicator can be included in the department size restrictions.
Physical arrangement	
Origin/Destination points of distance measure	If the processor generates the location of input/output (I/O) stations, it will use them for the measure of distance; if not, most processors in the literature use a measure between centroids.
Distance measure	Since the type of measure is related to the generation of aisles and the location of I/O stations, this indicator is redundant.
Aisle congestion	Processors in the macro design literature do not yet consider aisle congestion.
Layout generation nature	Using a simple 2-way exchange procedure, any construction processor can be transformed to an improvement processor.

tion algorithms, improvement algorithms, hybrid algorithms and graph-theoretic heuristics. Based on their work, Raoot and Rakshit (1991) and Delmaire et al. (1995b) present another classification. Meller and Gau (1996) establish their own classification based on three main areas of research: block layout (covered by Kusiak and Heragu, 1987), extensions, and other types of layout. In order to analyze the processors in terms of the predefined characteristics, we prefer to use a classification scheme based on the solution procedure instead of the application domain. Our classification scheme of OR processors for the macro design problem is:

(1) Exact optimization methods
(2) Heuristic methods, including:
    (a) Iterative heuristics
        (i) Simulated annealing algorithms
        (ii) Tabu search algorithms
        (iii) Genetic algorithms
    (b) Artificial intelligence heuristics

Table 5.4. Complexity levels per characteristic

Characteristics	**Complexity**			
	level 1	**level 2**	**level 3**	**level 4**
Materials				
1- probabilistic nature	deterministic			stochastic
2- evaluation nature	quantitative	qualitative	qual. and quant. with math. terms	qual. and quant. with fuzzy sets
3- time nature	static			dynamic
Activities				
4- perimeter sizing	one-unit size	fixed	variable—grid	variable—continuous
5- dept. size restrictions	no			yes
6- dept. location restrictions	no			yes
Physical arrangement				
7- aisle network	block layout		predetermined[a]	net layout
8- materials handling equipment selection	no			yes
9- building shape	without constraints			with constraints
10- department shape	irregular			regular
11- number of floors	singie-floor			multi-floor

[a] e.g., spine, O-shaped, L-shaped

(c) Miscellaneous heuristics

(3) Graph-theoretic methods

The *exact optimization methods* correspond to (mixed) integer linear programming models. The solution of those models is based essentially on the Branch & Bound technique, which is an *implicit* enumeration of all feasible solutions which allows finding the optimal one.

Among heuristic methods, *iterative heuristics* and *artificial intelligence heuristics* correspond to the most recent trends in research. The iterative heuristics present algorithms generating a solution or a set of solutions at each iteration. The subsequent iteration is based on the set of solutions generated at the previous iteration. The procedure can generate at a given iteration a solution of worse quality than at the previous one, but the final solution should not be affected by a local optima. The iterative heuristics are quite insensitive to the starting solution, which is not the case for other types of heuristics. Artificial intelligence is used for solving problems where the uncertainty of the initial data is expressed as probabilities or value intervals. The artificial intelligence heuristics, including expert systems and fuzzy set theory, allow a multi-criteria evaluation, both for quantitative and qualitative criteria. The decision rules for layout design are established by experts knowledgeable about all the parameters of the specific setting and especially about the decision rules of the company. *Miscellaneous heuristics* refers to author-specific processors for selecting and locating departments.

Finally, the *Graph-theoretic methods* use a graph composed of nodes and edges. There are two lines of research, one based on adjacency graphs and the other on cut trees. In the first case, the dual of the adjacency graph facility is associated with the layout. In the second case, the edges of a cut tree can easily correspond to the aisle network.

The mathematical model most frequently associated with the plant layout problem is the quadratic assignment problem (QAP). This problem consists in assigning n activities to n potential sites in order to optimize an objective function such as the total distance or total cost of materials handling, operating variable costs, or deliveries costs (Akinc, 1985). Those types of processors use quantitative criteria only. However, some processors evade this constraint by transposing qualitative values, e.g., the closeness relationships, into quantitative terms.

Urban (1987) defines the quadratic assignment problem as follows:

$$\text{MIN} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} c(i,j,k,l) * x(i,k) * x(j,l)$$

subject to

$$\sum_{j=1}^{n} x(i,j) = 1 \qquad \text{for } i = 1,\ldots,n$$

$$\sum_{i=1}^{n} x(i,j) = 1 \qquad \text{for } j = 1,\ldots,n$$

$$x(i,j) \text{ binary} \qquad \text{for } i = 1,\ldots n,\ j = 1,\ldots,n,$$

where

$c(i,j,k,l)$: cost for assigning activities i and $j$to sites k and l respectively
$x(i,k)$: equals 1 if activity i is assigned to site k, 0 otherwise.

The next subsections present the state of the art of research for each category using the characteristics and complexity levels listed in Table 5.4. This review does not intend to present all the details of each processor. As previously stated, the main objective is the evaluation of applicability of categories of processors to real-life facilities layout problems.

4.1 Exact optimization methods

The use of this category of processors allows the modeling of several important characteristics such as a regular shape of departments, location of the I/O stations (one of the first processors dedicated to this problem is by Montreuil and Ratliff, 1988), the evaluation of the real distances, or a layout that includes aisles. Distinct input and output stations allow considering directed flows (e.g., Barbosa-Povoa et al., 2001, 2002). Most processors consider constraints on department dimensions (e.g., lower and upper bounds on the length, width, perimeter or area, aspect ratio), or on their location (zoning constraints). Montreuil et al. (2002a) include also an a priori selection of materials handling equipment, which has an impact on the distance calculation, e.g., a rectilinear distance on the floor and a Euclidean distance for overhead handling. Their processor is aimed at translating a block layout into a net layout, i.e., with an aisle network and I/O stations. On the other hand, the use of such processors may be difficult for someone without a strong mathematical background. The results are always optimal and the processors of this category generate only one layout; this can be a drawback when manual fine-tuning is necessary to take into account characteristics not considered by the processor.

Several processors are associated with the QAP. The first processors for the QAP considered departments of one-unit size (e.g., Gavett and Plyter, 1966; Rosenblatt, 1986; Rosenblatt and Kropp, 1992). Processors considering fixed dimensions are mainly associated with the machine layout problem (e.g., Love and Wong, 1976; Kim and Kim, 2000; Barbosa-Povoa et al., 2001). In this context, parts routing is also considered. More complex processors use variable dimensions on a grid or continuous dimensions. Based on their previous work, Barbosa-Povoa et al. (2002) complicate their processor by considering three-dimensional characteristics: a three-dimensional department size and a multi-floor setting. Also, the probabilistic nature of the macro design problem is addressed by several authors, e.g., Rosenblatt and Kropp (1992), Montreuil and Laforge (1992), McKendall et al. (1999), and Benjaafar and Sheikhzadeh (2000).

Authors like Malakooti and D'Souza (1987), Meller and Bozer (1997) (for the multi-floor layout problem), Lacksonen (1994, 1997), and Kim and Kim (2000) propose hybrid-type processors, that is, heuristics for the macro design problem, which may include an exact optimization method. Another example of a hybrid-type processor due to Montreuil et al. (1989), combines linear programming and graph theory. Their main idea is to generate an aisle network skeleton using graph theory in which the inter-department links correspond to aisle segments. From this skeleton, a mathematical program is used to design the final layout that includes the aisles. Another hybrid processor (Ho and Moodie, 2000) combines a heuristic for generating a block layout and linear programming to obtain the net layout.

A total of 34 processors have been found: Gavett and Plyter (1966), Bazaraa (1975), Love and Wong (1976), Picard and Queyranne (1981), Rosenblatt (1986), Malakooti and D'Souza (1987), Montreuil and Ratliff (1988), Montreuil and Venkatadri (1988), Montreuil et al. (1989, 2002a,b), Wang and Wong (1990), Heragu and Kusiak (1991), Van Camp et al. (1992), Butler et al. (1992), Heragu (1992), Ketcham (1992), Kouvelis et al. (1992b), Rosenblatt and Kropp (1992), Montreuil and Laforge (1992), Houshyar and White (1993), Lacksonen (1994, 1997), Banerjee et al. (1997), Meller and Bozer (1997), Urban (1998), McKendall et al. (1999), Benjafaar and Sheikhzadeh (2000), Ho and Moodie (2000), Kim and Kim (2000), Barbosa-Povoa et al. (2001, 2002), Anjos and Vannelli (2002), and Castillo and Peters (2003).

In Figure 5.4, a *radar* graph is used to present a synthetic view of the progress in research according to each characteristic and in terms of levels of complexity. The graph is defined by eleven axes, one for each characteristic, and a scale of values from 1 to 4 according to the complexity level. As previously defined, the eleven characteristics depicted are the probabilistic nature (prob), the evaluation nature (eval), the time nature (time), the perimeter sizing (perim), the department size restrictions (size), the department location restrictions (loc), the aisle network (network), the materials handling equipment selection (mh), the building shape (b.shape), the department shape (d.shape), and the number of floors (floor). On the graph, the grey zone represents the levels of complexity for the majority of processors found for this category. We have also included in the graph examples of authors distinguished from the others by the level of complexity of some characteristics. Note that these are not necessarily the most recent ones. A *radar* graph is presented for each category.

It can be observed from this figure that:

- The perimeter characteristic (perim) tends to be continuous rather than on a grid. As a result, department shapes (d.shape) are more regular. Heragu and Kusiak (1991), Van Camp et al. (1992), and Montreuil and Laforge (1992) are among the first authors considering this level of complexity.

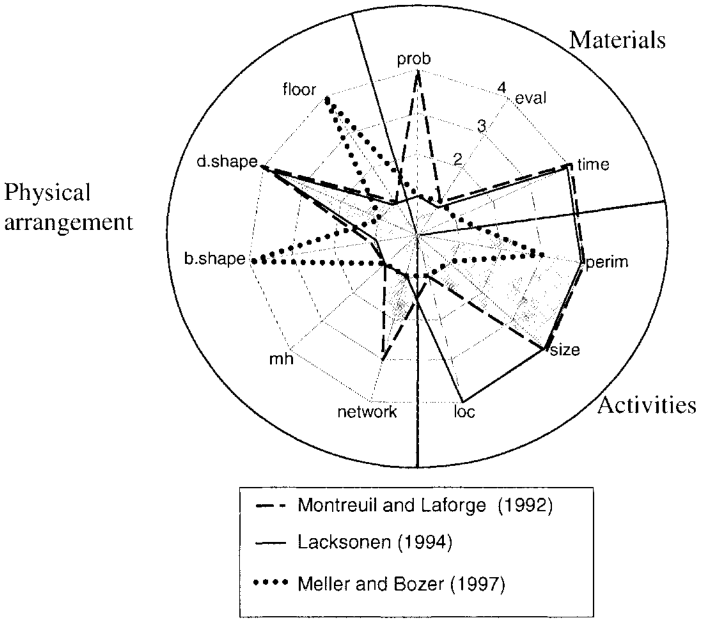


Figure 5.4. Complexity levels of characteristics for the exact optimization methods

- Constraints related to the department size characteristic (size), such as upper and lower bounds and aspect ratio, are taken into account more and more, e.g., Montreuil and Venkatadri (1988), Montreuil et al. (1989, 2002a,b), Montreuil and Laforge (1992), Lacksonen (1997), McKendall et al. (1999), Barbosa-Povoa et al. (2001), and Anjos and Vannelli (2002).
- For the aisle network characteristic (network), some authors address the spine layout (Love and Wong, 1976; Picard and Queyranne, 1981; Heragu and Kusiak, 1991), or a predetermined aisle skeleton (Montreuil and Laforge, 1992). Despite this work, the type of layout usually considered is a block layout.
- For the number of floors characteristic (floor), the few processors found in the literature address the problem in two steps: 1. group the departments in clusters (one for each floor); 2. for each cluster, locate the departments.

4.2 Simulated annealing algorithms

Simulated annealing (SA) is used to improve an existing layout. It is an iterative heuristic. At each iteration, a layout is generated and the objective function evaluated. The new layout is kept or discarded based

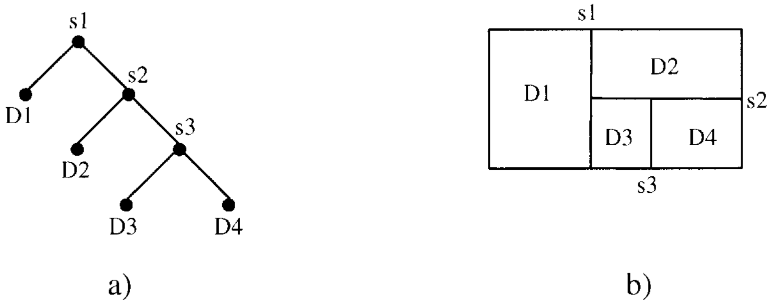


Figure 5.5. Example of a slicing tree and the corresponding block layout

on a number of criteria. The algorithm stops when there is no more possible improvements in the objective function. The QAP has been largely addressed with SA, e.g., Kouvelis and Chiang (1992), Kouvelis et al. (1992a), Shang (1993), Chiang and Chiang (1998), and Castillo and Peters (2002). However, the very nature of this type of heuristic leads to the analysis of more complex problems with features such as variable department size and restrictions on location.

Even though Kirkpatrick et al. (1983) were the first authors to apply SA to the facilities layout problem, the use of SA for the macro design problem grew considerably in the 1990's, with authors such as Meller and Bozer (1991) and Jajodia et al. (1992). Few characteristics are however considered. Block layouts are obtained and the probabilistic issue is not considered. To our knowledge, Baykasoglu and Gindy (2001) are the first to propose a SA-based processor for the dynamic layout problem. Chiang and Chiang (1998) propose a combination of a SA algorithm (for generating a facilities layout) and a Tabu search algorithm (for evaluating the solution).

Wu and Appleton (2002a) use slicing trees to represent the departments' arrangement. Figure 5.5 depicts a slicing tree and the related block layout where each cut (s1, s2, s3) corresponds to a branching node in the slicing tree. There are vertical and horizontal cuts and the zones defined by the cuts correspond to the department locations (D1, D2, D3, D4). The slicing lines can be the basis of an aisle structure.

A total of 17 processors fall in this category: Kirkpatrick et al. (1983), Meller and Bozer (1991), Heragu and Alfa (1992), Jajodia et al. (1992), Kouvelis and Chiang (1992), Kouvelis et al. (1992a), Tam (1992a), Shang (1993), Lin et al. (1994), Chiang and Chiang (1998), Chwif et al. (1998), Kim and Kim (1998), Matsuzaki et al. (1999), Mir and Imam (2000), Baykasoglu and Gindy (2001), Castillo and Peters (2002), and Wu and

Appleton (2002a). Figure 5.6 presents the most frequent characteristics (grey zone) for the simulated annealing algorithms.

One can observe from Figure 5.6:

- For the perimeter size characteristic (perim), Tam (1992a), Kim and Kim (1998) and Matsuzaki et al. (1999) are among authors using a continuous representation of departments, but even today, some authors, such as Mir and Imam (2000) are still addressing the fixed size case.
- For the department size characteristic (size), the aspect ratio and the lower and upper bounds on a department area are constraints used by some authors, such as Chwif et al. (1998), Kim and Kim (1998), Matsuzaki et al. (1999), and Castillo and Peters (2002).
- For the department location restrictions characteristic (loc), some authors (Kouvelis et al., 1992a; and Tam, 1992a) use zoning constraints. However, most authors do not consider those constraints.
- For the number of floors characteristic (floor), SABLE, from Meller and Bozer (1991), which is a variant of their previous MULTIPLE (Bozer et al., 1994) — which in turn is based on CRAFT — addresses the single- and the multi-floor facilities layout problem. Meller and

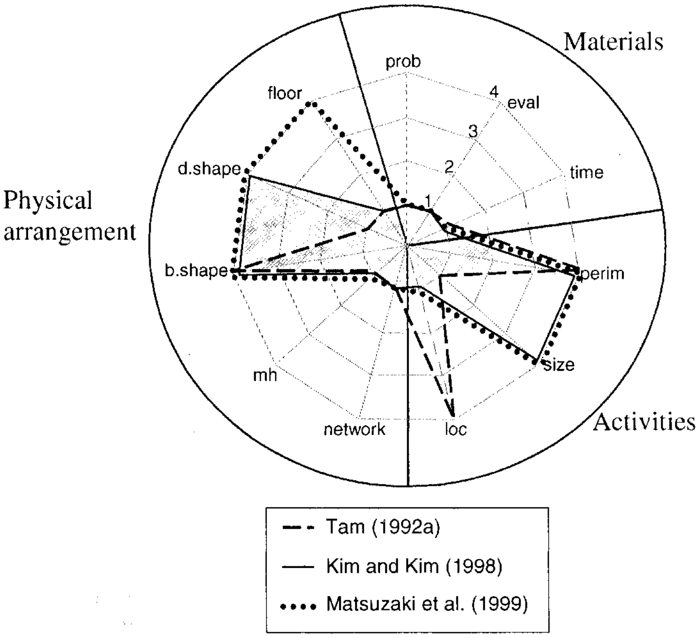


Figure 5.6. Complexity levels of characteristics for the simulated annealing algorithms

Bozer (1997) present an extension of their multi-floor processor defined in two steps: an exact optimization method and a SA-based step. It minimizes the total distance instead of the distance per floor. A real-life problem is used for the performance evaluation of their processor. Based on their work, the Matsuzaki et al. (1999) processor is a combination of simulated annealing for generating a facilities layout and a genetic algorithm for the optimization of the elevators in terms of number and location, and the assignment of flow to the elevators.

4.3 Tabu search algorithms

Tabu search (TS), introduced by Hansen (1986) and Glover (1989), is an iterative heuristic that allows deterioration of a solution in order to escape from a local optimum. The difference between SA and TS is that the latter restrains the number of intermediate solutions. A list of previous solutions is generated and restricts their selection as a new solution for a number of iterations. Skorin-Kapov (1990, 1991) applies the method to the QAP. Chiang and Kouvelis (1996) and Chiang and Chiang (1998) define a dynamic tabu list size, i.e., of variable length.

This research avenue is rather recent and TS as well as SA have been used on facilities layout problems of low complexity. Very few actual setting constraints are taken into account. The generated layouts are block layouts. Even though some authors, such as Chittratanawat and Noble (1999), Abdinnour-Helm and Hadley (2000), and Chiang (2001) use a representation of departments other than of one-unit size, the majority of processors consider the one-unit size representation. A few authors such as Chiang and Kouvelis (1996) present a processor for the dynamic context but their processor is still for the QAP. The probabilistic aspect has not yet been considered.

Chittratanawat and Noble (1999) is one of the very few works on the macro design problem addressing the materials handling equipment selection. They include qualitative relations, expressed in numeric terms, and restrictions related to the location of I/O stations.

Nine processors have been found: Hansen (1986), Glover (1989), Skorin-Kapov (1990, 1991), Chiang and Kouvelis (1996), Chiang and Chiang (1998), Chittratanawat and Noble (1999), Abdinnour-Helm and Hadley (2000), and Chiang (2001). Figure 5.7 summarizes the most frequent characteristics (grey zone) for the tabu search algorithms.

The following observations can be made:

- For the department size characteristic (size), the lower and upper bounds and the aspect ratio are the only constraints used by some authors, such as Abdinnour-Helm and Hadley (2000) and Chiang (2001).

- The shape characteristics (b.shape and d.shape) are directly related to the perimeter size characteristics which are mainly of one-unit size.
- For the number of floors characteristic (floors), the processor of Abdinnour-Helm and Hadley (2000) is based on MULTIPLE of Bozer et al. (1994).

4.4 Genetic algorithms

Proposed by Holland (1975), genetic algorithms (GA) use the genetic operators: reproduction, selection and mutation. These operators are applied to a set (population) of strings (solutions). It is an iterative heuristic. As opposed to SA and TS, several regions of the solution domain can be explored simultaneously. Michalewicz (1992) applies this approach to the macro design problem.

Research using this method has evolved rapidly towards a variable perimeter sizing, actually mainly on a grid. A few processors take into consideration a fixed aisle skeleton, for example Delmaire et al. (1997) for the spine layout, the T-shaped, and the O-shaped aisle network. Few others generate their own aisle network (Banerjee et al., 1992b, 1997, and Tavakkoli-Maghaddain and Shayan, 1998). The processor of Delmaire et al. (1997) uses a genetic algorithm where each string corresponds to a

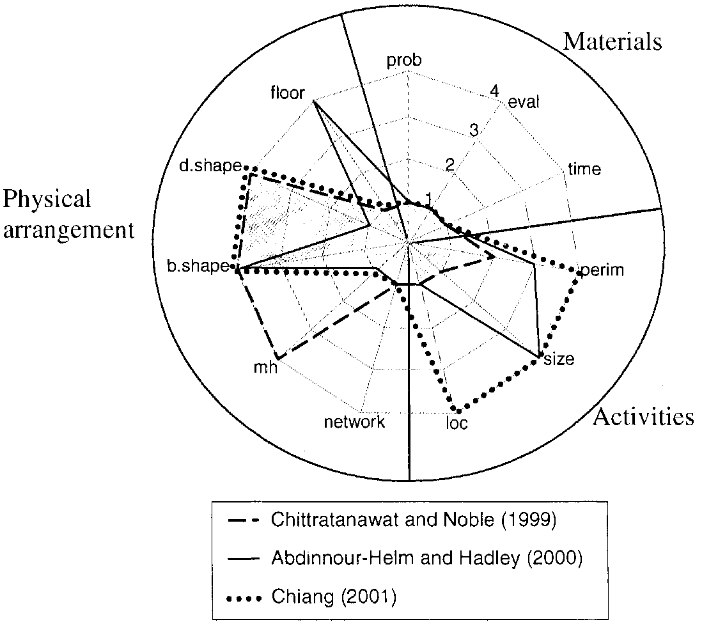


Figure 5.7. Complexity levels of characteristics for the tabu search algorithms

relative arrangement of the departments and the evaluation of the string is obtained through a linear programming model giving the optimal department dimensions and I/O stations location. Rao et al. (1999) link a genetic algorithm with AutoCAD. Physical limitations and restrictions on location or orientation are considered in generating a final layout. Hamamoto et al. (1999) use a simulation software (for the evaluation stage of the facilities layout) with a genetic algorithm for locating the departments. They consider a double objective: minimizing the duration of handling and maximizing the throughput rate.

In the continuous case for department dimensions, the slicing tree technique, previously defined, is used by Al-Hakim (2000), Azadivar and Wang (2000), and Wu and Appleton (2002b). A related formulation is the bay structure where transversal cuts create bays which are then subdivided into departments by perpendicular cuts. Two borders for each department are defined either by transversal cuts or by one transversal cut and the external border of the layout. This technique is derived from the well-known ALDEP and is used by Norman and Smith (1999), Gómez et al. (2003), Lee et al. (2003), and Ozdemir et al. (2003). Unlike ALDEP, these processors consider bays with variable widths. The processor of Lee et al. (2003) takes as inputs the location of main aisles and inner walls structure, which respectively correspond to transversal cuts and cuts within a bay.

Norman and Smith (1999) consider the probabilistic nature of the facilities layout problem using the demand standard deviation. The final layout is a block layout. Azadivar and Wang (2000) include an evaluation of the solutions by simulation.

Another recent research avenue addresses the management of directed flow in a machine loop-layout problem. The objective function of such processors, e.g., Cheng et al. (1996), Cheng and Gen (1998), and Rajasekharan et al. (1998) is related to aisle congestion.

In the last few years, the trend in research on the macro design problem is to combine the various approaches. For example, Banerjee et al. (1997) present an iterative method using a genetic algorithm, an LP-solver, and a graph-theoretic method. We refer the interested reader to the review on genetic algorithm processors by Pierreval et al. (2003) of articles published from 1995 to 2000.

A total of 36 processors have been proposed: Holland (1975), Michalewicz (1992), Banerjee et al. (1992b, 1997), Tam (1992b), Lin et al. (1994), Suresh et al. (1995), Tate and Smith (1995), Delmaire et al. (1995a, 1997), Cheng et al. (1996), Castell et al. (1998), Cheng and Gen (1998), Islier (1998), Kochhar and Heragu (1998, 1999), Kochhar et al. (1998), Mak et al. (1998), Rajasekharan et al. (1998), Tam and Chan (1998), Tavakkoli-Moghaddain and Shayan (1998), Gau and Meller (1999), Hamamoto et al. (1999), Norman and Smith (1999), Rao et al. (1999), Al-Hakim (2000), Azadivar and Wang (2000), Balakrishnan and Cheng (2000), Li and Love (2000), Norman et al. (2001), Lee and Lee (2002), Wu and Appleton (2002b), Balakrishnan et al. (2003a), Gómez et al. (2003), Lee et al. (2003), and Ozdemir et al. (2003). Figure 5.8 summarizes the characteristics for the majority of genetic algorithms (grey zone).

The following observations can be made:

- Kochhar and Heragu (1999), Balakrishnan and Cheng (2000), and Balakrishnan et al. (2003a) are among the few authors considering the dynamic nature (time). This characteristic is usually combined with a simplified perimeter sizing characteristic. Those authors, in fact, address the QAP or the modified QAP (i.e., departments with unequal sizes on a grid).
- For the perimeter sizing characteristic (perim), most processors work on a grid. Recent processors tend to use continuous dimensions of departments, e.g., Kochhar and Heragu (1999), Gau and Meller (1999), and Norman et al. (2001).

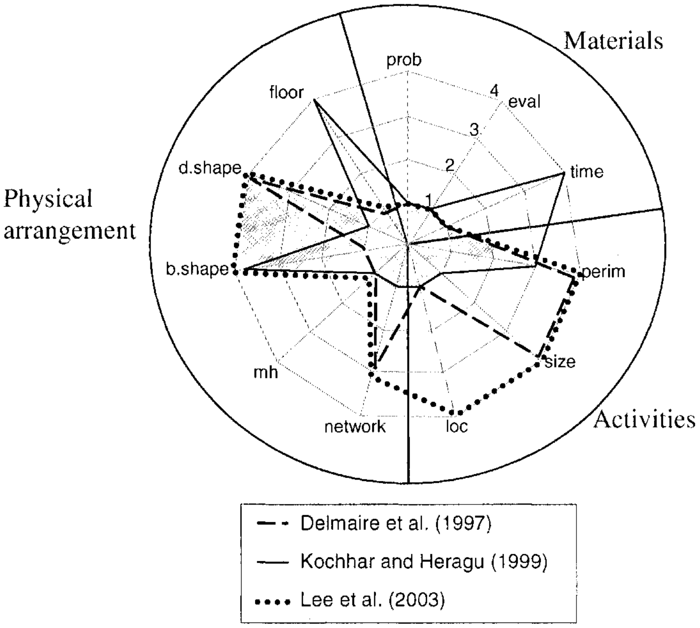


Figure 5.8. Complexity levels of characteristics for the genetic algorithms

- For the department sizing characteristic (size), Banerjee et al. (1997) use upper and lower bounds for the departments' perimeter, length, and width. Also, Tate and Smith (1995) use a set of possible shapes of departments. Tam and Chan (1998), Norman and Smith (1999), and Lee et al. (2003) use the aspect ratio as a constraint for their processor. However, most processors do not address this characteristic.
- For the aisle network characteristic (network), some processors use a fixed aisle skeleton (Delmaire et al., 1997) and some others generate a network (Banerjee et al., 1992b, 1997; Tavakkoli-Maghaddain and Shayan, 1998).
- For the number of floors characteristic (floor), Kochhar and Heragu (1998) use fixed elevator locations. Departments can be of various dimensions and a department cannot be assigned to more than one floor. An extension (DHOPE, by Kochhar and Heragu, 1999) addresses the dynamic aspect of the layout problem by considering two periods. This algorithm generates irregular departments and building shapes.

4.5 Artificial intelligence heuristics

The main advantage of artificial intelligence (AI) is the evaluation of the various quantitative and qualitative parameters, such as materials, personnel and information flow, need for proximity, processes interference, and supervision. Every parameter is graded as very important, average, or weak. For every combination of activities, the parameters are evaluated. Then, decision rules are applied, according to the proposed algorithms, and a value is assigned to each combination of activities. Most authors use distance or a closeness rating scale as defined by Muther (1973). Finally, a selection and location heuristic is used for the layout design. A characteristic of AI is its applicability in a stochastic context within a static environment. The consequences of the lack of temporal data and of the variability of these are thus minimized.

Expert systems and fuzzy sets have been applied quite recently to the macro design problem. The first works date from the end of the 1980's, e.g., Grobelny (1987, 1988), Kumara et al. (1987, 1988), Evans et al. (1987), and Malakooti and Tsurushima (1989). Grobelny (1987 and 1988) addresses respectively the QAP and the MLP (machine layout problem), the latter including fixed dimensions and activities with orientation constraints. In Kumara et al. (1988) the departments have unequal sizes and the solution method is simple: 1. find the department areas' common denominator, 2. divide each area in blocks of equal size. All blocks of a department must have a strong artificial relation in order to be adjacent. Malakooti and Tsurushima (1989) use a method proposed by Kumara et al. (1988) in the context of a modified QAP. They add a priority factor to the decision rules. Several layouts can be generated with different sets of priorities. This constitutes a beginning of interaction for the industrial engineer in layout analysis and selection. On the other hand, Evans et al. (1987) consider departments of unequal sizes on a grid, which is more realistic for the macro design problem.

The well known expert system processors of Abdou and Dutta (1990) and Heragu and Kusiak (1990) were developed for machine layout. Decision rules pertain to the materials handling equipment selection. An aisle skeleton, determined according to each possible materials handling equipment, is then generated by the processor. This method assumes that for each type of handling equipment there is an ideal aisle network.

A specific feature of the processor of Raoot and Rakshit (1993) — based on fuzzy set theory — is to generate multiple solutions. Those solutions are then sorted according to the objective value. The industrial engineer can hence select the final layout among this set of solutions. Badiru and Arif (1996) present a 3-step method. The first step use FLEXPERT, a fuzzy set algorithm to evaluate the department quotations based on inter-department relations. In the second step, several layouts are generated using one of the heuristics from the literature. In the last step, the layouts are evaluated and a final selection is made. Dweiri and Meier (1996) present a similar processor. They use a modified version of CORELAP at the layout generation step. Yang and Kuo (2003) also use heuristics to generate layouts. Shape ratio constraints are added to the processor.

Deb and Bhattacharyya (2003) propose a processor for a machine layout problem where the machine dimensions and the location of input and output stations are given a priori. The separate management of the two types of stations allows consideration of directed flows.

A rather new avenue of research is the use of neural networks. The characteristics of Tsuchiya et al. (1996) are classical, i.e., a quantitative evaluation (distance and cost) in a QAP context, and generation of a block layout.

A total of 21 processors using AI has been surveyed: Evans et al. (1987), Grobelny (1987, 1988), Kumara et al. (1987, 1988), Malakooti and Tsurushima (1989), Abdou and Dutta (1990), Heragu and Kusiak (1990), Cambron and Evans (1991), Banerjee et al. (1992a), Shih et al. (1992), Raoot and Rakshit (1991, 1993), Sirinaovakul and Thajchayapong (1994), Badiru and Arif (1996), Dweiri and Meier (1996), Tsuchiya et al. (1996), Dweiri (1999), Aiello and Enea (2001), Deb and Bhattacharyya (2003), and Yang and Kuo (2003). Figure 5.9 presents the most frequent characteristics (grey zone) for AI heuristics.

We observe that:

- For the probabilistic nature (prob), most processors consider a deterministic environment. Despite the appropriateness of AI for stochastic issues, only a few authors have addressed them (Abdou and Dutta, 1990; Badiru and Arif, 1996; Dweiri and Meier, 1996; and Aiello and Enea, 2001).
- For the perimeter sizing characteristic (perim), most processors use a grid. A few processors use a continuous representation of departments, e.g., Badiru and Arif (1996) and Dweiri (1999).
- For the materials handling equipment selection characteristic (mh), Heragu and Kusiak (1990) and Abdou and Dutta (1990) are among the few to include the selection in their processors.

4.6 Miscellaneous heuristics

This category includes all the heuristics not covered in the previous sections. Developed in the sixties, CRAFT, CORELAP, ALDEP, PLANET, and COFAD are among the first processors proposed. In particular, CRAFT is the keystone for numerous processors developed over the years. Also, the methodology of analysis of Muther (1973) for the design of a plant layout is still in use. The computerization of Muther's

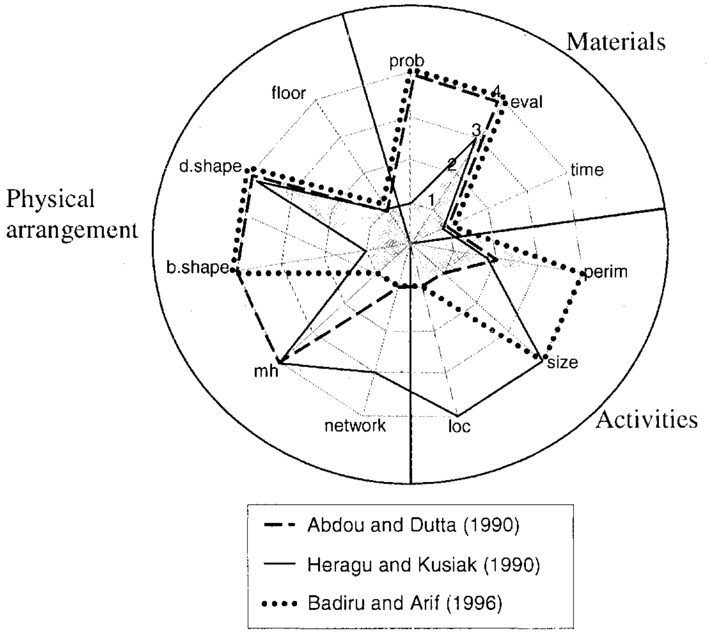


Figure 5.9. Complexity levels of characteristics for the artificial intelligence heuristics

method started with the work of O'Brien and Abdel Barr (1980). MICROLAY (Wäscher and Chamoni, 1987) allows for the design of a new layout or for starting from an existing layout. Then, each improvement step (using the well known CRAFT method) is validated by an industrial engineer. MOCRAFT (Svestka, 1990) is a variant of CRAFT with a multi-criteria objective. Here again, it is possible to intervene in the generation process for both the intermediate layouts and the final one. Given the characteristics addressed by most processors, layout review by an industrial engineer is still necessary for verifying the realism of the generated solutions.

Processors of this category allow more flexibility of taking into account specific characteristics, e.g., the processor of Johnson (1982) with the location of building columns. However, most processors generate a single final layout. Some processors use the two-way or three-way pair exchange as improvement procedure. A new generation of processors combines one or various heuristic methods for the generation of layouts, e.g., the processor of Balakrishnan et al. (2003b) uses 2 heuristic methods: one SA-based and the other GA-based. However, a single layout is obtained.

The emphasis of the research in this category is on the department dimensions (perim and size) and on a pre-selection of the aisle network (network). The stochastic nature (prob) of data is addressed in very few works. Rosenblatt and Lee (1987) model the block layout design problem with variable demands. A probability is assigned to each value of the demand. Their main objective is to evaluate the robustness (i.e., the flexibility) of the layout generated by their method. Yaman et al. (1993) take into account the variation of the quantities to be manufactured. The departments are of equal sizes and their location is done with a space filling curve method.

In a dynamic context, the Balakrishnan et al. (2000) processors are based on Urban (1993), which in turn is based on CRAFT. Among others, the characteristic of the first processor is the use of a backward method: a backward pass (from period t to period $t-1$) is performed on each layout plan generated from a forward pass (from period t to period $t+1$).

Tompkins and Reed (1976) and Welgama and Gibson (1996) are among the few authors considering the materials handling equipment selection (mh). The number of floors (floor) is also rarely considered. SPACECRAFT by Johnson (1982), based on CRAFT, includes a multi-floor feature with different speeds for vertical and horizontal movements. However, this processor permits locating a department on several floors and the departments can have irregular shapes. The processor of Kaku et al. (1988) clusters the departments into different cells. The objectives are to minimize the intercellular interactions and to maximize the intracellular interactions. Then, each cell is associated with a specific floor, and an overall layout for each floor is designed. MULTIPLE by Bozer et al. (1994) limits the departmental dimensions with lower and upper bounds. Exchanges between floors are allowed if there is sufficient available space on each floor. A space filling curve method is used with constraints on the department width.

As previously mentioned, four levels of complexity are defined for the evaluation nature characteristic (eval). The first level, and the most used for this category of processors, corresponds to the use of an objective function with quantitative data only. Decision variables represent generally the traveling distances between each pair of departments. One exception is FLAT (Kusiak and Heragu, 1987), which is based on adjusted flows for all triplets of departments. The second level corresponds to the use of qualitative data only, such as the closeness rating scale of Muther (1973). The third level corresponds to the use of a multi-criteria objective function, often weighted, with quantitative and qualitative data. A scale of values is used to convert the qualitative data into quantitative, e.g., Muther (1973), Rosenblatt (1979), Fortenberry and Cox (1985), Malakooti and D'Souza (1987), Urban (1987), Malakooti (1989), Svestka (1990), Wang et al. (1991), Houshyar (1991), Harmonosky and Tothero (1992), Partovi and Burton (1992), and Chen and Sha (1999). The processor of Chen and Sha (1999) permits standardization in the objective function of the quantitative and qualitative criteria. The last level is not encountered in this category of processors.

DISCON by Drezner (1987), uses scatter diagrams and considers department shapes other than rectangular. Departments are represented by circles with appropriate area. Then, adding rectangular bounds for each department permits generating several different final solutions. Safizadeh and McKenna (1996) propose a revised version of this processor.

A total of 43 processors fall into this category: Armour and Buffa (1963), Nugent et al. (1968), Khalil (1973), Muther (1973), Neghabat (1974), Tompkins and Reed (1976), Rosenblatt (1979), O'Brien and Abdel Barr (1980), Liggett (1981), Johnson (1982), Gaston (1984), Fortenberry and Cox (1985), Scriabin and Vergin (1985), Hassan et al. (1986), Drezner (1987), Kusiak and Heragu (1987), Jacobs (1987), Malakooti and D'Souza (1987), Rosenblatt and Lee (1987), Urban (1987, 1993), Wäscher and Chamoni (1987), Kaku et al. (1988), Malakooti (1989), Chhajed et al. (1990), Svestka (1990), Houshyar (1991), Wang et al. (1991), Harmonosky and Tothero (1992), Kaku and Rachamadugu (1992), Partovi and Burton (1992), Rosenblatt and Golany (1992), Das

(1993), Welgama and Gibson (1993, 1996), Yaman et al. (1993), Bozer et al. (1994), Langevin et al. (1994), Tretheway and Foote (1994), Safizadeh and McKenna (1996), Chen and Sha (1999), and Balakrishnan et al. (2000, 2003b). Figure 5.10 presents the most frequent characteristics (grey zone) for the miscellaneous heuristics.

The following observations can be made:

- For the evaluation nature (eval), more than half of the processors use a quantitative single criterion objective function. However, the use of quantitative and qualitative weighted multi-criteria objective functions is increasing.
- For the perimeter sizing characteristic (perim), most processors use a grid. A few processors use continuous dimensions of departments, e.g., Langevin et al. (1994) and Tretheway and Foote (1994).
- Constraints related to department sizing characteristic (size), such as upper and lower bounds and aspect ratio, are more and more addressed, e.g., Jacobs (1987), Das (1993), Bozer et al. (1994), and Welgama and Gibson (1996).
- For the department location restrictions characteristic (loc), only a few authors (Jacobs, 1987, Svestka, 1990, and Bozer et al., 1994) use

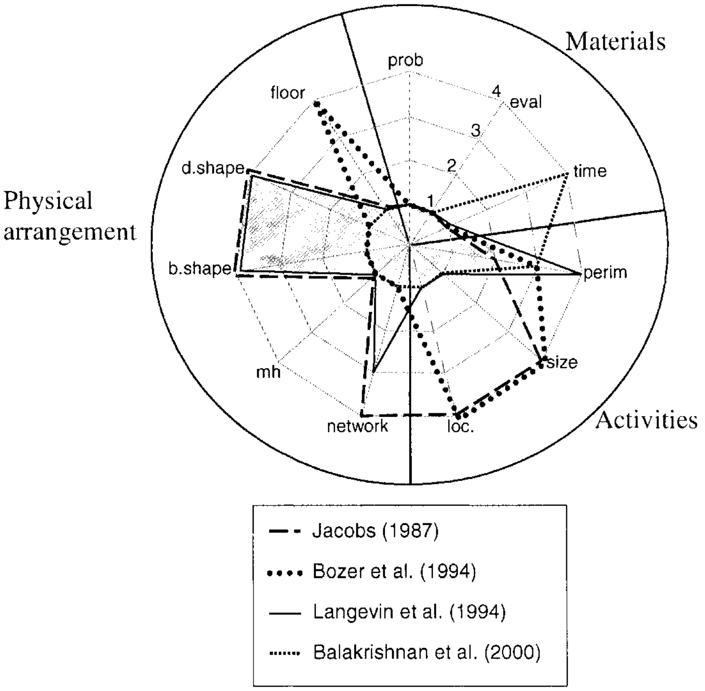


Figure 5.10. Complexity levels of characteristics for the miscellaneous heuristics

constraints such as forbidden zones and assignment to a specific location — or a specific floor for the multifloor layout problem — for a given department.

- For the aisle network characteristic (network), the processor of Langevin et al. (1994) is applicable to the spine layout. Their processor uses a list of pairs of departments ordered by the frequency of travel.

4.7 Graph-theoretic methods

Processors related to graph-theoretic methods (GA) are used for layout design without considering non-desirable proximity links between departments. All the methods of this category refer to a single-floor facilities layout problem and some can take into account department location restrictions, e.g., Montreuil et al. (1987), Montreuil and Ratliff (1989), Banerjee et al. (1990), Hassan and Hogg (1991), and Al-Hakim (1992). The methods are either based on adjacency graphs or on cut trees. The adjacency graphs based approach, as stated by Hassan and Hogg (1991), is defined by three steps leading to the generation of a block layout:

(1) construction of the adjacency graph, defined as a planar graph
(2) construction of the dual graph
(3) conversion of the dual graph into a facilities layout.

In an adjacency graph, a node corresponds to a department and an edge to the relation between two nodes (their adjacency being qualitative or quantitative and non-negative). The maximal planar weighted graph (MPWG) considers a maximum of $3n - 6$ interdepartmental relations where n is the number of departments. The dual graph allows representing the common boundaries of the departments. Therefore, each vertex represents a department, including a vertex corresponding to the building perimeter, and an edge, a common boundary (or wall). There are a number of processors integrating the three steps, e.g., Seppänen and Moore (1975), Hassan and Hogg (1991), and Goetschalckx (1992). However, the three steps are usually addressed separately, and step 1 and 3 are the most frequently found in the literature.

Numerous processors have been developed for step 1, e.g., TESSA (Boswell, 1992) and Goldschmidt et al. (1996). Some authors, such as Green and Al-Hakim (1985) and Giffin and Foulds (1987), consider qualitative relations translated into quantitative terms. Foulds and Giffin (1985) take into consideration materials handling costs. More recently, Pesch et al. (1999) include the evaluation of qualitative and negative data such as undesirable proximity relationship.

Step 3 is the most difficult one. Many layouts can be derived from a given dual graph. Hassan and Hogg (1989, 1991) insist on the fact that results have to be supervised, reviewed, and adjusted manually by the analyst. In the same vein, many other authors propose interactive processors, e.g., Hassan and Hogg (1989, 1991) and Irvine and Rinsma-Merchert (1997). The distinguishing feature of Irvine and Rinsma-Merchert (1997) is the use of predetermined shapes of departments.

For step 1, there is a total of 12 processors: Seppänen and Moore (1970), Foulds and Giffin (1985), Foulds et al. (1985), Green and Al-Hakim (1985), Giffin and Foulds (1987), Hassan and Hogg (1987), Al-Hakim (1991), Boswell (1992), Goldschmidt et al. (1996), Cimikowski and Mooney (1997), Wächer and Merker (1997), and Pesch et al. (1999). For step 3, there are a total of 7 processors: Montreuil et al. (1987), Hassan and Hogg (1989), Rinsma et al. (1990), Al-Hakim (1992), Leung (1992), Irvine and Rinsa-Melchert (1997), and Watson and Giffin (1997).

The difficulty in building planar graph leads to the use of cut trees, which are applicable for continuous representation of departments. With cut trees, only $n - 1$ interdepartmental relations are considered and a cut tree is easily related to a layout where the aisles are defined by the interdepartmental links. Seppänen and Moore (1975) are among the first to use cut trees for the facility layout problem. Based on the generated cut tree, edges are added to develop a maximal planar graph. Then, steps 2 and 3 are carried out. Montreuil and Ratliff (1989), Banerjee et al. (1990), and Kim et al. (1995) separate the cut tree technique from the maximal planar graph technique. Their methodology consists of two steps leading to the generation of a block layout with an aisle network:

(1) construction of a cut tree
(2) conversion into a facilities layout.

The mathematical aspect of the first step has been looked at frequently in the literature. However, this methodology has not been used much for facilities layout design. Banerjee et al. (1990) generate a cut tree and use linear programming to obtain a block layout. Kim et al. (1995) focus on generating the cut tree.

Figure 5.11 presents the most frequent characteristics for this category. The following observations can be made:

- For the evaluation nature (eval), most authors use a quantitative objective function. Qualitative criteria are translated into quantitative ones.
- The only department restrictions applied are related to their location (loc).

- For the aisle network characteristic (network), most authors that use adjacency graphs propose block layouts, even if graph theory would lead naturally to links defining an aisle network. With the cut tree approach, the processors include the definition of an aisle network.

4.8 Summary

This section presents for each characteristic a summary of the analyses of OR processors.

Materials: probabilistic nature. Most processors use deterministic data. A few exceptions can be found for the machine layout problem where the stochastic aspect is related to the product demands. This allows considering real data and evaluating the robustness (i.e., the flexibility) of the generated layout. As previously stated, despite the appropriateness of AI for stochastic issues, only a few authors have addressed this level of complexity for this characteristic.

Materials: evaluation nature. Most processors use quantitative input data. Actually, many categories of OR processors for the macro design problem rely only on quantitative data: the exact optimization methods, the iterative heuristics SA, TS, and GA, and graph-theoretic

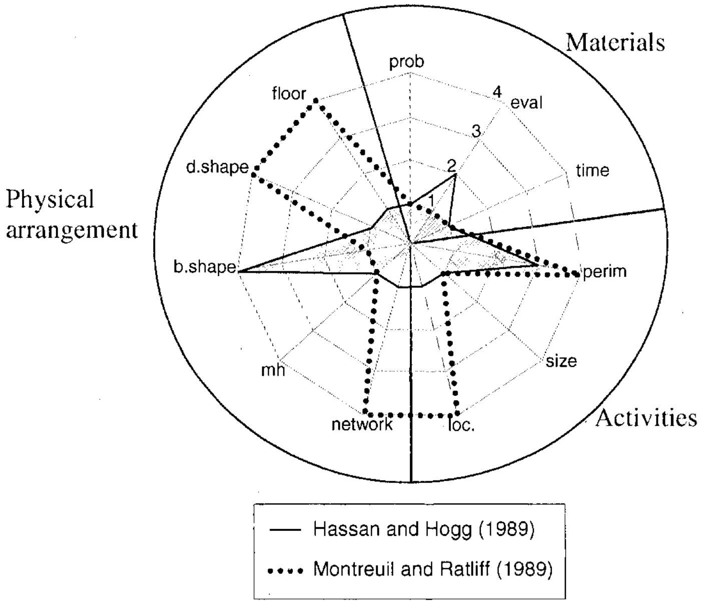


Figure 5.11. Complexity levels of characteristics for the graph-theoretic methods

methods. Many authors claim they are taking into account qualitative data, whereas these data are in fact translated into quantitative ones using a scale of values. The last level of complexity is related to AI, which combines a scale of values for each parameter with a set of decision rules.

In a multi-criteria context, authors usually use a weighted objective function. However, to our knowledge Montreuil and Ratliff (1988) are the only ones to discuss the need for sensitivity analysis. No processor includes a sensitivity analysis to find all the possible layouts that have little influence on the value of the objective function.

Materials: time nature. The static environment is the one most investigated. A dynamic approach is usually combined with a decrease in the complexity of another characteristic, often the perimeter sizing. Very few processors consider both a reactive analysis (layouts for periods $t = 1$ to n) — the case usually addressed — and a proactive analysis (layouts for periods $t = n$ to 1). The interested reader can find a literature review on this topic in Balakrishnan and Cheng (1998).

Activities: perimeter sizing. More than half of the processors address the QAP. Several authors have circumvented the difficulty of unequal departments sizes by using a layout grid and assigning one or many grid units to each department. A very large value is then given to the relations between all the grid units of a department. A fixed perimeter sizing is usually used for the machine layout problem. Generally, a continuous spatial representation of the departmental dimensions is combined with constraints pertaining to the characteristic of department size restrictions. For all the categories of processors presented in this literature review, the degree of complexity of this characteristic increases with time.

Activities: department size restrictions. Various restrictions for the department size have been considered. The most popular are: upper and lower bounds for the length of the sides of a department, for its perimeter, or for its surface area, and the aspect ratio (width/length). For the machine layout problem, a specific orientation for a machine is frequent. This can be translated into a restriction on the length of the machine sides.

Activities: department location restrictions. Restrictions on the location of some departments are taken into account by many processors, the most frequent being the assignment of a department to a specific location and the definition of fixed activities such as elevators and docks. Omitting such restrictions can have a significant impact on the final solution. Zoning constraints can also be used in the case of assignment to an area or a specific floor, or in the case of forbidden zones. For the first case, locating docks or offices on the building perimeter is a common requirement found in the manufacturing industry. For the second case, an activity with special needs, e.g., structure capacity or height, could be quite limited in its possible locations. It would thus be interesting to analyze the impact of this type of restriction and to include such flexibility in any processor.

Physical arrangement: aisle network. Apart from the GT category, most processors generate a block layout. Spine layout is also sometimes used. Works on spine layout have increased considerably. The QAP (with departments of equal dimensions and predefined potential sites) permits use of a predetermined aisle network. However this case is not addressed in the literature. For the GT category, by definition, the processors generate an aisle network skeleton, i.e., an aisle network without the width of segments. Few processors using exact optimization methods also generate an aisle network skeleton.

Physical arrangement: materials handling equipment selection. Only a few processors consider the materials handling equipment selection for the design of a new layout. For most methods, the handling equipment is determined at another step of the analysis, before or after the design of the layout.

Physical arrangement: building shape. Usually, the external shape of the building is given. For the QAP, the modified QAP (i.e., departments with unequal sizes on a grid), and for layouts with a continuous spatial representation of departments, a rectangular shape is assumed. To our knowledge, only one work considers a priori nonrectangular shape. Nevertheless, fixed dummy departments could be used to represent the anomalies of a building's shape.

Physical arrangement: department shape. Most processors generate departments with regular shape. However, the use of a grid requires a method of assigning the department units, e.g., a space filling curve method. This type of representation leads to irregular department shapes.

Physical arrangement: number of floors. More and more processors take into account more than one floor. The most frequent approach is the two-stage algorithm by Kochhar and Heragu (1998) which is the assignment of the departments to floors followed by the design of each floor by a single-floor facilities layout processor. Dummy activities correspond to fixed features such as elevators or staircases. Another approach consists of iteratively alternating between the two steps. Yet another one consists of assigning a specific location on a specific floor.

5. Conclusion

The main objective of this chapter was to present a survey of the processors proposed in the literature, with a emphasis on their *applicability* for real-life problems. It is interesting to note that the most influential processors are not necessarily the most recent.

A complementary research area to the macro design problem is the use of simulation for evaluating layouts. However, only a few processors, mainly related to the machine layout problem, use simulation.

This extensive survey has allowed the identification of a number of gaps in the literature which could trigger new research avenues for the macro design problem. The stochastic environment, the construction/improvement dichotomy, the "several proposals/single recommendation" dichotomy, and the testing issue are elements worth considering.

The stochastic environment needs to be considered if processors aspire to be used for real-life problems. As mentioned by Heragu (1997), as of today, processors use explicit assumptions such as the knowledge of the future of manufacturing activities — including what products will be produced on what processing equipment — and the negligible variability of the product mix and volume. Those assumptions will need to be re-examined.

Even today, several processors are still developed for the construction of a layout. A new building, a location move, or a major enlargement, requires a construction method. Yet, in many cases simply a review of the existing layout is needed to optimize the productivity. This type of review constitutes a large part of the macro design problem. For instance, the classical procedures for layout improvement (2-way or 3-way exchanges of departments) should be integrated in any construction processor. The future is in meta-processors allowing multiple intermediate solutions.

Given that the processors for the macro design problem can handle only a few criteria, it is necessary to generate a set of layout proposals with the same quality. Based on those layouts, other criteria not easily quantifiable or not easily transferable into decision rules could be considered by an industrial engineer for recommending a layout. Here again, developing meta-processors constitutes an interesting research avenue.

Less than 50 percent of authors use the typical problem instances from the literature for evaluating and comparing generated layouts. The instance set of Nugent et al. (1968), still used by many, is not well adapted to the complexity of several processors. Several authors have generated their own instance set, some being used by others, e.g., Bazaraa (1975), Rosenblatt (1979), Dutta and Sahu (1982), Fortenberry and Cox (1985), Malakooti and D'Souza (1987), Golany and Rosenblatt (1989), and Skorin-Kapov (1990). In order to standardize the evaluation of the processors, it would be beneficial to the research community to gather together a set of typical problem instances, issuing from the manufacturing industry and as complete as possible. This set could include instances with departments of unequal sizes and dimension bounds, stochastic flows for a number of consecutive periods, diverse directed links, and other characteristics. Additionally, it is important to point out the significant role of the flow dominance factor of any instance used for evaluating a processor. Flow dominance is related to the amount of flow between two points with respect to the total amount of flow. Das (1993) summarizes the issue well by stating that the level of difficulty of a layout problem is inversely proportional to the degree of dominance of the circulation flows.

Appendix: Input parameters and variables used in engineering analysis and design methods

The list uses eight themes presented by Apple (1963, 1977). To simplify the presentation, only the input parameters and variables not cited by Apple (1963, 1977) are referenced.

Legend.

S: Subset of input parameters and variables used in engineering analysis and design methods used in our analysis (see Table 5.2)

IP: Input parameters

V: Variables

S	IP	V	Elements by theme
A. Materials or products (including scrap, removal, and waste products, and coolant, Askin and Standridge, 1993)			
			1. Characteristics (receipts and shipments)
	X		a. Size
	X		b. Shape
	X		c. Weight and density or bulkiness (Muther, 1973)

S	IP	V	Elements by theme
	X		d. Value or cost (Muther, 1973)
	X		e. Degree of palletization or of containerization
	X		f. Risk of damage (Muther and Haganäs, 1969)
	X		g. Condition
	X		h. Special control (Muther and Haganäs, 1969)
	X		i. Physical stage (solid, liquid, gas and unit, contained, bulk) (Muther and Haganäs, 1969)
√	X		2. Volume of production
	X		3. Timing, including seasonality
	X		4. Number of different parts and subassemblies
	X		5. Number and sequence of operations
	X	X	6. Storage requirements
B. Moves			
√	X		1. Frequency
	X	X	2. Speed
	X		3. Rate
	X		4. Volume
√		X	5. Distance
√	X		6. Sources
√	X		7. Destinations
√	X	X	8 Cross-traffic
	X		9. Required flow between work areas
√	X	X	10. Location of receiving & shipping
√		X	11. General linear flow
C. Handling methods			
√	X		1. Unit load
	X	X	2. Possible use of gravity
			3. MH principles (updated in Tompkins et al., 2003)
	X		a. Standardization principle
	X		b. Ergonomic principle
	X	X	c. Space utilization
	X	X	d. Automation principle
	X		e. Environmental principle
	X	X	f. Life cycle cost principle
	X		4. Desired flexibility
√	X	X	5. Equipment required
	X	X	6. Capacity requirements
	X		7. Limitations imposed by handling methods (Muther and Haganäs, 1969)
	X		8. Frequency and seriousness of potential breakdowns (Muther and Haganäs, 1969)
	X		9. Rapidity of repair (Muther and Haganäs, 1969)
√	X		10. Volume of spare parts required to stock (Tompkins et al., 2003)
	X		11. Availability of repair parts (Tompkins et al., 2003)
	X		12. Safety (materials, equipment, personnel)
√	X	X	13. Possible alternatives

S	IP	V	Elements by theme
	X		14. Ability to pace, or keep pace with, production requirements (Muther and Haganäs, 1969)
	X		15. Integration with and ability to serve the process operations (Muther and Haganäs, 1969)
D. Process			
√	X		1. Type (e.g., products layout, process layout)
√	X	X	2. Possible alternatives
	X		3. Possibility of performing during move
	X		4. Specific requirements of activities
	X		5. Quantity of equipment
√	X		6. Space requirements (size, shape, type, characteristics)
√	X		7. Adjacency restrictions (Heragu, 1997)
√	X	X	8. Equipments or departments location
√	X		9. Desired location of production services areas
√	X		10. Capacity requirements (Wrennal, 2001)
	X		11. Daily activity level (Reed, 1961)
	X		12. Work schedule (Reed, 1961)
√	X	X	13. Location of utilities and auxiliaries (including maintenance, repair, housekeeping, and fixed feature)
	X	X	14. Storage facilities (including raw materials, work-in-progress, and finished goods)
	X		15. Procedures and controls
	X		16. Safety (materials, equipment, personnel)
E. Building			
	X	X	1. Location on site and orientation
√	X	X	2. Building size and shape
	X	X	3. Construction type
	X	X	4. Structural design
√	X	X	5. Docks and doors — number, opening, size, location, height
√	X	X	6. Floors — numbers, condition, load capacity, type of flooring, resistance (e.g., to shock, abrasion, heat, vibration, humidity, solvents, salt, water, etc.), color, sanitary, odourless, static electricity, sound absorbent (Muther, 1973), and flatness (Sule, 1994)
	X	X	7. Walls characteristics, inside and outside (Sule, 1994)
√	X	X	8. Possible use of mezzanines, balconies, basement, roof
	X	X	9. Ceiling height
	X	X	10. Overhead load capacity
	X	X	11. Columns — location and spacing
	X	X	12. Windows (type, location, size)
√	X	X	13. Space availability and characteristics, including limits (Reed, 1961)
√	X	X	14. Elevators, ramps
√	X	X	15. Loading and unloading facilities
√	X	X	16. Aisle requirements — quantity, type, location, width
√	X	X	17. Aisle congestion (Heragu, 1997)
	X		18. Safety requirements
	X	X	19. Expansion possibilities

S	IP	V	Elements by theme
F. Site			
	X		1. Size and location
	X		2. Topography, including slope of the land
	X		3. Transportation facilities (road, rail, air, water)
	X		4. Expansion possibilities
	X		5. Weather conditions (prevailing wind, southern exposure, North light)
	X		6. Surroundings, including adjacent plants (dirt, fumes, etc.)
	X		7. Available power
	X		8. Within plant conditions (e.g., spread of contaminating materials, winter draft, glare from welding arcs, vibrations)
	X		9. Existing buildings
	X		10. Regulations — governments, city, building codes, and for insurance company (Heragu, 1997), including on waste disposal
	X		11. Company's own impact on the community (e.g., noise, hazards, traffic)
G. Personnel			
	X		1. Number
	X	X	2. Movement
	X	X	3. Working conditions (e.g., lighting, ventilation, heating, noise, vibration and temperature, natural light, fresh air, colors; Sule, 1994)
	X	X	4. Provision for fire protection — extinguishers, sprinkler systems, exits, etc.
√	X	X	5. First aid facilities
√	X	X	6. Aisle location and width
√	X	X	7. Desired location of personnel services areas (entrances, locker room, food service, etc.)
√	X		8. Supervisory requirements
			9. Personnel characteristics problems
	X		a. Available workers with proper skills (Tompkins et al., 2003)
	X		b. Training capability (Tompkins et al., 2003)
	X		c. Disposition of redundant workers (Tompkins et al., 2003)
	X		d. Job description changes (Tompkins et al., 2003)
	X		e. Union contracts (Tompkins et al., 2003)
	X		f. Work practices (Tompkins et al., 2003)
	X	X	10. Safety
H. Miscellaneous			
	X		1. Nature of business and economic cyclic effects (Reed, 1961)
	X		2. Company policies, including make or buy policy (Reed, 1961)
	X	X	3. Flexibility
	X		4. Degree of automation (Tompkins et al., 2003)
	X		5. Software requirements (Tompkins et al., 2003)

Acknowledgements This work was supported by the Canadian Natural Sciences and Engineering Council and by the Fonds de recherche sur la nature et les technologies — Québec.

References

Abdinnour-Helm, S. and Hadley, S.W. (2000). Tabu search based heuristics for multi-floor facility layout. *International Journal of Production Research*, 38:365 – 383.

Abdou, G. and Dutta, S.P. (1990). An integrated approach to facilities layout using expert systems. *International Journal of Production Research*, 28:685 – 708.

Aiello, G. and Enea, M. (2001). Fuzzy approach to the robust facility layout in uncertain production environments. *International Journal of Production Research*, 39:4089 – 4101.

Akinc, U. (1985). Multi-activity facility design and location problems. *Management Science*, 31:275 – 283.

Al-Hakim, L.A. (1991). Two graph-theoretic procedures for an improved solution to the facilities layout problem. *International Journal of Production Research*, 29:1701 – 1718.

Al-Hakim, L.A. (1992). A modified procedure for converting a dual graph to a block layout. *International Journal of Production Research*, 30:2467 – 2476.

Al-Hakim, L.A. (2000). On solving facility layout problems using genetic algorithms. *International Journal of Production Research*, 38:2573 – 2582.

Anjos, M.F. and Vannelli, A. (2002). *A New Mathematical Programming Framework for Facility Layout Design*. Working paper.

Apple, J.M. (1963). *Plant Layout and Materials Handling*. The Ronald Press Company.

Apple, J.M. (1977). *Plant Layout and Materials Handling*, 3rd edition. John Wiley & Sons Inc.

Armour, G.C. and Buffa, E.S. (1963). A heuristic algorithm and simulation approach to relative location of facilities. *Management Science*, 9:294 – 309.

Askin, R.G. and Standridge, C.R. (1993). *Modeling and Analysis of Manufacturing Systems*. John Wiley & Sons.

Azadivar, F. and Wang, J.J. (2000). Facility layout optimization using simulation and genetic algorithms. *International Journal of Production Research*, 38:4369 – 4383.

Badiru, A.B. and Arif, A. (1996). FLEXPERT: Facility layout expert system using fuzzy linguistic relationship codes. *IIE Transactions*, 28:295 – 308.

Balakrishnan, J. and Cheng, C.H. (1998). Dynamic layout algorithms: A state-of-the-art survey. *OMEGA — International Journal of Management Science*, 26:507 – 522.

Balakrishnan, J. and Cheng, C.H. (2000). Genetic search and the dynamic layout problem. *Computers and Operations Research*, 27:587 – 593.

Balakrishnan, J., Cheng, C.H., and Conway, D.G. (2000). An improved pair-wise exchange heuristic for the dynamic plant layout problem. *International Journal of Production Research*, 38:3067 – 3077.

Balakrishnan, J., Cheng, C.H., Conway, D.G., and Lau, C. M. (2003a). A hybrid genetic algorithm for the dynamic plant layout problem. *International Journal of Production Economics*, 86:107 – 120.

Balakrishnan, J., Cheng, C.-H., and Wong, K.-F. (2003b) FACOPT: a user friendly FACility layout OPTimization system. *Computers and Operations Research*, 30: 1625 – 1641.

Banerjee, P., Montreuil, B., Moodie, C.L., and Kashyap, R.L. (1990). A qualitative reasoning-based interactive optimization methodology for layout design. *International industrial engineering conference proceedings*, pp. 230 – 235.

Banerjee, P., Montreuil, B., Moodie, C.L., and Kashyap, R.L. (1992a). A modelling of interactive facilities layout designer reasoning using qualitative patterns. *Inter-*

national Journal of Production Research, 30:433 – 453.

Banerjee, P., Zhou, Y., and Montreuil, B. (1992b). *Conception d'aménagement d'installations : Optimisation par induction génétique.* Technical Report 92-66, Groupe de recherche en gestion de la logistique, Faculté des sciences de l'administration de l'Université Laval.

Banerjee, P., Zhou, Y., and Montreuil, B. (1997). Genetically assisted optimization of cell layout and material flow path skeleton.*IIE Transactions*, 29:277 – 291.

Barbosa-Povoa, A.P. Mateus, R., and Novais, A.Q. (2001). Optimal two-dimensional layout of industrial facilities. *International Journal of Production Research*, 39: 2567 – 2593.

Barbosa-Povoa, A.P., Mateus, R., and Novais, A.Q. (2002). Optimal 3D layout of industrial facilities. *International Journal of Production Research*, 40:1669 – 1698.

Baykasoglu, A. and Gindy, N.N.Z. (2001). A simulated annealing algorithm for dynamic layout problem. *Computers and Operations Research*, 28:1403 – 1426.

Bazaraa, M.S. (1975). Computerized layout design: A branch and bound approach. *AIIE Transactions*, 7:432 – 438.

Benjaafar, S. and Sheikhzadeh, M. (2000). Design of flexible plant layouts. *IIE Transactions*, 32:309 – 322.

Boswell, S.G. (1992). TESSA — A new greedy heuristic for facilities layout planning. *International Journal of Production Research*, 30:1957 – 1968.

Bozer, Y.A., Meller, R.D., and Erlebacher, S.J. (1994). An improvement-type layout algorithm for single and multiple-floor facilities. *Management Science*, 40:918 – 932.

Butler, T.W., Karwan, K.R., Sweigart, J.R., and Reeves, G.R. (1992). An integrative model-based approach to hospital layout. *IIE Transactions*, 24:144 – 152.

Cambron, K.E. and Evans, G.W. (1991). Layout design using the analytic hierarchy process. *Computers and Industrial Engineering*, 20:211 – 229.

Castell, C.M.L., Lakshmanan, R., Skilling, J.M., and Banares-Alcantara, R. (1998). Optimisation of process plant layout using genetic algorithms. *Computers and Chemical Engineering*, 22:S993 - S996.

Castillo, I., and Peters, B.A. (2002). Unit load and material-handling considerations in facility layout design. *International Journal of Production Research*, 40:2955 – 2989.

Castillo, I. and Peters, B.A. (2003). An extended distance-based facility layout problem. *International Journal of Production Research*, 41:2451 – 2479.

Cedarleaf, J. (1994). *Plant Layout and Flow Improvement.* McGraw-Hill Book Company Inc.

Chen, C.-W. and Sha, D.Y. (1999). A design approach to the multi-objective facility layout problem. *International Journal of Production Research*, 37:1175 – 1196.

Cheng, R. and Gen, M. (1998). Loop layout design problem in flexible manufacturing systems using genetic algorithms. *Computers and Industrial Engineering*, 34:53 – 61.

Cheng, R., Gen, M., and Tosawa, T. (1996). Genetic algorithms for designing loop layout manufacturing systems. *Computers and Industrial Engineering*, 31:587 – 591.

Chhajed, D., Montreuil, B., and Lowe, T.J. (1990). *Flow network design for manufacturing systems layout.* Technical Report 90-12, Faculté des sciences de l'administration de l'Université Laval.

Chiang, W.-C. (2001). Visual facility layout design system. *International Journal of Production Research*, 39:1811 – 1836.

Chiang, W.-C. and Chiang, C. (1998). Intelligent local search strategies for solving facility layout problems with the quadratic assignment problem formulation. *Eu-*

ropean Journal of Operational Research, 106:457 – 488.

Chiang, W.-C. and Kouvelis, P. (1996). An improved tabu search heuristic for solving facility layout design problems. *International Journal of Production Research*, 34:2565 – 2585.

Chittratanawat, S. and Noble, J.S. (1999). An integrated approach for facility layout, P/D location and material handling system design. *International Journal of Production Research*, 37:683 – 706.

Chwif, L., Pereira Barretto, M.R., and Moscato, L.A. (1998). A solution to the facility layout problem using simulated annealing. *Computers in Industry*, 36:125 – 132.

Cimikowski, R. and Mooney, E. (1997). Proximity-based adjacency determination for facility layout. *Computers and Industrial Engineering*, 32:341 – 349.

Das, S.K. (1993). A facility layout method for flexible manufacturing systems. *International Journal of Production Research*, 31:279 – 297.

Deb, S.K. and Bhattacharyya, B. (2003). Facilities layout planning based on Fuzzy multiple criteria decision-making methodology. *International Journal of Production Research*, 41:4487 – 4504.

Delmaire, H., Langevin, A., and Riopel, D. (1995a). *Evolution Systems and the Quadratic Assignment Problem.* Technical Report G-95-24, Les Cahiers du GERAD, Montréal, Canada.

Delmaire, H., Langevin, A., and Riopel, D. (1995b). *Le problème du design d'implantation d'usine: une revue.* Technical Report G-95-09, Les Cahiers du GERAD, Montréal, Canada.

Delmaire, H., Langevin, A., and Riopel, D. (1997). Skeleton-based facility layout design using genetic algorithms. *Annals of Operations Research*, 69:85 – 104.

Drezner, Z. (1987). A heuristic procedure for the layout of a large number of facilities. *Management Science*, 33:907 – 915.

Dutta, K.N. and Sahu, S. (1982). A multigoal heuristic for facilities design problems: MUGHAL. *International Journal of Production Research*, 20:147 – 154.

Dweiri, F. (1999). Fuzzy development of crisp activity relationship charts for facilities layout. *Computers and Industrial Engineering*, 36:1 – 16.

Dweiri, F. and Meier, F.A. (1996). Application of fuzzy decision-making in facilities layout planning. *International Journal of Production Research*, 34:3207 – 3225.

Evans, G.W., Wilhelm, M.R., and Karwowski, W. (1987). A layout design heuristic employing the theory of fuzzy sets. *International Journal of Production Research*, 25:1431 – 1450.

Fortenberry, J.C. and Cox, J.F. (1985). Multiple criteria approach to the facilities layout problem. *International Journal of Production Research*, 23:773 – 782.

Foulds, L.R., Gibbons, P.B., and Giffin, J.W. (1985). Facilities layout adjacency determination: An experimental comparison of three graph theoretic heuristics. *Operations Research*, 33:1091 – 1106.

Foulds, L.R. and Giffin, J.W. (1985). A graph-theoretic heuristic for minimizing total transport cost in facilities layout. *International Journal of Production Research*, 23:1247 – 1257.

Gaston, G.K. (1984). Facility layout optimizes space, minimizes costs. *IE*, May:22 – 28.

Gau, K.-Y. and Meller, R.D. (1999). An iterative facility layout algorithm. *International Journal of Production Research*, 37:3739 – 3758.

Gavett, J.W. and Plyter, N.V. (1966). The optimal assignment of facilities to locations by branch and bound. *Operations Research*, 14:210 – 232.

Giffin, J.W. and Foulds, L.R. (1987). Facilities layout generalized model solved by n-boundary shortest path heuristics *European Journal of Operational Research*,

28:382 – 391.

Glover, F. (1989). Tabu search, Part I. *ORSA Journal on Computing*, 1:190 – 206.

Goetschalckx, M. (1992). An interactive layout heuristic based on hexagonal adjacency graphs. *European Journal of Operational Research*, 63:304 – 321.

Golany, B. and Rosenblatt, M.J. (1989). A heuristic algorithm for the quadratic assignment formulation to the plant layout problem. *International Journal of Production Research*, 27:293 – 308.

Goldschmidt, O., Takvorian, A., and Yu, G. (1996). On finding a biconnected spanning planar subgraph with applications to the facilities layout problem. *European Journal of Operational Research*, 94:97 – 105.

Gómez, A., Fernández, Q.I., De la Fuente García, D., and García, P.J. (2003). Using genetic algorithms to resolve layout problems in facilities where there are aisles. *International Journal of Production Economics*, 84:271 – 282.

Green, R.H. and Al-Hakim, L. (1985). A heuristic for facilities layout planning. *OMEGA — International Journal of Management Science*, 13:469 – 474.

Grobelny, J. (1987). On one possible 'fuzzy' approach to facilities layout problems. *International Journal of Production Research*, 25:1123 – 1141.

Grobelny, J. (1988). The 'linguistic pattern' method for a workstation layout analysis. *International Journal of Production Research*, 26:1779 – 1798.

Hamamoto, S., Yih, Y., and Salvendy, G. (1999). Development and validation of genetic algorithm-based facility layout: A case study in the pharmaceutical industry. *International Journal of Production Research*, 37:749 – 768.

Hansen, P. (1986). The steepest ascent, mildest descent heuristic for combinatorial programming. *Congrès sur les méthodes numériques en Optimisation Combinatoire*, Capri, Italy.

Harmonosky, C.M. and Tothero, G.K. (1992). A multi-factor plant layout methodology. *International Journal of Production Research*, 30:1773 – 1789.

Hassan, M.M.D. and Hogg, G.L. (1987). A review of graph theory application to the facilities layout problem. *OMEGA — International Journal of Management Science*, 15:291 – 300.

Hassan, M.M.D. and Hogg, G.L. (1989). On converting a dual graph into a block layout. *International Journal of Production Research*, 27:1149 – 1160.

Hassan, M.M.D. and Hogg, G.L. (1991). On constructing a block layout by graph theory. *International Journal of Production Research*, 29:1263 – 1278.

Hassan, M.M.D., Hogg, G.L., and Smith, D.R. (1986). SHAPE: A construction algorithm for area placement evaluation. *International Journal of Production Research*, 24:1283 – 1295.

Heragu, S.S. (1992). Recent models and techniques for solving the layout problem. *European Journal of Operational Research*, 57:136 – 144.

Heragu, S.S. (1997). *Facilities Design.* PWS Publishing Company.

Heragu, S.S. and Alfa, A.S. (1992). Experimental analysis of simulated annealing based algorithms for the layout problem. *European Journal of Operational Research*, 57:190 – 202.

Heragu, S.S. and Kusiak, A. (1990). Machine layout: An optimization and knowledge-based approach. *International Journal of Production Research*, 28:615 – 635.

Heragu, S.S. and Kusiak, A. (1991).Efficient models for the facility layout problem. *European Journal of Operational Research*, 53:1 – 13.

Ho, Y.-C. and Moodie, C.L. (2000). A hybrid approach for concurrent layout design of cells and their flow paths in a tree configuration. *International Journal of Production Research*, 38:895 – 928.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. Technical Report, Michigan University.

Houshyar, A. (1991). Computer aided facility layout: An interactive multi-goal approach. *Computers and Industrial Engineering*, 20:177 – 186.

Houshyar, A. and White, B. (1993). Exact optimal solution for facility layout: Deciding which pairs of locations should be adjacent. *Computers and Industrial Engineering*, 24:177 – 187.

Immer, J.R. (1953). *Material Handling*. McGraw-Hill Book Company Inc.

Irvine, S.A. and Rinsma-Melchert, I.(1997). A new approach to the block layout problem. *International Journal of Production Research*, 35:2359 – 2376.

Islier, A.A. (1998). A genetic algorithm approach for multiple criteria facility layout design. *International Journal of Production Research*, 36:1549 – 1569.

Jacobs, F.R. (1987). A layout planning system with multiple criteria and a variable domain representation. *Management Science*, 33:1020 – 1034.

Jajodia, S., Minis, I., Harhalakis, G., and Proth, J.-M. (1992). CLASS: Computerized LAyout Solutions using Simulated annealing. *International Journal of Production Research*, 30:95 – 108.

Johnson, R.V. (1982). SPACECRAFT for multi-floor layout planning. *Management Science*, 28:407 – 417.

Kaku, B.K. and Rachamadugu, R. (1992). Layout design for flexible manufacturing systems. *European Journal of Operational Research*, 57:224 – 230, 1992.

Kaku, B.K., Thompson, G. L., and Baybars, I. (1988). A heuristic method for the multi-story layout problem. *European Journal of Operational Research*, 37:384 – 397.

Ketcham, M.G. (1992). A branch and bound approach to facility design for continuous flow manufacturing systems. *International Journal of Production Research*, 30:573 – 597.

Khalil, T.M. (1973). Facilities relative allocation technique (FRAT). *International Journal of Production Research*, 11:183 – 194.

Kim, C.-B., Foote, B.L., and Pulat, P.S. (1995). Cut-tree construction for facility layout. *Computers and Industrial Engineering*, 28:721 – 730.

Kim, J.-G. and Kim, Y.-D. (1998). A space partitioning method for facility layout problems with shape constraints. *IIE Transactions*, 30:947 – 957, 1998.

Kim, J.-G. and Kim, Y.-D. (2000). Layout planning for facilities with fixed shapes and input and output points. *International Journal of Production Research*, 38:46354653.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. *Management Science*, 29:671 – 680.

Kochhar, J.S., Foster, B.T., and Heragu, S.S. (1998). HOPE: A genetic algorithm for the unequal area facility layout problem. *Computers and Operations Research*, 25:583 – 594.

Kochhar, J.S. and Heragu, S.S. (1998). MULTI-HOPE: A tool for multiple floor layout problems. *International Journal of Production Research*, 36:3421 – 3435.

Kochhar, J.S. and Heragu, S.S. (1999). Facility layout design in a changing environment. *International Journal of Production Research*, 37:2429 – 2446.

Kouvelis, P. and Chiang, W.-C. (1992). A simulated annealing procedure for single row layout problems in flexible manufacturing systems. *International Journal of Production Research*, 30:717 – 732.

Kouvelis, P., Chiang, W.-C., and Fitzsimmons, J. (1992a). Simulated annealing for machine layout problems in the presence of zoning constraints. *European Journal* *of Operational Research*, 57:203 – 223.

Kouvelis, P., Kurawarwala, A.A., and Gutiérrez, G.J. (1992b). Algorithms for robust single and multiple period layout planning for manufacturing systems. *European Journal of Operational Research*. 63:287 – 303.

Kumara, S.R.T., Kashyap, R.L., and Moodie, C.L. (1987). Expert system for industrial facilities layout planning and analysis. *Computers and Industrial Engineering*, 12:143 – 152, 1987.

Kumara, S.R.T., Kashyap, R.L., and Moodie, C.L. (1988). Application of expert systems and pattern recognition methodologies to facilities layout planning. *International Journal of Production Research*, 26:905 – 930.

Kusiak, A. and Heragu, S.S. (1987). The facility layout problem. *European Journal of Operational Research*, 29:229 – 251.

Lacksonen, T.A. (1994). Static and dynamic layout problems with varying areas. *Journal of the Operational Research Society*, 45:59 – 69.

Lacksonen, T.A. (1997). Preprocessing for static and dynamic facility layout problems. *International Journal of Production Research*, 35:1095 – 1106.

Langevin, A., Montreuil, B., and Riopel, D. (1994). Spine layout design. *International Journal of Production Research*, 32:429 – 442.

Lee, K.-Y., Han, S.-N., and Roh, M.-I. (2003). An improved genetic algorithm for facility layout problems having inner structure walls and passages. *Computers and Operations Research,* 30:117 – 138.

Lee, Y. H. and Lee, M.H. (2002). A shape-based block layout approach to facility layout problems using hybrid genetic algorithm. *Computers and Industrial Engineering*, 42:237 – 248.

Leung, J. (1992). A new graph-theoretic heuristic for facility layout. *Management Science*, 38:594 – 605.

Levary, R.R. and Kalchik, S. (1985). Facilities layout: A survey of solution procedures. *Computers and Industrial Engineering*, 9:141 – 148.

Li, H. and Love, P.E.D. (2000). Genetic search for solving construction site-level unequal-area facility layout problems. *Automation in Construction*, 9:217 – 226.

Liggett, R.S. (1981). The quadratic assignment problem: An experimental evaluation of solution strategies. *Management Science*, 27:442 – 458.

Lin, J.-L., Foote, B.L., Pulat, S., Chang, C.-H., and Cheung, J.Y. (1994). *SMILE: An Algorithm for Plant Layout with Constraints on Department Shape Change.* Working, paper, University of Oklahoma.

Love, R.F. and Wong, J.Y. (1976). On solving a one-dimensional space allocation problem with integer programming. *INFOR,* 14:139 – 143.

Mak, K.L., Wong, Y.S., and Chan, F.T.S. (1998). A genetic algorithm for facility layout problems. *Computer Integrated Manufacturing Systems*, 11:113 – 127.

Malakooti, B. (1989). Multiple objective facility layout: A heuristic to generate efficient alternatives. *International Journal of Production Research*, 27:1225 – 1238.

Malakooti, B. and D'Souza, G.I. (1987). Multiple objective programming for the quadratic assignment problem. *International Journal of Production Research*, 25:285 – 300.

Malakooti, B. and Tsurushima, A. (1989). An expert system using priorities for solving multiple-criteria facility layout problems. *International Journal of Production Research*, 27:793 – 808.

Marcoux, N. (1999). *Implantation et manutention : indicateurs de performance et relations type-forme.* Thesis, École Polytechnique de Montréal.

Matsuzaki, K., Irohara, T., and Yoshimoto, K. (1999). Heuristic algoritm to solve the multi-floor layout problem with the consideration of elevator utilization. *Computers and Industrial Engineering*, 36:487 - 502.

McKendall, A.R., Noble, J.S., and Klein, C.M. (1999). Facility layout of irregular-shaped departments using a nested approach. *International Journal of Production Research*, 37:2895 - 2914.

Meller, R.D. and Bozer, Y.A. (1991). *Solving the Facility Layout Problem with Simulated Annealing.* Technical Report 91-20, University of Michigan.

Meller, R.D. and Bozer, Y.A. (1997). Alternative approaches to solve the multi-floor facility layout problem *Journal of Manufacturing Systems*, 16:192 - 203.

Meller, R.D. and Gau, K.-Y. (1996). The facility layout problem: Recent and emerging trends and perspectives. *Journal of Manufacturing Systems*, 15:351 - 366.

Michalewicz, Z. (1992). *Genetic Algorithms + Data Structures = Evolution Programs.* Springer-Verlag.

Mir, M. and Imam, M.H. (2000). A hybrid optimization approach for layout design of unequal-area facilities. *Computers and Industrial Engineering*, 39:49 - 63.

Montreuil, B., Brotherton, É., and Marcotte, S. (2002b). Zone-based facilities layout optimization. *Proceedings of the Industrial Engineering Research Conference, IIE Annual Conference*, Orlando, FL.

Montreuil, B. and Laforge, A. (1992). Dynamic layout design given a scenario tree of probable futures. *European Journal of Operational Research*, 63:271 - 286.

Montreuil, B. and Ratliff, H.D. (1988). Optimizing the location of input/output stations within facilities layout. *Engineering Costs and Production Economics*, 14:177 - 187.

Montreuil, B. and Ratliff, H.D. (1989). Utilizing cut trees as design skeletons for facility layout. *IIE Transactions*, 21:136 - 143.

Montreuil, B., Ratliff, H.D., and Goetschalckx, M. (1987). Matching based interactive facility layout. *IIE Transactions*, 19:271 - 279.

Montreuil, B. and Venkatadri, U. (1988). *From Gross to Net Layouts: An Efficient Design Model.* Technical Report 88-56, Faculté des sciences de l'administration de l'Université Laval.

Montreuil, B., Venkatadri, U., and Blanchet, E. (2002a). Generating a Net Layout from a Block Layout with Superimposed Flow Networks. Working paper, Faculté des sciences de l'administration de l'Université Laval.

Montreuil, B., Venkatadri, U., and Ratliff, H.D. (1989). *Generating a Layout from a Design Skeleton.* Technical Report 89-01, Faculté des sciences de l'administration de l'Université Laval.

Moore, J.M. (1962). *Plant Layout and Design.* MacMillan Publishing Company Inc.

Muther, R. (1973). *Systematic Layout Planning.* CBI Publishing Company Inc.

Muther, R. and Haganäs, K. (1969). *Systematic Handling Analysis.* Management & Industrial Research Publications.

Muther, R. and Hales, L. (1979). *Systematic Planning of Industrial Facilities.* Management & Industrial Research Publications.

Nadler, G. (1967). *Work Systems Design: The IDEALS Concept.* Richard D. Irwin.

Neghabat, F. (1974). An efficient equipment-layout algorithm. *Operations Research*, 22:622 - 628.

Norman, B.A., Arapoglu, R.A., and Smith, A.E. (2001). Integrated facilities design using a contour distance metric. *IIE Transactions*, 33:337 - 344.

Norman, B.A. and Smith, A.E. (1999). *Considering Production Uncertainty in Block Layout Design.* Working paper.

Nugent, C.E., Vollmann, T.E., and Ruml, J. (1968). An experimental comparison of techniques for the assignment of facilities to locations. *Operations Research*, 16:150 - 173.

O'Brien, C. and Abdel Barr, S.E.Z. (1980). An interactive approach to computer aided facility layout. *International Journal of Production Research*, 18:201 - 211.

Ozdemir, G., Smith, A.E., and Norman, B.A. (2003). Incorporating heterogeneous distance metrics within block layout design. *International Journal of Production Research*, 41:1045 - 1056.

Partovi, F.Y. and Burton, J. (1992). An analytical hierarchy approach to facility layout. *Computers and Industrial Engineering*, 22:447 - 457.

Pesch, E., Glover, F., Bartsch, T., Salewski, F., and Osman, I. (1999). Efficient facility layout planning in a maximally planar graph model. *International Journal of Production Research*, 37:263 - 283.

Philips E.J. (1997). *Manufacturing Plant Layout: Fundamentals and Fine Points of Optimum Facility Design.* Society of Manufacturing Engineers.

Picard, J.-C. and Queyranne, M. (1981). On the one-dimensional space allocation problem. *Operations Research*, 29:371 - 391.

Pierreval, H., Caux, C., Paris, J.L., and Viguier, F. (2003). Evolutionary approaches to the design and organization of manufacturing systems. *Computers and Industrial Engineering,* 44:339 - 364.

Rajasekharan, M., Peters, B.A., and Yang, T. (1998). A genetic algorithm for the facility layout design in flexible manufacturing systems. *International Journal of Production Research*, 36:95 - 110.

Rao, H.A., Pham, S.N., and Gu, P. (1999). A genetic algorithms-based approach for design of manufacturing systems: An industrial application. *International Journal of Production Research*, 37:557 - 580.

Raoot, A.D. and Rakshit, A. (1991). A 'fuzzy' approach to facilities lay-out planning. *International Journal of Production Research*, 29:835 - 857.

Raoot, A.D. and Rakshit, A. (1993). A 'linguistic pattern' approach for multiple criteria facility layout problems. *International Journal of Production Research*, 31:203 - 222.

Reed, R., Jr. (1961). *Plant layout: Factors, Principles, and Techniques.* Richard D. Irwin Inc.

Reed, R., Jr. (1967). *Plant Location, Layout, and Maintenance.* Richard D. Irwin Inc.

Rinsma, I., Giffin, J.W., and Robinson, D.F. (1990). Orthogonal floorplans from maximal planar graphs. *Environment and Planning B: Planning and Design*, 17:57 - 71.

Rosenblatt, M.J. (1979).The facilities layout problem: A multi-goal approach. *International Journal of Production Research*, 17:323 - 332.

Rosenblatt, M.J. (1986). The dynamics of plant layout. *Management Science*, 32:76 - 86.

Rosenblatt, M.J. and Golany, B. (1992). A distance assignment approach to the facility layout problem. *European Journal of Operational Research*, 57:253 - 270.

Rosenblatt, M.J. and Kropp, D.H. (1992). The single period stochastic plant layout problem. *IIE Transactions*, 24:169 - 176.

Rosenblatt, M.J. and Lee, H.L. (1987). A robustness approach to facilities design. *International Journal of Production Research*, 25:479 - 486.

Safizadeh, M.H. and McKenna, D.R. (1996). Application of multidimensional scaling techniques to facilities layout. *European Journal of Operational Research*, 92:54 - 62.

Scriabin, M. and Vergin, R.C. (1985). A cluster-analytic approach to facility layout. *Management Science*, 31:33 – 49.

Seppänen, J. and Moore, J.M. (1970). Facilities planning with graph theory. *Management Science*, 17:B242 – B253.

Seppänen, J. and Moore, J.M. (1975). String processing algorithms for plant layout problems. *International Journal of Production Research*, 13:239 – 254.

Shang, J.S. (1993). Multicriteria facility layout problem: An integrated approach. *European Journal of Operational Research*, 66:291 – 304.

Sheth, V.S. (1995). *Facilities Planning and Materials Handling: Methods and Requirements* Marcel Dekker.

Shih, L.C., Enkawa, T., and Itoh, K. (1992). An AI-search technique-based layout planning method. *International Journal of Production Research*, 30:2839 – 2855.

Sirinaovakul, B. and Thajchayapong, P. (1994). A knowledge base to assist a heuristic search approach to facility layout. *International Journal of Production Research*, 32:141 – 160.

Skorin-Kapov, J. (1990). Tabu search applied to the quadratic assignment problem. *ORSA Journal on Computing*, 2:33 – 41.

Skorin-Kapov, J. (1991). *Extensions of a Tabu Search Adaptation to the Quadratic Assignment Problem.* Technical Report HAR-90-006, W.A Harriman School for Management and Policy, State University of New York at Stony Brook.

Sule, D.R. (1994). *Manufacturing facilities: Location, Planning and Design*, 2nd edition. PWS Publishing Company.

Suresh, G., Vinod, V.V., and Sahu, S. (1995). A genetic algorithm for facility layout. *International Journal of Production Research*, 33:3411 – 3423.

Svestka, J.A. (1990). MOCRAFT: A professional quality micro-computer implementation of CRAFT with multiple objectives. *Computers and Industrial Engineering*, 18:13 – 22.

Tam, K.Y. (1992a). A simulated annealing algorithm for allocating space to manufacturing cells. *International Journal of Production Research*, 30:63 – 87.

Tam, K.Y. (1992b). Genetic algorithms, function optimization, and facility layout design. *European Journal of Operational Research*, 63:322 – 346.

Tam, K.Y. and Chan, S.K. (1998). Solving facility layout problems with geometric constraints using parallel genetic algorithms: experimentation and findings. *International Journal of Production Research*, 36:3253 – 3272.

Tate, D.M. and Smith, A.E. (1995). Unequal-area facility layout by genetic search. *IIE Transactions*, 27:465 – 472.

Tavakkoli-Moghaddain, R. and Shayan, E. (1998). Facilities layout design by genetic algorithms. *Computers and Industrial Engineering*, 35:527 – 530.

Tompkins, J.A. and Reed, R., Jr. (1976). An applied model for the facilities design problem. *International Journal of Production Research*, 14:583 – 595.

Tompkins, J.A. and White, J.A. (1984). *Facilities Planning.* John Wiley and Sons.

Tompkins, J.A., White, J.A., Bozer, Y.A., Frazelle, E.H., Tanchoco, J.M.A., and Trevino, J. (1996). *Facilities Planning*, 2nd edition. John Wiley & Sons.

Tompkins, J.A., White, J.A., Bozer, Y., and Tanchoco, J.M.A. (2003). *Facilities Planning*, 3rd edition. John Wiley & Sons.

Tretheway, S.J. and Foote, B.L. (1994). Automatic computation and drawing of facility layouts with logical aisle structures. *International Journal of Production Research*, 32:1545 – 1555.

Tsuchiya, K., Bharitkar, S., and Takefuji, Y. (1996). A neural network approach to facility layout problems. *European Journal of Operational Research*, 89:556 – 563.

Urban, T.L. (1987). A multiple criteria model for the facilities layout problem. *International Journal of Production Research*, 25:1805 - 1812.

Urban, T.L. (1993). A heuristic for the dynamic facility layout problem. *IIE Transactions*, 25:57 - 63.

Urban, T.L. (1998). Solution procedures for the dynamic facility layout problem. *Annals of Operations Research*, 76:323 - 342.

Van Camp, D.J., Carter, M.W., and Vannelli, A. (1992). A nonlinear optimization approach for solving facility layout problems. *European Journal of Operational Research*, 57:174 - 189.

Wang, M.J., Liu, C.M., and Pan, Y.S. (1991). Computer-aided panel layout using a multi-criteria heuristic algorithm. *International Journal of Production Research*, 29:1215 - 1233, 1991.

Wang, T.-C. and Wong, D.F. (1990). An optimal algorithm for floorplan area optimization. *27th ACM/IEEE design automation conference*, pp. 180 - 186.

Wäscher, G. and Chamoni, P. (1987). MICROLAY: An interactive computer program for factory layout planning on microcomputers. *European Journal of Operational Research*, 31:185 - 193.

Wäscher, G. and Merker, J. (1997). A comparative evaluation of heuristics for the adjacency problem in facilities layout planning. *International Journal of Production Research*, 35:447 - 466, 1997.

Watson, K.H. and Giffin, J.W. (1997). The vertex splitting algorithm for facilities layout. *International Journal of Production Research*, 35:2477 - 2492.

Welgama, P.S. and Gibson, P.R. (1993). A construction algorithm for the machine layout problem with fixed pick-up and drop-of points. *International Journal of Production Research*, 31:2575 - 2590.

Welgama, P.S. and Gibson, P.R. (1996). An integrated methodology for automating the determination of layout and materials handling system. *International Journal of Production Research*, 34:2247 - 2264.

Wrennal, W. (2001). *Chapter 8.2 — Facilities Layout and Design.* In: K. B. Zandin (ed.), *Maynard's Industrial Engineering Handbook*, pp. 8.21 - 8.62, McGraw-Hill Book Company Inc.

Wu, Y. and Appleton, E. (2002a). Integrated design of the block layout and aisle structure by simulated annealing. *International Journal of Production Research*, 40:2353 - 2365.

Wu, Y. and Appleton, E. (2002b). The optimisation of block layout and aisle structure by a genetic algorithm. *Computers and Industrial Engineering*, 41:371 - 387.

Yaman, R., Gethin, D.T., and Clarke, M.J. (1993). An effective sorting method for facility layout construction. *International Journal of Production Research*, 31:413 - 427, 1993.

Yang, T. and Kuo, C. (2003). A hierarchical AHP/DEA methodology for the facilities layout design problem. *European Journal of Operational Research,* 147:128 - 136.

Chapter 6

THE DESIGN, PLANNING, AND OPTIMIZATION OF REVERSE LOGISTICS NETWORKS

Nathalie Bostel
Pierre Dejax
Zhiqiang Lu

Abstract Reverse logistics is concerned with the return flows of products or equipment back from the consumer to the logistics network for reuse, recovery or recycling for environmental, economic or customer service reasons. In this paper, we review applications, case studies, models and techniques proposed for the design, planning and optimization of reverse logistics systems. We consider both cases of separate and integrated handling of original products and return flows throughout the logistics network. According to the hierarchical planning framework for logistics systems, the works are described in relation to their contribution to strategic, tactical or operational planning. Major contributions concern facility location, inventory management, transportation and production planning models. Directions for further research are indicated in all of these areas as well as for the general development of reverse logistics activities in a supply chain network.

1. Introduction

1.1 Basic concepts

In recent years, many companies have begun to pay attention to used products and materials, because of legislative, economic and commercial factors (Fleischmann et al., 2000b). Reduction of waste has become a major concern for industrial countries in view of declining landfill and incineration capacity. In addition to growing disposal costs, governmental legislation requires producers to take charge of their products throughout their life cycle. Environmentally concerned customers now expect "green companies" to reduce the quantity of waste generated and to recycle resources encompassed within used products.

Recovery programs have also demonstrated an economic interest for industry: a reduction in the cost of raw materials due to recycling, a reduction in the cost of manufacturing packages by reutilization, a decrease in disposal costs because of reduced quantities (Lu et al., 2001). For enterprises whose products are particularly costly and sophisticated, the reuse of products or components may represent a reduction of 50% of production costs (Fontanella, 1999).

Several definitions of reverse logistics (RL) have been proposed by various authors, such as the American Reverse Logistics Executive Council (Rogers and Tibben-Lembke, 1998), but also Philipp (1999); Stock (1999); Beaulieu et al. (1999); Browne and Allen (1999). In order to emphasize the links between traditional forward flows and reverse flows in an integrated logistics system, we propose the following definition: "Reverse Logistics can be viewed as an evolution of traditional forward logistics in an environmentally-conscious industry or due to other commercial drives; it encompasses all the logistics activities and management functions necessary for reintroducing valued-objects, which have finished or are not suitable to perform their primary function any more, into certain recovery systems for either recapturing their value or proper disposal" (Lu, 2003).

De Brito and Dekker (2002) have compared existing reverse logistics definitions. They distinguish several types of recovery activity:

product recovery (products may be recycled directly into the original market or into a secondary market, or repaired and sent back to the user under conditions of warranty),

component recovery (products are dismantled and parts can be remanufactured into the same kind of product or different products),

material recovery (materials are recuperated and recycled into raw materials like metal, paper or glass),

energy recovery (incineration).

At this point, it is important to emphasize the global nature of the reverse logistics concepts and their differences from concepts such as:

- waste management, because for these products there is no new use or no recovery value,
- green logistics, which considers environmental aspects of forward logistics,
- transportation of empty materials such as containers or movements of empty vehicles, transport activities being complementary to logistics activities.

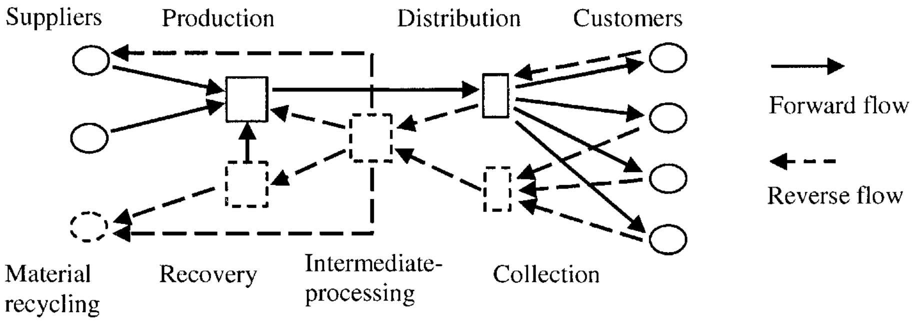


Figure 6.1. Framework of an integrated logistics system with forward and reverse flows

A reverse logistics system consists of a series of activities such as: collection, cleaning, disassembly, testing and sorting, storage, transport and recovery operations. An integrated logistics system with forward and reverse flows can be represented as shown in Figure 6.1.

The nodes of the network represent forward or reverse activities; solid arrows represent forward flows whereas dashed arrows represent reverse flows between nodes.

The design and management of such an integrated network is more complex than that of traditional logistics networks limited to direct flows. Two factors cause these difficulties:

- the simultaneous existence and mutual impact of the two types of flow: the possible coordination /integration and interfering constraints between forward and reverse flows must be considered;
- the existence of numerous uncertainties about the return flows: choice of recovery options, quality of return objects, quantity, reprocessing time (Dekker and van der Laan, 1999; Jayaraman et al., 1999).

Reverse logistics systems can be classified into various categories depending on the characteristics that are emphasized thus several classifications may be found in the literature: Fleischmann et al. (1997); Rogers and Tibben-Lembke (1998); Beaulieu et al. (1999). If two important factors are considered, the types of return items([Fleischmann et al., 1997) and the main options of recovery (Thierry et al., 1995), four kinds of typical reverse logistics networks can be proposed as follows:

directly reusable network: return items (like pallets, bottles or containers) can be directly reused without major operations on them (only cleaning or minor maintenance). This is a closed-loop system because forward flows are closely associated with reverse flows.

remanufacturing network: products at the end of their life or needing maintenance (such as copy machines or aircraft engines) are returned and some parts or components are remanufactured to be used like new parts. This is also a closed-loop system, because remanufacturing is often implemented by the original producer.

repair service network: defective products (like durable products or electronic equipment) are returned and repaired in service centers. In this type of network, there are few links with the forward channel so it can be considered an open-loop system.

recycling network: raw materials (such as metal, glass and paper) are recycled and, as this operation is often carried out by specialized third parties, it can be considered an open-loop system. The collection and elimination of waste is also found in this category.

1.2 Earlier reviews, applications and case studies

In the area of transportation planning, Dejax and Crainic (1987) carried out a review of problems related to the transportation of empty equipments or vehicles , such as containers for reutilization, separately or jointly with the transportation of loaded containers. They classified problems and published work according to the hierarchical planning framework into strategic, tactical and operational problems. The survey focused on purely transportation problems without considering manufacturing activities.

Fleischmann et al. (1997) and Fleischmann (2001) published a review of quantitative models for reverse logistics. They discussed the various dimensions of the reverse logistics context and they analyzed works pertaining to reverse distribution, inventory control in systems with return flows and production planning with reuse of parts and materials.

Several case studies have been reported providing descriptions of reverse logistics organizations in companies as well as management and optimization methods.

De Brito et al. (2003) have published a review of case studies in reverse logistics. They have analyzed over 60 cases, pointing out the variety of real life situations, and have presented comparison tables explaining how reverse logistics activities are undertaken. They have made numerous propositions and pointed out research opportunities. They have identified four different themes for study:

- reverse logistics network structures,
- reverse logistics relationships,
- inventory management techniques,
- planning and control of recovery activities,

- information techniques for reverse logistics.

Gungor and Gupta (1999) provided a literature review on the different aspects of reverse logistics systems, in the context of Environmentally Conscious Manufacturing and Product Recovery (ECMPRO). Environmentally conscious methods suggest taking into account environmental factors when designing products: *design for recycling* (choosing materials better so that the process of material separation and recovery becomes more efficient) and *design for remanufacturing* (designing to disassemble more easily). These concepts have a direct impact on reverse logistics performance. Then they reviewed studies on material and product recovery methods, pointing out:

- the collection of returned products (reverse distribution),
- disassembly with two related problems: disassembly leveling (how far to disassemble) and disassembly process planning,
- inventory control,
- production planning.

In addition to the above reviews, several authors have described specific industrial practices, as well as definitions of processes and strategies of management. The study on copier recovery by Thierry et al. (1995) provides a complete description of the steps to be followed in a product recovery strategy. Clendenin (1997) has reported on a business process reengineering approach to optimize the reverse logistics channel at Xerox. Festinger (1998) and Rohlich (1999) have cited numerous examples of practices of reverse logistics in industry, focusing on the economic interest and pointing out the lack of management science support. Rogers and Tibben-Lembke (1998) presented a monograph introducing reverse logistics and practices in industry. Rogers and Tibben-Lembke (1999) showed the increasing importance of reverse flows and discussed the strategies to reduce the associated costs. Browne and Allen (1999) presented the package recovery activity in Great Britain. Fleischmann (2001) described the organization of the IBM Corporation for product recovery and spare parts management.

1.3 Goal of the chapter

In this chapter, we focus on the methodologies for the design and optimization of networks for logistics systems, including reverse flows, and particularly on the activities regarding the different levels of the hierarchical planning of such systems. We emphasize the specificities imposed upon logistics networks by the consideration of reverse flows and activities. Our approach is based upon the consideration of RL at the different levels of hierarchical planning of the logistics network.

At the strategic level of planning, it is necessary to take into account the recovery option during the design of the product, by including the design for recovery, as well as to consider the costs related to the direct and reverse channels. For the network design, one has to decide where to locate plants and warehouses, but also where to locate the different units for recovery (collection points and remanufacturing plants).

At the tactical level, return flows must be integrated within the overall activities: freight transport (by combining routes for the delivery and collection of returns), handling and warehousing, procurement (recycled parts are an alternative to the procurement of new parts), production planning and inventory management taking into account returns.

At the operational level, production scheduling and control related decisions, such as disassembly and reassembly operations, must be taken in relation to traditional production decisions, as well as the management of all forward and reverse flows for distribution and collection activities.

Section 2 of this chapter is devoted to the hierarchical planning concepts of logistics systems including reverse activities. Section 3 describes strategic planning methodologies for the design of a reverse logistics network. We make a distinction between the qualitative analysis of reverse logistics networks and quantitative methods based on mathematical models. Section 4 of the chapter presents methods for the management of reverse logistics systems at the tactical and operational levels. We discuss inventory management methods and flow optimization models. The final section contains our conclusions and general directions for future work in the area of reverse logistics.

2. Hierarchical planning of reverse logistics systems

The management of logistics networks and their activities is very complex because of their large dimensionality, wide variety of decisions of different scope, focus and time horizon, and disturbance factors (Harhalakis et al., 1992). However, the structure of decisions for such systems presents, in practice, a natural hierarchy of interconnections. According to Anthony's taxonomy (Anthony, 1965), these management decisions can be classified into three categories: strategic planning, tactical planning and operational planning decisions. Generally, these decisions at different categories (levels) are responsive respectively to the managerial functions at different echelons of the organization (Bitran and Tirupati, 1993; Dejax, 2001; Dupont, 1998). In the context of RL, such a hierarchical planning structure can still be employed while the consideration of reverse flows of return products is introduced into the system (Lu, 2003).

Strategic planning decisions are mostly concerned with the design of the network, the establishment of managerial policies and the development of resources to satisfy external requirements in a manner that is consistent with the organizational goals. Such decisions, which are made at fairly high managerial levels, involve large investments and have long-term implications (e.g., more than one year) (Bitran and Tirupati, 1993; Crainic and Laporte, 1997; Dejax, 2001). Typical decisions in the areas of logistics and production systems design consist of the location and sizing of new plants, storage and transfer facilities, acquisition of new equipment, selection of new product lines, and design of the transportation network (Owen and Daskin, 1998). At this level, data and decisions are highly aggregated (Hax and Meal, 1975; Boskma, 1982; Herrmann et al., 1994) on the basis of product types and long-term time horizons (at least three years). The location and sizing of facilities and the design of the network must take into account the impact of the reverse flows and activities and this problem depends on the level of integration of forward and reverse activities. Lu (2003) proposes a hierarchical framework for RL planning in which the strategic planning level covers:

- the design of the logistics network considering both forward and reverse flows: determination of the number and location of all types of logistics facilities, including plants, warehouses, distribution centers, etc., in the forward logistics channel, as well as the corresponding facilities in the reverse channel, e.g., collection and sorting centers, recovery centers, etc.
- the determination of the capacities/resources needed for all the respective facilities.
- the allocation of service areas to each facility for the distribution/collection activities.

The consideration of reverse flows introduces certain important specificities into the structure of logistics systems, and these aspects must be considered in the specific environment of RL applications (Fleischmann et al., 1997; Lu, 2003). Among these questions we can mention:

- What is the objective of a RL system? Who are the actors in the system and their responsibilities (analysis of reverse flow policy and application environment)? Should both forward and reverse activities be jointly or separately considered in the system? Which relationships are to be considered between the two channels and what will be their levels of integration?

- What are the specific functions to be covered all along the network channels?
- What are the relationships and respective importance of the economic and environmental factors and goals of the system in a specific application context?

Tactical planning decisions focus on the resource utilization process within the framework of the strategic plan. The basic problem to solve is the allocation of the resources determined at the strategic level, such as capacity, work force availability, storage and distribution resources to be effectively utilized. Taking the production/distribution capacities as constraints, the tactical planning function tries to establish a plan to meet the demand as effectively and profitably as possible. Typically, a tactical plan will be determined on periods of one month (covering the first part of the strategic planning horizon) over a medium-range planning horizon of up to a year. This timing permits the consideration of yearly seasonality in customer demand. The product structure will still be aggregated but at the level of product types or families only. Data is still aggregated, and decisions are sensitive only to broad variations in data and system parameters without consideration of the shorter term, or day to day information (Bitran and Tirupati, 1993; Crainic and Laporte, 1997; Dejax, 2001). In turn, tactical decisions will be used as a framework for the decision process at the operational level or for day-to-day operations. In the reverse logistics environment, the integration of forward and reverse flows at the production and distribution management levels of the network is important. The problem may be viewed as a combined mathematical model of aggregate production planning and flow optimization with the following features (Lu, 2003):

- the planning horizon depends on the seasonality of demand and should cover at least one cycle, while unit planning periods reflect the dynamic character of the system.
- the optimization of flows is done through a distribution network covering a multi-supply system and a multi-demand system.
- special constraints are designed for the coordination of forward flows and reverse flows.

Operational planning, or operations control, deals with short-term or day-to-day operational and scheduling problems to meet customer requirements within the guidelines established by the more aggregate plans of higher levels. It often requires disaggregation of information generated at higher levels and detailed planning of both forward and reverse flows (Vicens et al., 2001; Jörnsten and Leisten, 1995; Rogers, D.F. et al., 1991). The planning decisions are usually based on specific product items, and the time horizon usually covers a few weeks or days. Operational planning decisions act as a framework and lead to the real time management of operations.

The hierarchical planning process constitutes a framework for top to down decisions. Once the design of the system structure has been defined, the top strategic decisions impose the major constraints (location of facilities, available resources and capacities, network structure) to the tactical planning level (Schneeweifl, 1995); in turn, the tactical level determines a more detailed plan within these constraints, while keeping the consistency of the strategic decisions (Axsater, 1980; Bitran and Tirupati, 1993; Erschler et al., 1986; Mercé, 1987; Gfrerer and Zapfel, 1995). Normally, the basic physical structure of the system is considered stable throughout a rather long time period compared to the time horizon of the tactical plan, and thus the strategic plan could be viewed as static (i.e., covering a single long time period). However, applications of dynamic or multi-period strategic planning have been proposed in certain specific cases (Canel et al., 2001; Chardaire and Sutter, 1996; Melachrinoudis and Min, 2000). Conversely, the strategic planning should be undertaken while considering the impact from the tactical level by anticipation [Schneeweifl, 1995], and the parameters of data aggregation. The effectiveness and efficiency of activities at lower levels could be viewed as possible measures to evaluate the design of the system for the potential future evolution of its structure. Lu (2003) studies the correlation between strategic and tactical planning in the RL context and shows that the coordination of production and recovery processes is necessary to insure the consistency of decisions at the strategic and tactical levels of the hierarchical planning framework.

3. Strategic planning for the design of a reverse logistics system

The basic questions of strategic planning concern the organization and design of the network system, in which we need to address the constitution of flow channels and to identify the relationships between the actors. Such decisions are situated at the top managerial level and depend largely on the policies of the firm. The activities for recovering/reusing used products or materials bring a new complexity to the planning of logistics systems (Dekker and van der Laan, 1999).

As it is a new research field, the results published to date on the design problem of RL systems are rather scarce and isolated. We review here the works reported in the literature in two categories: qualitative analysis based on case studies and quantitative analysis based on optimization models.

3.1 Qualitative analysis of RL networks

In the literature, the analyses of RL networks by most authors are based on case studies in order to provide a clear view of the RL process. Such work can be found for example in Stock (1999), Festinger (1998), Rohlich (1999), Browne and Allen (1999), Rogers and Tibben-Lembke (1998, 1999).

In the design of RL systems, possible actors can be members of the forward channel, e.g., producers, distributors, retailers, logistics service providers, or special third parties, e.g., secondary material dealers, recovery facilities providers, or special reverse logistics operators (Rogers and Tibben-Lembke, 1999; Beaulieu et al., 1999). There are also multiple options of network type. A system in which returns are not sent back to the original producer can be classified as an *open-loop system* as opposed to a *closed-loop system* for the converse situation (Fleischmann et al., 1997). Thus, the enterprises have to clearly determine which kind of system they want to establish and which ones are feasible because, for the same products or activities, an enterprise could decide to build up a closed-loop recovery system (recovery in-house) or consign the relevant activities to a specific reverse logistics operator.

In the choice of recovery options (Thierry et al., 1993), there are usually numerous possibilities even for the same product. The decision may be influenced by different factors such as economic and ecological objectives, recovery technology and experience, possible volumes of return flows, demand for recovered products, legislation, types of reused parts or products and so on. A suitable economic and ecological evaluation method needs to be developed and applied to the design process of a reverse logistics network. After determining the process of recovery operations, an important problem of system design is to choose the locations of the facilities of all the recovery activities, e.g., collection, testing and sorting, and recovery centers.

Reverse logistics also leads to a greater need for cooperation with other partners in the logistics channel (Philipp, 1999). Based on common objectives and interests, it requires a greater exchange of information, joint recovery operations (e.g., some of the necessary technologies can come from the suppliers because they produce and supply the parts or components of the product), joint product design, and cooperation in the common market of recovered products or parts.

Fleischmann et al. (1997) discuss the new issues arising in the context of reverse logistics according to three aspects: distribution planning, inventory control and production planning. They briefly review the mathematical models proposed in the literature on these three topics, and point out that "*in a number of situations, the two flows (reverse and forward) cannot be treated independently but have to be considered simultaneously to achieve adequate planning, and it is the interaction of these two flows that adds complexity to the system involved.*" Meanwhile, handling the uncertainty in the system is also a major task in the planning of reverse activities. The authors conclude that efforts for further research in this area are needed, particularly about the influence of return flows on supply chain management.

Two categories of modeling of network design are identified by the authors, i.e., separate modeling of reverse flows, where only reverse activities (channels) are considered and integration of forward and reverse distribution, which considers two distributions simultaneously. For this second problem, the authors also remark that very few models have been proposed in the current literature.

Then, in Fleischmann et al. (2000a), the authors summarize the general characteristics of logistics networks for product recovery, based on an analysis of diverse examples of applications. They indicate that a RL network is not normally the symmetrical image of a traditional network, which means that new actors and new functions can be involved in such a system. According to their paper, the activities composing a typical network structure can be grouped into collection, inspection/separation, reprocessing, disposal and re-distribution. The main differences between a RL system and a traditional logistics system are the following:

- *The collection phase of the system characterizes a convergent structure, where reverse flows are converging from the disposal market to recovery facilities. Conversely, for the re-distribution phase, flows are diverging from recovery facilities to demand points in the reuse market (Ginter and Starling, 1978).*
- *The geographical distribution and volume of both supply and demand are considered as exogenous variables.*
- *The network structure may be more complex because of a multiple sequence of processing steps.*

They also conclude that a further analysis of the aspects characterizing different network types is worthwhile and that more mathematical models based on the specificities of reverse logistics systems are desirable.

Reverse logistics being a relatively new research field, the topics addressed and results obtained are still far from satisfactory (Jayaraman et al., 1999, Fleischmann et al., 1997). Sometimes, the definitions of concepts and denominations proposed by various authors are different even in this field because of the differences in focus of different researchers. Contrary to the numerous practices in other industrial fields, the theoretical results have not so far provided systematic methods and supports for these systems.

3.2 Quantitative analysis of design problems of RL systems based on mathematical models

Most of the studies found in the published literature suggest extending classical facility location models to support the analysis of design problems for RL systems. All works reviewed below, except for Marín and Pelegrín (1998) and Lu (2003), are devoted to specific application cases. We divide these models into three categories depending on the integration or not of forward and backward flows and the consideration of a weak or strong correlation between the two types of flow.

3.2.1 Independent modes for revers activities. We describe below several analyses of reverse logistics systems, independent of possible corresponding forward flows. Much of this work is based upon the development and experimentation of mixed integer programming models (MIP) and is either generic or devoted to a particular type of application.

Spengler et al. (1997) propose a sophisticated operations research model for recycling industrial by-products in the steel industry. The by-products arising from the production stage of steel have to be totally further treated, and a two-echelon location model (MIP) is formulated to help select the recycling process and to determine the locations and capacities of treatment facilities. Maximum facility capacities are given, and the amounts and sites of by-products are assumed known. The particularity of this model is the integration of the decisions of selection of the process chains (process technologies) into the facility location problem. The model has been applied successfully in the fields of recycling of waste and by-product management.

Barros et al. (1998) consider the problem of establishing an efficient sand-recycling network from construction waste, which is taken charge of by an independent syndicate of construction waste processing companies. The sieved sand from crushing facilities is separated into three categories of clean, half-clean and polluted sand at regional depots. The sand in the first two categories is stored at these depots to be reused in different projects, where half-clean sand has a use restriction. The polluted sand is shipped to treatment facilities and, after treatment, this sand can also be used to satisfy the demand for clean sand. Because the total volume of treated sand is assumed to be greater than the total demand in the system, the storage of processed sand at facilities is necessary. A capacitated three-echelon location model (MIP) with two different types of facility to locate (depot and treatment facilities) is developed to solve this problem. The solution proposed is a heuristic procedure based on linear relaxation, where the lower bound is derived from the linear relaxed problem adding classical valid inequalities and the upper bound is made by heuristics on the basis of iterative rounding rules. The results obtained by the proposed model for the sand problem in the Netherlands are also discussed in this paper.

Jayaraman et al. (1999) present a two-echelon capacitated location model (MIP) for solving the location problem of remanufacturing facilities, where the used electronic products are acquired at collection zones and then transported to remanufacturing facilities. After the value-added process of remanufacturing, the remanufactured products are sent back to serve the demands for these products at customers (same location as the collection zone). The storage of remanufactured products at facilities is assumed if not all of them are demanded. The modeled chain links the supply of used products and the demand for recovered products, both of which are assumed to come from the same customers. The forward production/distribution activities are not considered in this model. The resulting model is solved by the modeling package GAMS, and tested on a set of problems using industrial data.

Louwers et al. (1999) treat a network design problem for the collection, reprocessing, and redistribution of carpet waste in Europe. The carpet waste from different sources is transported to regional preprocessing centers and, after being sorted and separated there, the palletized homogenized materials are transported to different customers for further processing. All reprocessed carpets are assumed to be fully used or disposed of either by customers or at waste disposal units, thus there is no storage at reprocessing centers. The authors propose a continuous location model to determine appropriate locations and capacities of regional centers. An iterative procedure using a standard software package to calculate the capacities, numbers and locations of facilities is given. The modeling of the problem can be categorized as two-echelon with a single type of facility to locate.

Shih (2001) describes the design problem of a reverse logistics system for recycling electrical appliances and computers in Taiwan. The system structure involves collecting points, storage sites, disassembly/recycling plants, and the final disposal/reclaimed material market. The problem has been formulated as a three-echelon model (MIP) with three different types of facility to locate (storage sites, disassembly/recycling plants, and final disposal facilities). The modeled chain connects the disposal market and the reuse market. All the collected returns are recycled or disposed of. Although some numerical results are given, no solution method to the model is explicitly presented.

Lu (2003) presents generic MIP models for both the uncapacitated and capacitated two-echelon location problems with one type of facility to be located for recycling used products as reusable materials, which consists of a pure reverse channel. In such a system, used products are collected from customer zones and recycled as reusable materials at recycling centers, and then sent to a market for reusable materials. According to the parameters of the system, the author proposes the model for two distinct cases, depending on whether the quantity of recycled materials in the system is greater than the demand at the reuse market (case a) or not (case b). A solution algorithm based on Lagrangian heuristics is developed. The model is tested by numerical experiments.

Lu (2003) also studies the facility location problem in a particular case of a repair service network, in which three kinds of flow exist (returned failed products, repaired products and spare parts). Returned products (rotable parts or durable products) need repair or preventive maintenance and are sent back from customers to "service centers," and then returned to customers. The main features of this network and its differences with the traditional network are discussed: the demand for the repair service at customers can be satisfied by any of the available service centers; however, the repaired products at the service centers must be shipped back to the original customer; the requirement for replacing parts at service centers can be met by any of the available suppliers within the constraint of supply capacity. Two uncapacitated and capacitated location models (MIP) are proposed, and algorithms based on Lagrangian heuristics are developed for solving them. For numerical experiments, the author develops a data set consisting of 44 large French cities serving both as customers and as candidate locations for repair service centers, among which 10 cities are selected as suppliers. Through testing, the algorithm proves to have a good performance in terms of quality of solution as well as computing time.

All of the above models are devoted to the design of systems as a location-allocation problem in which only recovery activities are covered. They do not include "forward" activities. The system structure is a priori defined according to the specific application context, which is often represented as a chain to link two demand markets (supply of returns and demand for recovered products). One exception is Spengler et al. (1997). In this work, the reverse flows originate from the "forward" stage of production, but no other interaction with the "forward" channel is further considered in the model. On the other hand, multiple reprocessing stages along reverse channels are often included in these works. Thus, before the returned products enter the recovery facility, a preprocessing activity is required in most cases so as to select/store the recoverable returns from reverse materials. This system structure of multiple stages can be found in the works of Barros et al. (1998), Louwers et al. (1999) and Shih (2001). Even in Jayaraman et al. (1999), although a single recovery stage is included in the model, the authors claim that they made a simplification assumption. It should also be noted that both the supply of returns and the demand for recovered products confine the limits of the reverse chain at its two ends. Moreover, if there is exogenous control (Fleischmann et al., 2000a) on the volume of both supply and demand, then storage is generally necessary in the system, as in Barros et al. (1998), Jayaraman et al. (1999) and the case of recycling network (a) in Lu (2003). However, if the demand for recycled material in the system is assumed sufficient, storage at an intermediate node becomes unnecessary, e.g., in Louwers et al. (1999), Shih (2001), Spengler et al. (1997), and the case of recycling network (b) in Lu (2003).

3.2.2 Integrated models with weak correlation between forward and reverse flows. In comparison with the models in the preceding section, in Spengler et al. (1997), the reverse flows (by-products) generated from the production stage were proportionally related to the volume of production, but the authors considered the reverse activities as an independent system according to their specific application case. We present below some models designed for the simultaneous consideration of forward and return flows in those cases where these two types of flow are only loosely correlated.

Bloemhof-Ruwaard et al. (1996) describe a two-echelon location problem in which two types of facility (production plants and disposal units) need to be simultaneously located. Two types of flow are assumed to exist in the system and to be coordinated at the production stage: forward flows that distribute products to satisfy the demands of customers and waste flows that arise from production and are shipped to disposal units. The authors formulate the problem as a MIP model and try to minimize the total system costs as the sum of fixed costs and variable costs. Lower and upper bound procedures are proposed, in which linear relaxation and Lagrangian relaxations are used to generate the lower bounds. The model and algorithms are tested on generated test problems. The authors claim that the proposed model can be applied in practice to problems like locating feedstock breeding farms and manure processing plants.

Marín and Pelegrín (1998) formulate and analyze, from a purely theoretical point of view, a facility location problem that they name the Return Plant Location Problem. Here, primary products are transported to satisfy the demands of customers, and secondary products available at customer sites are sent back to plants. At the plants, the outbound quantities of flows are proportionally restricted to the inbound amounts. The problem is formulated as a MIP and both a Lagrangian decomposition based heuristic and exact solution method are developed. The results of applying the model and algorithms on test problems are given.

In Fleischmann et al. (2000b), after an analysis of the case studies published in the literature, the authors present a "generic" uncapacitated facility location model (MIP) for logistics network design in the reverse logistics context and discuss its differences with the traditional logistics setting. This model is formulated on the basis of a two-echelon forward chain and two-echelon reverse chain. The coherence of the two kinds of flow at the production facilities is ascertained. Two application examples of the model are illustrated, i.e., copier remanufacturing and paper recycling. The standard solver CPLEX is used to solve the problems.

Lu et al. (2004) present a strategic model for the facility location problem and network design in the general framework of combined forward and reverse flows, in the case of directly reusable products. The producers provide products of a single type shipped to distributors to satisfy their demand. In return, a supply of reusable materials, e.g., containers, bottles, pallets, handling equipment or packages, need to be shipped back from distributors to producers for reutilization, on the basis of economic and environmental considerations. The returns to the production sites will be directly reused in the process of production and forward transportation. The authors propose a capacitated and an uncapacitated location model (MIP) comprising a special linking constraint for the correlation of forward and reverse flows. A specific solution algorithm based on the Lagrangian heuristic technique is developed. Numerical experiments on a sizable example adapted from the OR-Library (http://www.ms.ic.ac.uk/jeb/orlib/) are conducted and their results are presented, including a discussion on the impact of the return flows on the facility locations.

. In all of the cases considered above, at the site of plant the return flows (or waste leaving for disposal units in the case of Bloemhof-Ruwaard et al., 1996) are assumed to be coherent with the flows of product shipped to customers. Two types of relationship between these two flows have been considered: *proportionally balanced flows* (the flows at the two ends of the production facility are proportional, as in Bloemhof-Ruwaard et al., 1996; Marín and Pelegrín, 1998); *unbalanced flows* or flows conditioned by an inequality constraint ("greater than") as in Fleischmann et al. (2000b) and Lu et al. (2004). It is important to note that, in these models, the decisions about "forward" and "reverse" flows are made simultaneously. However, the impacts of reverse flows on decisions like the location of production facilities are implicit, rather than explicit, restrictions. In fact, such a correlation relationship between the two types of flow is similar to the quantitative conservation of flow conditions at intermediate network nodes found in the classical multi-stage facility location problem.

3.2.3 Integrated models with strong correlation between forward and reverse flows. Crainic et al. (1989, 1993a) addressed the problem of facility location for the combined distribution of loaded containers and the collection and transfer of empty containers in a container transportation planning system. They proposed multilevel facility location models with inter-depot balancing constraints and a branch-and-bound based solution technique. Their work focused on transportation planning and not on manufacturing activities.

Lu (2003) discusses the problem of the design of an integrated production and remanufacturing system, in which new products are manufactured at production sites and distributed through the "forward channel" while used products (or their components or parts) are sent back through the "backward channel" to be recovered to meet the original quality standard at a relatively low cost. Such a recovery process is designed to be implemented "in-house" and integrated into the forward logistics process in a so-called closed-loop system, combining forward and reverse flows. At customer sites, there is a demand for products and supplies of used products ready to be recovered. Intermediate reprocessing centers are responsible for some necessary preprocessing activities, such as cleaning, disassembly, checking, sorting and so on, before the return products are shipped back to remanufacturing centers. Remanufacturing centers accept the checked returns from intermediate centers and are responsible for the process of remanufacturing. Producers are in charge of the "traditional" production to serve the product demands of customers together with the remanufacturing centers. Such a system is modeled as a MIP location model, comprising two echelons in the forward channel and three echelons in the reverse channel to decide simultaneously on three different types of facility to be located in the network (production facilities, remanufacturing centers and intermediate centers). Solution algorithms based on Lagrangian heuristics are developed, and numerical results are also presented from a large data set.

Like the models presented in the preceding section, this category of models integrates both forward and reverse channels and the decisions related to these two channels are made simultaneously. Normally, in these models the demands for "forward" products and supplies of "reverse" returns originate from the same market (customers) and therefore the flows in the system consist of a closed-loop. It is important to note that, in these models, the flows at the two ends of the production facilities are constrained by a quantitative relationship (e.g., *unbalanced*) and also some other explicit *restrictions* (impacts) from reverse flows are imposed on the decisions about the location, number and capacity of production facilities. For example, in Lu (2003), because part of the reverse flow can become a component of the forward flow after recovery of used products, a "greater than" unbalanced relationship has been imposed to constrain the two flows. Furthermore, the quantity of recovered products at a potential location site directly impacts on the necessity to locate another production facility at this site. In Fleischmann et al. (2000b) (see section above), two different types of facility, for initial production and product recovery respectively, are distinguished but their possible locations are unified to formulate the model under a certain hypothesis. Therefore, the influence of reverse flows on the location decisions of production facilities is not explicit.

3.2.4 Synthesis and future directions. Table 6.1 summarizes all the quantitative models we have reviewed in Section 3.2 with a description of their main characteristics. Except for Louwers et al. (1999), all the models shown are 0-1 mixed integer programming models, and can be viewed as extensions of the traditional facility location model by introducing the specificities of reverse logistics systems. The objective is to propose the location of facilities (forward and reverse) and the quantities of production, trans-shipment, disposal and storage at minimum total cost. In four works (Bloemhof-Ruwaard et al., 1996; Lu, 2003, in the cases of a directly reusable network and a remanufacturing network; Marín and Pelegrín, 1998; Fleischmann et al., 2000b), the authors consider simultaneously the forward and reverse activities in one single system and the ensured coherence of the two types of flow.

As can be seen in this section, a number of strategic models have been developed for facility location and logistics network design to take account of reverse logistics. Some of these models are purely devoted to the management of reverse flows and others integrate traditional production and forward flows for product distribution with forward flows and

Table 6.1. Summary of quantitative location-allocation models for RL systems

Authors	**Model type**[a]	**Stages/ Types of facility** ($S^r/F^r-S^f/F^f$)[b]	**Correlation**[c]	**Solution technique**	**Application**
Barros et al. (1998)	D/C MIP	3/2 – 0/0	No	Linear relaxation + heuristics	Recycling sand
Bloemhof-Ruwaard et al. (1996)	D/C MIP	1/1 – 1/1	Yes, weak (balanced)	Linear, Lagrangian relaxation	Breeding farm
Fleischmann et al. (2000b)	D/U MIP	2/1 – 2/2	Yes, weak (unbalanced)	Standard package CPLEX	Copier remanufacturing and paper recycling
Jayaraman et al. (1999)	D/C MIP	2/1 – 0/0	No	Standard package GAMS	Remanufacturing of electronic product
Louwers et al. (1999)	D/C *	2/1 – 0/0	No	Standard package E04UCF	Recycling carpet materials
Crainic et al. (1989, 1993b)	D/U MIP	2/1 – 2/1	Yes	Branch-and-bound	Container transport planning
Lu (2003) in the case of recycling	D/U/C MIP	2/1 – 0/0	No	Lagrangian heuristics	Generic model
Lu (2003) in the case of repair service	D/U/C MIP	2/1 – 0/0	No	Lagrangian heuristics	Generic model
Lu et al. (2004) in the case of direct reuse	D/U/C MIP	1/0 – 1/1	Yes, weak (unbalanced)	Lagrangian heuristics	Generic model
Lu (2003) in the case of remanufacturing	D/U/C MIP	2/1 – 1/1	Yes, strong (unbalanced)	Lagrangian heuristics	Generic model

[a] D/U/C: D represents a deterministic model; U stands for uncapacitated; C stands for capacitated; MIP: Mixed integer programming model; *: continuous location model.

[b] $S^r/F^r-S^f/F^f$: S^r and S^f represent the number of stages (echelons) structured in the system for forward and reverse channels respectively; F^r and F^f stand for the number of types of facility to be located related to forward and reverse channels respectively.

[c] Type of correlation considered between forward and reverse flows.

Table 6.1 (continued).

Authors	Model type	Stages/ Types of facility (S^r/F^r – S^f/F^f)	Correlation	Solution technique	Application
Marín and Pelegrín (1998)	D/U MIP	1/1–1/0	Yes, weak (balanced)	Lagrangian decomposition based heuristics	Generic model
Shih(2001)	D/C MIP	3/3–0/0	No	Not indicated	Recycling electrical appliances
Spengler et al. (1997)	D/C MIP	2/1–0/0	No	Standard package GAMS	Recycling of by-products in steel production

product recovery activities. Some of these models are aimed at specific applications and others are generic. The influence of reverse logistics activities on the overall system is discussed.

These works report the growing need for models in the area of logistics systems design for reverse logistics and applications in different sectors. A major difficulty arises from the uncertainty about the rate of product returns and the estimation of supply and the necessity to include uncertainty in these models. There is also a need for the development of more realistic models, capable of handling complex industrial cases, and efficient solution techniques for handling large cases. In this respect, some authors rely on standard MIP software packages and others have developed specific solution techniques for large problems with complex constraints.

4. Tactical and operational planning including reverse flows

The goal of tactical planning is to ensure an efficient use of resources over a medium time horizon (e.g., one year) within the framework of the strategic plan, allocating resources and optimizing the activities, forward and reverse flows and inventory levels throughout the logistics system at short (e.g., monthly) time periods. Tactical planning consists mainly in the determination of the allocation of activities at all levels of the system and time periods in order to ensure the overall goals of the strategic plan. It is necessary to meet the constraints of production, transportation and recovery and achieve a good control of storage levels at each period of the planning horizon to ensure an overall optimization over planning horizon.

Operational planning models aim at short-term optimization of activities such as production scheduling or vehicle routing. Very little work has been published in this area and some models are at the interface between tactical and operational planning. For this reason, we have included published operational planning models with tactical planning models in a single section of this review.

In this section, we distinguish work pertaining to inventory management and work pertaining to flow optimization.

4.1 Inventory management with reverse flows

Appropriate planning and control methods are required to integrate the return flows of used products into the producer material management. The major difficulty is due to the uncertainty of the timing, quantity and quality of return flows. A limited number of case studies have been published on this problem but, since the 1960's, researchers have proposed many quantitative models (Schrady, 1967). Inventory management in reverse logistics has been receiving growing attention in the past decade with the rise in concern for the environment. After a general discussion on inventory management with reverse flows, we discuss deterministic models and stochastic models.

We have selected some representative case studies on inventory management to introduce the subject and their content is subsequently outlined. Toktay et al. (2000) consider the inventory management at Kodak. For a specific product (single-use cameras), printed circuit boards can be bought from suppliers or remanufactured from camera returns. Rudi et al. (2000) study the returns of medical devices for the Norwegian National Insurance Administration, in order to control the purchase of new devices and to decide what to do with returns (refurbished or scrapped). Fleischmann (2001) presents the case of IBM Machines that can be refurbished or dismantled to recover valuable parts. van der Laan (1997) investigates the case of automotive exchange parts, where remanufactured parts are sold more cheaply than new ones.

Compared with a traditional inventory control system, inventory models with reverse flows have two main characteristics:

- an exogenous inbound flow,
- multiple supply options for serviceable stock.

The general structure of inventory control with reverse flows involves two distinct inventories as illustrated in Figure 6.2:

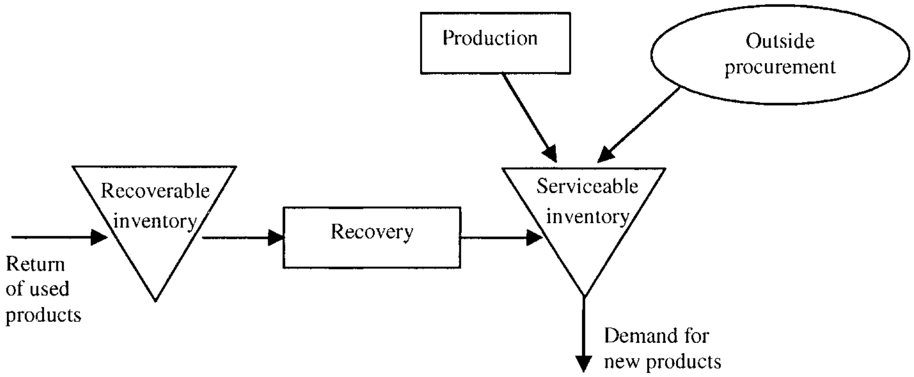


Figure 6.2. Recoverable inventory system (adapted from Fleischmann, 2001)

- the recoverable items returned from the market,
- the serviceable inventory supplies both from outside procurement and a recovery process.

However, specific models depend on the particular characteristics of the return activity. They may consider only one stock (single-echelon model) for the end-item stock, like in the case of directly reusable items, or be two-echelon models distinguishing recoverable inventory and serviceable inventory, as in the case of remanufacturing systems.

In the particular case of a recycling network, returned products are recycled as new raw materials and have no interaction with direct flows (open-loop system) so this can be considered as a classical inventory model.

The repair network is also a particular one. In certain situations, it may involve no closed-loop at all (when defective products are repaired in specialized service centers and then returned to the customer) or there may be no direct relation between demand and returns (for example, returns of used computer equipment for dismantling and demand for spare parts (Fleischmann, 2001)).

The well-known class of repairable item inventory models (where a return necessarily generates a simultaneous demand for a replacement item) has been investigated since the 1960's (see the classical METRIC model (Sherbrooke, 1968)). Many other works have been carried out on this particular subject, such as Pierskalla and Voelker (1976), Nahmias (1981), Cho and Parlar (1991), and Guide, Srivastava and Spencer (1997).

The cases of recycling networks and repair networks having been dealt with above, the rest of this section is devoted to the two other cases in our typology, i.e., directly reusable items and remanufacturing networks.

In a product recovery setting, the correlation between the two types of flow (forward and backward) tends to be much weaker and mainly reflects the dependence of returns on previous demand.

Models with multiple supply options can be a starting point to study inventory models with reverse flows. Moinzadeh and Nahmias (1988) present a two-supplier model, in a continuous review process. These models typically address the trade-off between a fast but expensive supplier and a cheaper, slower one (emergency supplies) and have the particularity that both the regular and emergency supplier are always available. In the recovery process, the cheaper channel (recovery) is often also the fastest one, availability of product returns for recovery is exogenously determined and the recovery supply mode is capacitated. In a recovery context, rather than lead time reduction, it is the restricted availability of the cheaper recovery channel that calls for an alternative supply source.

4.1.1 Deterministic models. Deterministic and static models including reverse flows are derived from classical EOQ models as proposed by Schrady (1967), where demand and returns are constant, lead times are fixed, and disposal is not allowed. Mabini et al. (1992) have extended the previous model by introducing a stockout service level constraint and considering the case of different items sharing the same repair facility. Ritcher (1996) studies the problem of lot-size coordination for production and remanufacturing in an EOQ framework. Teunter (2001) extends this work to the case of different holding costs for produced and recovered products.

Dynamic approaches based on the classical Wagner-Within model have also been proposed in the reverse logistics context. Beltran and Krass (2002) consider dynamic lot-sizing for an inventory point facing both demands and returns and controlled by procurement and disposal. They show that the complexity is increased by the inclusion of returns. Minner and Klerber (2001) also propose a dynamic model for an optimal control formulation in the case of a simple deterministic inventory system with a linear cost structure.

Most models consider linear cost functions. Dobos (2003) proposes a two-echelon system including reverse logistics, where holding costs for the two stores (serviceable and returns) are quadratic, as well as production, remanufacturing and disposal costs. He derives an optimal solution with two state variables (inventory status of the two stocks) and three control variables (production, remanufacturing and disposal rates) to minimize the total cost.

4.1.2 Stochastic models.

These are classified into single period, periodic review and continuous review systems. Two distinct strategies can be studied:

- the push strategy, where returned products are remanufactured as soon as a sufficient quantity of product is available. A manufacturing order is placed only when the serviceable inventory appears to be too low to satisfy future demand.
- the pull strategy, where returned products are remanufactured only when they are needed to satisfy the demand. If the remanufacturing output is too low, a manufacturing order is placed.

4.1.2.1 A single period review system.

This problem concerns products sold through e-commerce or mail sales. Returns may be used to satisfy new demand but only if they are available before the end of the selling period. Otherwise, returns can only be disposed of or sold at a lower price in a secondary market.

Vlachos and Dekker (2003) have extended the classical newsboy problem to set the optimal order quantity by taking into account the returns. They have shown that the classical newsboy solution is inadequate when return rates are significant. They propose different mathematical models for various return handling options depending on cost parameters, collection time and the need for recovery operations.

A related problem is studied by Mantrala and Raman (1999). The fashion goods market has become more unpredictable and competitive so retailers negotiate that suppliers buy back unsold inventory at the end of the selling season. They have studied how the retailer's optimal order quantity decisions are affected by demand uncertainty and how a supplier's returns policy can influence these decisions. Compared to previous works, they have introduced multiple retailer stores with possibly correlated demands.

4.1.2.2 Periodic review systems.

The first stochastic model was developed by Whisler (1967) who proposed an optimal control policy based on two parameters (disposal and new supply) for a single stock inventory with stochastic demand and return. Simpson (1978) extended this to a two-echelon model, where the optimal policy is controlled by three parameters (disposal, remanufacturing and new supply). These models consider no fixed cost and lead time. Inderfurth (1997) shows that the previous results are still valid if repaired and procurement channels have fixed and identical lead times and have zero fixed ordering costs.

Cohen et al. (1980) assume that a fixed fraction of products issued in a given period is returned after a fixed sojourn time in the market and can be reused. They propose an optimal periodic review "order up to" policy, without fixed costs and procurement lead time.

Kelle and Silver (1989) propose an extension of the previous research with a model for container reuse, considering fixed order costs and a stochastic sojourn time in the market. They transform the model into a dynamic lot-sizing problem.

Mahadevan et al. (2003) study a periodic review push policy where returns and demand follow a Poisson process. They develop several heuristics based on traditional inventory models tested by simulation, to determine when to remanufacture returns and how many new products to manufacture. They show that the problem can be reduced to choosing an appropriate order-up-to level. They also investigate the performance of the system as a function of return rates, backorder costs and lead times.

Kiesmüller (2003) analyses a stochastic recovery system with two stocking points, different lead times for production and remanufacturing, and no disposal of returned items. He proposes heuristic policies to make the decisions on production and remanufacturing quantities in a periodic review system, with a pull policy.

Kiesmüller and Scherer (2003) propose a method for the exact computation of parameters for an optimal periodic policy with two stocking points. They consider stochastic and dynamic demands and returns and equal deterministic lead times, taking into account production, remanufacturing, disposal backorder costs and also holding costs for serviceable and remanufacturable inventories. This exact method is very time-consuming and could not be used in practice. However, they propose two approximation methods and study their performance: a dynamic programming model and a deterministic model. They distinguish two cases: without stock of returns (corresponding to a push policy) and with stock of returns.

The study of a more general network has been carried out by Minner (2001). He studies a multi-stage supply chain with regular replenishment in which two types of return are considered: external product returns and internal by-products. This analysis aims to show how these additional external and internal material flows impact on the required amount of safety stock. These safety stocks depend on the service level and are required in relation to uncertain customer demand and returns.

4.1.2.3 Continuous review systems. Fleischmann et al. (2002) propose a model for a single stock point (directly reusable packages), extending a traditional single-item Poisson demand inventory model with a Poisson return flow of items. Procurement orders arrive after a fixed lead time and fixed order costs are considered. They propose an optimal (s, Q) policy for ordering new items and derive optimal values for the control parameters s and Q. They also discuss the impact of the return ratio on costs. Fleischmann and Kuik (2003) also consider a single inventory point system with independent stochastic demand and return items. They transform the model into a traditional (s, S) model without return flows and are able to propose an optimal (s, S) order policy.

van der Laan and Salomon (1997) investigate both continuous push and pull policies, with a disposal option, and analyze the influence of the return rate. van der Laan et al. (1999) evaluate the effects of the lead-time duration and variability on the total expected cost for push and pull control strategies in a system with two stocking points. They propose non-optimal strategies that are easy to implement and use in practice. Earlier, Muckstadt and Isaac (1981) studied a similar model but without considering stochastic manufacturing lead times nor holding costs for returns, and with no disposal possibility. They consider a single stocking point with a (s, Q) strategy to control the procurement.

4.1.3 Synthesis and future directions. Table 6.2 summarizes all the deterministic inventory management models reviewed in Section 4.1.1 with a description of their main characteristics.

Table 6.3 summarizes all the stochastic inventory management models reviewed in Section 4.1.2, except for single period models, with a description of their main characteristics.

As can be seen from these tables, a significant number of models of different types have been developed for inventory management including reverse flows in the deterministic as well as stochastic cases, considering one or two echelons. However, optimal solution techniques have been derived only under limiting assumptions, such as no lead times, equal lead times or no fixed costs. Problems with more realistic hypotheses have led to models that could only be solved approximately.

There is therefore a need for the development of more realistic models for inventory management in multi-echelon logistics channels. Furthermore, all models analyzed here consider single product inventory management. Models for multi-product inventory management should be developed to deal with the recovery of the various components of return products on the basis of their bill of materials. Actually, different components of a returned product may be remanufactured to be used again in different products.

Table 6.2. Synthesis of deterministic inventory models with reverse flows

Authors	Demand[a]	Model type[b]	Stocking points	Number of decision variables	Disposal option	Fixed costs	Lead time	Type of model
Schrady (1967)	C	S	2	3	No	Yes	Yes	EOQ
Mabini et al. (1992)	C	S	2	3	No	Yes	Yes	EOQ
Ritcher (1996)	C	S	2	4	Yes	Yes	Yes	EOQ
Teunter (2001)	C	S	1	4	Yes	Yes	Yes	Simulation
Minner & Klerber (2001)	C	D	2	3	Yes	No	No	Optimal
Beltran & Krass (2002)	D	D	1	2	Yes	Yes	No	Wagner-within
Dobos (2003)	C	S	2	3/5	Yes	Yes	No	Optimal quadratic

[a] D/C: discrete or continuous demand
[b] S/D: static or dynamic model

4.2 Flow optimization models in reverse logistics

After discussing inventory management models involving return products, we will now review tactical models pertaining to flow optimization in logistics networks involving return flows. We distinguish collection and distribution models and models involving disassembly or production activities.

4.2.1 Transportation planning models for product collection and distribution. This area concerns the collection and transportation of used products and packages, integrated or not with forward flows. Models dealing with forward and reverse distribution simultaneously consider the possible location of joint facilities (see strategic models) but there exist few models dealing with the combined routing of new products (forward flows) and return products (reverse flows) with a specific consideration for return flows.

In the area of transportation planning of containerized goods, Crainic et al. (1993b) proposed a multi-periodic and stochastic model for the

Table 6.3. Synthesis of stochastic inventory models with reverse flows

Authors	Demand/ return distribution[a]	Stocking points	Decision variables	Disposal option	Fixed costs	Lead time	Back order	Type of model
Periodic review models								
Whisler (1967)	Gen/gen	1	2	Yes	No	No	No	Optimal
Simpson (1978)	Gen/gen	2	3	Yes	No	No	Yes	Optimal
Cohen et al. (1980)	Gen/gen	1	1	No	No	No	No	Optimal
Kelle & Silver (1989)	Gen/gen	1	1	No	Yes	No	Yes	Dynamic lot-sizing
Inderfurth (1997)	Gen/gen	2	3	Yes	No	Yes	Yes	Optimal
Mahadevan et al. (2003)	Poisson/ Poisson	2	1	No	Yes	Yes	Yes	Heuristics/ simulation
Kiesmüller (2003)	Gen/gen	2	2	No	Yes	Yes	Yes	Heuristics
Kiesmüller & Scherer (2003)	Gen/gen	1/2	3	Yes	Yes	Yes	Yes	Exact method/ heuristics
Continuous review models								
Muckstadt & Isaac (1981)	Poisson/ Poisson	1	2	No	Yes	Yes	Yes	Non-optimal (s, Q) policy
van der Laan & Salomon (1997)	Poisson/ Poisson	2	4/5	Yes	Yes	Yes	Yes	Non-optimal (s, S) policy
van der Laan et al. (1999)	Poisson/ Poisson	2	3/4	No	Yes	Yes	Yes	Non-optimal (s, S) policy
Fleischmann et al. (2002)	Poisson/ Poisson	1	2	No	Yes	Yes	Yes	Optimal (s, Q) policy
Fleischmann & Kuik (2003)	Gen/gen	1	2	No	Yes	Yes	Yes	(s, S) policy

[a] Distribution law for direct demand and returns: general (gen) or Poisson

allocation of empty containers. Their model was aimed at the land transportation of maritime containers for international trade.

Del Castillo and Cochran (1996) studied production and distribution planning for products delivered in reusable containers. In this case, the return of empty containers was a constraint for the production system.

Duhaime et al. (2001) analyze the problem of reusable containers at Canada Post. They show with a minimum cost flow model that stockout can be avoided if containers are returned quickly from customers.

Feillet et al. (2002) studied the problem of the tactical planning of interplant transport of containerized products. They developed vehicle routing models with gains aimed at determining interplant circuits for the combined transport of containers loaded with parts, the return of empty containers and the positioning of empty trucks. They applied their models to a real case in the automotive industry. Lu (2003) proposes an integrated production/distribution linear model including both forward and reverse flows. The model is detailed for the case of directly reusable products and can be easily extended to the case of a remanufacturing network, both cases considering the two channels (forward and reverse) simultaneously. In the case of an open-loop system, such as a repair service network or a recycling network, the reverse channel is independent of the forward one because the actors involved are different. Therefore, the reverse channel can be studied independently, which makes the problem much simpler and allows the application of classical methods for distribution planning in logistics.

Vehicle routing problems considering forward and reverse flows belong to the classes of vehicle routing problems with backhauls (VRPB) (Goetschalckx and Jacobs-Blecha, 1989) or pick up and delivery problems (VRPPD) (Savelsbergh and Sol, 1995). After Nagy and Salhi (2004), we can distinguish three classes of problems involving deliveries from a depot or pick ups to a depot:

- **delivery first, pick up second problems,** where customers can be divided into two categories (linehauls and backhauls) and vehicles can only pick up goods after they have finished delivering all their load,
- **mixed pick ups and deliveries,** where linehauls and backhauls can occur in any sequence on a vehicle route,
- **simultaneous pick ups and deliveries,** where customers may simultaneously receive and send goods.

Reverse logistics problems belong primarily to the last two categories of problems because clients can receive their deliveries and simultaneously return their reusable packages or products to be recycled or remanufactured. In this context, the following works can be quoted.

Min (1989) and Halse (1992) are among the first to study the VRP with simultaneous pick up and delivery. Min et al. (1992) explore the VRPPD in the context of multi depot. Wade and Salhi (2002) propose a practical compromise between the classical VRPB and the mixed VRPB. They allow mixed linehauls and backhauls under particular conditions. Nagy and Salhi (2004) propose heuristic algorithms for single and multidepot VRPPD. Dethloff (2001) studies the vehicle routing problem in the context of reverse logistics and claims that it can be viewed as a VRP with simultaneous pick up and delivery. He proposes a heuristic construction procedure to solve this problem. In his thesis, Vural (2004) proposes a genetic algorithm based metaheuristic for the capacitated VRP with simultaneous pick up and delivery. Rusdiansyah and Tsao (2003) present an integrated approach for solving the period VRP with simultaneous delivery and pick up. They address the problem of a fleet of capacitated vehicles used for delivering products from the warehouse to retailers and collecting the reusable empty containers in the reverse direction. Retailers can be visited once or several times over the period. The problem consists in finding a compromise between inventory costs and travelling costs.

4.2.2 Production planning.

The applicability of traditional production planning and scheduling methods to product recovery systems is very limited due to uncertainties (in the quantity, timing and quality of returns) and the specificity of recovery activities (disassembly operations must be planned to fulfil the demand for components for production). New methodologies must be developed for this purpose, such as MRP techniques with a reverse bill of materials and scheduling problems incorporating disassembly activities.

The different cases of recovery activity induce different problems. In the direct reuse activity, no production process additional to the initial production of products is necessary, and the main problem is an inventory management problem with uncertainty.

For material recycling, new production processes are necessary to transform returned products into raw materials. However, it is more a technological problem than a management problem and conventional methods can be used to plan and control recycling operations.

In addition, we can mention Hoshino et al. (1995) who proposed a model for a recycling-oriented manufacturing system, which takes into account the collection of used products to be recycled as raw materials or sold to raw-material suppliers for re-production of raw materials or disposal. The model includes two objectives (maximizing the recycling rate and maximizing the total profit) and is solved by goal programming techniques.

The more complex situation is that of the remanufacturing problem. In this case, there is no predetermined sequence of production steps and the repair operations on components depend on the state of the product and can be known only after testing. Thus, it induces a high level of uncertainty. The coordination of several interdependent activities is necessary: the disassembly operation is a procurement source of various parts simultaneously. Furthermore, a capacity problem may arise if several parts require the same repair equipment.

This analysis has allowed us to identify two distinct subjects of research: disassembly leveling and planning, and production planning including reverse flows.

4.2.2.1 Disassembly leveling and planning. Disassembly activities take place in various recovery operations including remanufacturing, recycling and disposal. The first decision to be made concerns the selection of an appropriate disassembly level (to determine at which level of the bill of materials the product must be disassembled) and processing options, in order to minimize the cost of disassembly (compared to the value of recovered components) with respect to technical constraints. As the number of components increases, the number of alternatives for the disassembly process planning grows quickly. Therefore, this problem is a very important and complex one for disassembly activities. Gungor and Gupta (1999) have provided a detailed review of existing techniques in this field: graph theory and tree representation are frequently used to treat this problem but also branch-and-bound, goal programming, simulation, and neural network techniques.

Johnson and Wang (1995) proposed a model for determining an optimal disassembly sequence for a given product structure, using a network flow algorithm. Penev and de Ron (1996) describe a static cost comparison tool to determine an economic disassembly level and sequence of a single product. Meacham et al. (1999) extend these approaches to multi-product models involving fixed costs and common parts. Krikke et al. (1998) propose a stochastic model taking into account uncertainty. They introduce quality classes for the different components and assign a reuse option to each class. The method is based on two steps:

- optimization at the product level (determination of a multiple product recovery and disposal (PRD) strategy for every product type returned, depending on different objectives or constraints),

- optimization at the group level of different types of product: they assign for each product of a group a PRD strategy to optimize an overall objective.

The authors apply this algorithm to a business case (Krikke et al., 1999) of recycling discarded computer monitors, to determine optimal PRD strategies with given group recovery and disposal policies.

Two other industrial cases related to disassembly and the choice of recycling options are presented by Spengler et al. (1997): dismantling and recycling of end-of-life products and recycling of industrial by-products. The first problem concerns the evaluation of integrated dismantling and recycling strategies for domestic buildings in Germany. They propose a mixed integer linear optimization model for the evaluation of costs and interactions of dismantling and recycling, and for the determination of dismantling procedures, recycling techniques and reuse options. In the second case, they develop a decision support system for by-product management in the iron and steel industry. In these industries, certain dusts and sludges can be recycled into the production chain. Companies have to decide which recycling process to use, and if it is possible to cooperate with other companies by sharing recycling plants to reduce costs.

4.2.2.2 Production planning and scheduling involving return products. The particularity of reverse flows prevents the use of the traditional MRP method for production planning. It needs to be adapted by, for example, a reverse bill of materials.

Authors like Gupta and Taleb (1994) propose an MRP algorithm for scheduling disassembly, taking into account dependencies between different components of the same product. In Taleb and Gupta (1997), they extended it to a multi-product situation. Flapper (1994) considers a situation where components can be obtained by purchase or disassembly of old products, and proposes a simultaneous schedule for disassembly, repair and assembly operations. An interesting approach taking into account uncertainty has been proposed by Thierry (1997) through a simulation study to compare different MRP approaches. Guide, Kraus and Scrivastava (1997) also use simulation to evaluate different scheduling policies in a remanufacturing system and Guide, Srivastava and Spencer (1997) extend the analysis to capacity planning using the techniques of rough cut capacity planning. Veerakamolmal and Gupta (2000) propose an adaptation of the MRP technique to determine the number of components needed, that is to say the number of products to disassemble to fulfil the demand for components for remanufacturing, at minimal disassembly and disposal costs. Kongar and Gupta (2000) describe an integer goal programming model to determine the number and type of products to be disassembled in order to fulfil a demand for used components or subsets for remanufacturing. This tool is relevant to decision-makers because it suggests several plans obtained by varying the objective criterion. Spengler (2002) proposes a decision support system for electronic scrap recycling companies. The recovery of electronic scrap is a multistage process. The model proposed is a mixed integer linear program based on activity analysis. It provides a daily plan to determine the recycling schedule of scrapped products, the levels of disassembly, the allocation of reusable parts and modules for the use of producers or suppliers.

4.2.3 Synthesis and future directions. At the tactical and operational levels of planning, flow optimization models are of course complementary to inventory management models. We have categorized flow optimization models into collection and distribution planning models involving only transport activities and planning and scheduling models involving different types of production activity. In both categories, there are models where the treatment of return products is integrated with that of new products, and models where the two types of flow are disconnected. In Table 6.4 we present a synthesis of all the flow optimization models reviewed in the section.

There is a growing need for models integrating all relevant factors, i.e., inventory management, production or disassembly activities as well as transport activities for collection or distribution planning. There is also a need for specialized models adapted to realistic specific cases. As we pointed out regarding strategic models, an important uncertainty factor exists regarding the data of the return flows and models should therefore incorporate stochastic features or be robust regarding the uncertainty of the supply of return products.

5. General conclusions

Although some of the concepts of reverse logistics, such as the recycling of products, have been put into practice for years, it is only fairly recently that the integration of reverse logistics activities has been a real concern for the management and organization of logistics systems. In the past ten years or so, a significant amount of work has been published regarding the management of return flows, independent of or integrated with the management of flows of new products.

This chapter has been an attempt to summarize the work pertaining to the design, planning and optimization of logistics systems according to a classification primarily based upon the three steps of systems planning, i.e., strategic, tactical and operational planning. Strategic plan-

Table 6.4. Synthesis of flow optimization models

Authors	Model type[a]	Type of flow	Mono/multi period	Method	Application
Crainic et al. (1993b)	D	reverse	multi		Allocation of empty containers
Del Castillo & Cochran (1996)	P/D	combined	multi	LP + simulation	Soft drink
Duhaime et al. (2001)	D	combined	mono	Minimum cost flow model	Canada Post
Feillet et al. (2002)	D	combined	multi	VRP with gains	Interplant transport of containerized production
Lu (2003)	P/D	combined	mono	Lagrangian relaxation	Generic model
Hoshino et al. (1995)	P	combined	multi	Goal programming	Numerical example
Johnson & Wang (1995)	P	reverse	mono	Network flow algorithm	
Penev & De Ron (1996)	P	reverse	mono	Graph theory and cost analysis	
Meacham et al. (1999)	P	reverse	mono	Graph theory and cost analysis	
Krikke et al. (1998)	P	reverse	mono	Stochastic model	Computer monitors
Spengler et al. (1997)	P	reverse	mono	MIP - Benders	Demolition waste
Gupta & Taleb (1994)	P	reverse	multi	MRP algorithm	
Taleb & Gupta (1997)	P	combined	multi	heuristics	
Flapper (1994)	P	combined	multi	MRP heuristic	
Thierry (1997)	P	combined	multi	MRP, simulation	Copier remanufacturing
Guide et al. (1997)	P	reverse	multi	simulation	
Veerakamolmal & Gupta (2000)	P	reverse	multi	MRP	
Kongar & Gupta (2000)	P	reverse	mono	goal programming heuristic	Numerical example
Spengler (2002)	P	reverse	mono	MILP	Electronic scrap

[a] P/D : Production, Distribution

ning models have focused on facility location models for the design of a logistics network including return flows. Tactical and operational models have been developed regarding the various logistics activities where return flows should be considered, i.e., inventory management models, production planning and scheduling models and transportation planning models for the distribution or collection of products.

As a general conclusion, we must acknowledge the amount of work already published in this area and the effort of researchers in the design and optimization of realistic logistics networks by considering environmental concerns, customer service or simply economic efficiency. However, there is a growing need for new models corresponding to generic or specific cases, focused on the logistics activities of a given firm or on the overall supply chain.

A major difficulty in adequately handling RL activities concerns the uncertainty of the reverse flows themselves. This uncertainty involves numerous factors like the quantity and quality of returns, the selection of the recovery methods, the supply of return products as well as the demand for recovered products. These factors are not appropriately addressed in general in most of the published work. It seems worthwhile to examine the impacts of these uncertainties on the decision factors for the logistics systems at all levels of planning.

At the strategic level of planning, there is room for the extension of the proposed models to more general cases. For example, static simple facility location models might be extended to multi-level, multi-period dynamic models if the evolution of system structure is important in the planning horizon of a specific application. The integration of routing factors may also be interesting for decisions about the location of certain types of facility, like collecting centers in reverse logistics. Other extensions, such as introducing technology or supplier and remanufacturer selection in the supply chain network and economies of scale in recovery activities, are further important directions for future work. This can be compared with the extension, in recent years, of classical "forward" logistics systems to include supply chains.

At the tactical level, research directions should focus on the integration of features that appear only in the different models, such as inventory management, production and transportation planning. An extension of proposed models to include logistics factors ignored so far may be necessary; for example, set-up constraints for production or shipment, shipments of integer vehicle sizes and utilization of empty-ride transportation capacity.

The design of short-term operational planning models for reverse logistics has been quite limited so far. This might be due to the necessary specificity of operational models. However, there is a need to develop a close coordination between forward shipments and reverse returns at this level. Operational planning models including RL activities should therefore be more widely studied for different types of problem, such as remanufacturing and production scheduling and vehicle routing for the collection of returns.

A first step in the development of new models adapted to industrial needs might be to develop the analysis of case studies to get a better knowledge of real problems and practices. One should also pay particular attention to the necessary consistency and complementarity between the various types of model developed, particularly between the different levels of planning.

The analysis presented in this chapter and the directions of research which we have derived confirm the positive contribution of the consideration of reverse logistics for the development of efficient logistics and supply chain systems.

References

Anthony, R.N. (1965). *Planning and Control Systems: A Framework for Analysis.* Harvard University, Graduate School of Business Administration, Cambridge, MA.

Axsater, S. (1980). *Feasibility and Optimality of Aggregate Plans.* OR Report No. 167, North Carolina State University.

Barros, A.I., Dekker, R., and Scholten, V. (1998). A two-level network for recycling sand: a case study. *European Journal of Operational Research*, 110:199 – 214.

Beaulieu, M., Martin, R., and Landry, S. (1999). Logistique á rebours: un portrait nord-amèricain. *Logistique & Management*, 7:5 – 14.

Beltran, J.L. and Krass, D. (2002). Dynamic lot sizing with returning items and disposals. *IIE Transactions*, 34:437-448.

Bitran, G.R. and Tirupati, D. (1993). Hierarchical production planning. In: S.C. Graves, H.H.G. Rinnooy Kan, and P.H. Zipkin (eds.), *Logistics of Production and Inventory.* HandbooKs in Operations Research and Management Science, Volume 4, Elsevier Science Publishers B.V.

Bloemhof-Ruwaard, J., Salomon, M. Van Wassenhove, L.N. (1996). The capacitated distribution and waste disposal problem. *European Journal of Operational Research*, 88:490 – 503.

Boskma, K. (1982). Aggregation and the design of models for medium term planning of production. *European Journal of Operational Research*, 10:244 – 249.

Browne, M. and Allen, J. (1999). Récupération des déchets d'emballage en Grande-Bretagne: quelles implications logistiques? *Logistique & Management*, 7(2):37 – 43.

Canel, C., Khumawala, B.M., Law, J., and Loh, A. (2001). An algorithm for the capacitated, multi-commodity multi-period facility location problem. *Computers & Operations Research*, 28:411 – 427.

Chardaire, P. and Sutter, A. (1996). Solving the dynamic facility location problem. *Networks*, 28:117 – 124.

Cho, D.I. and Parlar, M. (1991). A survey of maintenance models for multi-unit systems. *European Journal of Operations Research*, 44(6):1 – 23.

Clendenin, J.A. (1997). Closing the supply chain loop: Reengineering the returns channel process. *International Journal of Logistics Management*, 8:75 – 85.

Cohen, M.A., Nahmias, S., and Pierskalla, W.P. (1980). A dynamic inventory system with recycling. *Naval Research Logistics Quarterly*, 27(2):289 – 296.

Crainic, T.G., Dejax, P., and Delorme, L. (1989). Models for multimode multicommodity location problems with interdepot balancing requirements. *Annals of Operations Research*, 18:279 – 302.

Crainic T.G., Delorme, L., and Dejax, P. (1993a). A branch-and-bound approach for the multicommodity location problem with balancing requirements. *European Journal of Operational Research*, 65:368 – 382.

Crainic, T.G., Gendreau, M., and Dejax, P. (1993b). Dynamic and stochastic models for the allocation of empty containers. *Operations Research*, 41:102 – 126.

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

De Brito, M.P. and Dekker, R. (2002). *Reverse Logistics — A Framework*. Econometric Institute Report EI 2002-38, Erasmus University, Rotterdam.

De Brito, M.P., Dekker, R., and Flapper, S.D.P. (2003). *Reverse Logistics — A Review of Case Studies*. ERIM Report Series Research in Management ERS-2003-012-LIS, Working paper of Erasmus University, Rotterdam.

Del Castillo, E. and Cochran, J.K. (1996). Optimal short horizon distribution operations in reusable containers. *Journal of the Operational Research Society*, 47:48 – 60.

Dejax, P. and Crainic, T.G. (1987). A review of empty flows and fleet management models in fleet transportation. *Transportation Science*, 21(4):227 – 247.

Dejax, P. (2001). Stratégie, planification et implantation du système logistique. In: J.-P. Campagne and P. Burlat (eds.), *Maîtrise et organisation des flux industriels*, pp. 129 – 160. Hermes, Lavoisier.

Dekker, R. and van der Laan, E.A. (1999). Gestion des stocks pour la fabrication et la refabrication simultanées : synthèse de résultats récents. *Logistique & Management*, 7, 59-64.

Dethloff, J. (2001). Vehicle routing and reverse logistics: the vehicle problem with simultaneous delivery and pickup. *OR Spectrum*, 23:79 – 96.

Dobos, I. (2003). Optimal production–inventory strategies for a HMMS-type reverse logistics system, *International Journal of Production Economics* 81-82:351 – 360.

Duhaime, R., Riopel, D. and Langevin, A. (2001). Value analysis and optimization of reusable containers at Canada Post. *Interfaces*, 31:3 – 15.

Dupont, L. (1998) *La gestion industrielle*. Hermes, Paris.

Erschler, J., Fontan, G., and Mercer, C. (1986). Consistency of the disaggregation process in hierarchical planning. *Operations Research*, 34(3):464 – 469.

Feillet, D., Dejax, P., and Gendreau, M. (2002). Planification tactique du transport de marchandises inter-usines : application au secteur automobile, *Journal Européen des Systèmes Automatisés*, 36(1):149 – 168.

Festinger, J.C. (1998). L'arrivée de la "reverse logistics." *Stratégie Logistique*, 6:32 – 54.

Flapper, S.D.P. (1994). Matching material requirement availabilities in the context of recycling: An MRP-1 based heuristic. In: *Proceedings of the 8th International Working Seminar on Production Economics*, pp. 511 – 519. Innsbruck, Austria.

Fleischmann, M. (2001). *Quantitative Models for Reverse Logistics*. Lecture Notes in Economics and Mathematical Systems, volume 501. Springer.

Fleischmann, M., Beullens, P., Bloemhof-Ruwaard, J.M., and Van Wassenhove, L.N. (2000b). The Impact of Product Recovery on Logistics Network Design, *Working Paper of the Center for Integrated Manufacturing and Service Operations*, INSEAD, 2000/33/TM/CIMSO 11.

Fleischmann, M., Bloemhof-Ruwaard, J.M., Dekker, R. van der Laan, E. van Nunen, J.A.E.E., and Van Wassenhove, L.N. (1997). Invited review, quantitative models for reverse logistics: A review. *European Journal of Operational Research*, 103:1–17.

Fleischmann, M., Krikke, H.R., Dekker, R., and Flapper, S.D.P. (2000a). A characterization of logistics networks for product recovery. *Omega*, 28:653–666.

Fleischmann, M. and Kuik, R. (2003). On optimal inventory control with independent stochastic item returns. *European Journal of Operational Research*, in press.

Fleischmann, M., Kuik, R., and Dekker, R. (2002). Controlling inventories with stochastic item returns: A basic model. *European Journal of Operational Research*, 138:63–75.

Fontanella, J. (1999). La logistique reverse, ou comment transformer le plomb en or. *Logistiques Magazine*, 141.

Gfrerer, H.and Zapfel, G. (1995). Hierarchical model for production planning in the case of uncertain demand. *European Journal of Operational Research*, 86:142–161.

Ginter, P.M. and Starling, J.M. 1978. Reverse distribution channels for recycling. *California Management Review*, 20(3):73–82.

Goetschalckx, M. and Jacobs-Blecha, Ch. (1989). The vehicle routing problem with backhauls. *European Journal of Operational Research*, 42:39–51.

Guide, V.D.R., Kraus, M.E., and Srivastava, R. (1997). Scheduling policies for remanufacturing, *International Journal of Production Economics* 48, 187-204.

Guide, V.D.R. and Srivastava, R. (1997). Repairable inventory theory: models and applications. *European Journal of Operational Research*, 102:1–20.

Guide, V.D.R, Srivastava, R., and Spencer, M.S. (1997). An evaluation of capacity planning techniques in a remanufacturing environment. *International Journal of Production Research*, 35:67–82.

Gungor, A. and Gupta, S.M. (1999). Issues in environmentally conscious manufacturing and product recovery: A survey. *Computers and Industrial Engineering*, 36:811–853.

Gupta, S.M. and Taleb, K. (1994). Scheduling disassembly. *International Journal of Production Research*, 32(8):1857–1866.

Halse, K. (1992). *Modelling and Solving Complex Vehicle Routing Problems*. Ph.D. thesis. Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby.

Harhalakis, G., Nagi, R., and Proth, J.M. (1992). *Hierarchical Modeling Approach for Production Planning*. Technical Research Report of Systems Research Center No. 14, University of Maryland.

Hax, A.C. and Meal, H.C. (1975). Hierarchical integration of production planning and scheduling. In: M.A. Geisler (ed.), *Studies in Management Sciences*, Vol. 1: Logistics. Elsevier, New York.

Herrmann, J.W., Mehra, A., Minis, I., and Proth, J.M. (1994). *Hierarchical Production Planning with Part, Spatial and Time Aggregation*. Technical Research Report of Systems Research Center No. 32, University of Maryland.

Hoshino, T., Yura, K., and Hitomi, K. (1995). Optimization analysis for recycle-oriented manufacturing systems. *International Journal of Production Research*, 33(8):2069–2078.

Inderfurth, K. (1997). Simple optimal replenishment and disposal policies for a product recovery system with leadtimes. *OR Spectrum*, 19:111–122.

Jayaraman, V., Guide, V.D.R, Jr., and Srivastava, R. (1999). A closed-loop logistics model for remanufacturing. *Journal of the Operational Research Society*, 50:497–508.

Johnson, M.R. and Wang, M.H. (1995). Planning product disassembly for material recovery opportunities. *International Journal of Production Research* 33(11):3119–3142.

Jörnsten, K., Leisten, R. (1995). Decomposition and iterative aggregation in hierarchical and decentralized planning structures. *European Journal of Operational Research*, 86:120–141.

Kelle, P. and Silver, E.A. (1989). Purchasing policy of new containers considering the random returns of previously issued containers. *IIE Transactions*, 21(4):349–354.

Kiesmüller, G.P. (2003). A new approach for controlling a hybrid stochastic manufacturing/remanufacturing system with inventories and different leadtimes. *European Journal of Operational Research*, 147:62–71.

Kiesmüller, G.P. and Scherer, C.W. (2003). Computational issues in a stochastic finite horizon one product recovery inventory model. *European Journal of Operational Research*, 146:553–579.

Kongar, E. and Gupta, S.M. (2000). A goal programming approach to the remanufacturing supply chain model. *Environmentally Conscious Manufacturing*, pp. 167–178. Proceedings of SPIE, Volume 4193.

Krikke, H.R., van Harten, A., and Schuur, P.C. (1998). On a medium term product recovery and disposal strategy for durable assembly products. *International Journal of Production Research*, 36(1):111-139.

Krikke, H. R., van Harten, A., and Schuur P.C. (1999). Business case Roteb: Recovery strategies for monitors. *Computers & Industrial Engineering* 36:739–757.

Louwers, D., Kip, B.J., Peters, E., Souren, F., and Flapper, S. D. P. (1999). A facility location allocation model for reusing carpet materials. *Computer & Industrial Engineering* 36:855–869.

Lu, Z. (2003). *Planification hiérarchisée et optimisation des systèmes logistiques avec flux inverses.* Ph.D. thesis. Université de Nantes.

Lu, Z., Bostel, N., and Dejax, P. (2001). Planification hiérarchisée des systèmes logistiques incluant la logistique inverse : problématique et modèles stratégiques. *Actes du 4e Congrès International de Génie Industriel* (GI2001), pp. 1141–1151. Aix-en-Provence–Marseille.

Lu Z., Bostel, N, and Dejax, P. (2004). The simple plant location problem with reverse flows. In: A. Dolgui, J. Soldek, O. Zaikin (eds.), *Suply Chain Optimization.* Kluwer Academic Publishers. In press.

Mabini, M.C., Pintelon, L.M., and Gelders, L.F. (1992). EOQ type formulations for controlling repairable inventories. *International Journal of Production Economics* 28:21–33.

Mahadevan, B., Pyke, D.F., and Fleischmann, M. (2003). Periodic review, push inventory policies for remanufacturing. *European Journal of Operational Research.* In press.

Mantrala, M.K. and Raman, K. (1999). Demand uncertainty and supplier's returns policies for a multi-store style good retailer. *European Journal of Operational Research*, 115:270–284.

Marín, A. and Pelegrín, B. (1998). The return plant location problem: Modeling and resolution. *European Journal of Operational Research* 104:375–392.

Meacham, A., Uzsoy, R.., and Venkatadri, U. (1999). Optimal disassembly configurations for single and multiple products. *Journal of Manufacturing Systems*

18(5):311 – 322.

Melachrinoudis, E. and Min, H. (2000). The dynamic relocation and phase-out of a hybrid, two-echelon plant/warehousing facility: A multiple objective approach. *European Journal of Operational Research* 123:1 – 15.

Mercé, C. (1987). *Cohérence des décisions en planification hierarchisée.* Ph.D. thesis. Université Paul Sabatier, Toulouse.

Min, H (1989). The multiple vehicle routing problem with simultaneous delivery and pick up points. *Transportation Research A*, 23:377 – 386.

Min, H, Current, J., and Schilling, D. (1992). The multiple depot vehicle routing problem with backhauling. *Journal of Business Logistics*, 13:259 – 288.

Minner, S. (2001). Strategic safety stocks in reverse logistics supply chains. *International Journal of Production Economics*, 71:417 – 428.

Minner S. and Klerber, R. (2001). Optimal control of production and remanufacturing in a simple recovery model with linear cost functions, *OR Spectrum*, 23:3 – 24.

Moinzadeh, K. and Nahmias, S. (1988). A continuous review model for an inventory system with two supply modes. *Management Science*, 34(6):761 – 773.

Muckstadt, J.A. and Isaac, M.H. (1981). An analysis of single item inventory systems with returns. *Naval Research Logistics Quarterly*, 28:237 – 254.

Nagy, G. and Salhi, S. (2004). Heuristic algorithms for single and multiple depot vehicle routing problems with pickups an deliveries. Forthcoming in *European Journal of Operational Research.*

Nahmias, S. (1981). Managing repairable item inventory systems: a review. *TIMS Studies in the Management Sciences*, 16:253 – 277.

Owen, S.H. and Daskin, M.S. (1998). Strategic facility location: a review. *European Journal of Operational Research*, 111, 423-447.

Penev, K.D. and de Ron, A.J. (1996). Determination of a disassembly strategy. *International Journal of Production Research*, 34(2):495 – 506.

Philipp, B. (1999). Reverse logistics : les formes adéquates de coopération pour la chaine logistique de valorisation des produits en fin de vie. Développements théroriques et approche de terrain, *Logistique and Management*, 7(2):45 – 57.

Pierskalla, W.P. and Voelker, J.A. (1976). A survey of maintenance models: the control and surveillance of deteriorating systems. *Naval Research Logistics Quarterly*, 23:353 – 388.

Ritcher, K. (1996). The extended EOQ repair and waste disposal model, *International Journal of Production Economics* 45(1-3):443 – 448.

Rogers, D.F., Plante, R.D., Wong, R.T., and Evans, J.R. (1991). Aggregation and disaggregation techniques and methodology in optimization. *Operations Research*, 39(4):553 – 582.

Rogers, D.S. and Tibben-Lembke, R.S. (1998). *Going backwards: Reverse logistics trends and practices.* Center for Logistics Management, University of Nevada, Reno, Reverse Logistics Executive Council.

Rogers, D.S. and Tibben-Lembke, R.S. (1999). "Reverse logistique": stratégies et techniques. *Logistique and Management*, 7(2):15 – 25.

Rohlich, Ph., (1999). Grande enquête, reverse logistique. *Stratégie Logistique*, 20:78 – 93.

Rudi, N., Pyke, D.F., and Sporsheim, P.O. (2000). Product recovery at the Norwegian National Insurance Administration. *Interfaces* 30:166 – 179.

Rusdiansyah, A. and Tsao, D. (2003). An integrated heuristic approach for the period vehicle routing problem with simultaneous delivery and pickup. In: *Second International Workshop on Freight Transportation and Logistics.* Odysseus 2003, Mondello, Italy, May 27 – 30.

Savelsbergh, M.W.P. and Sol, M. (1995). The general pick up and delivery problem. *Transportation Science*, 29:17 – 29.

Schneeweifl, Ch. (1995). Hierarchical structures in organizations: A conceptual framework. *European Journal of Operational Research*, 86:4 – 31.

Schrady, D.A. (1967). A deterministic inventory model for repairable items. *Naval Research Logistics Quarterly* 14:391 – 398.

Sherbrooke, C.C. (1968). Metric: A multi-echelon technique for recoverable item control. *Operations Research* 16:122 – 141.

Shih, L. (2001). Reverse logistics system planning for recycling electrical appliances and computers in Taiwan. *Resources, Conservation and Recycling*, 32:55 – 72.

Simpson V.P. (1978). Optimum solution structure for a repairable inventory problem. *Operations Research* 26:270 – 281.

Spengler, T. (2002). Management of material flows in closed-loop supply chains: Decision support system for electronic scrap recycling companies. In: *Proceedings of the 36th Hawaii International Conference on System Sciences* (HICSS'03).

Spengler, Th., Püchert, H., Penkuhn, T., and Rentz, O. (1997). Environmental integrated production and recycling management. *European Journal of Operational Research* 97:308 – 326.

Stock, J.R. (1999). Développement et mise en úuvre des programmes de reverse logistics, *Logistique & Management*, 7(2):79 – 84.

Taleb, K and Gupta, S.M. (1997). Disassembly of multiple product structures. *Computers and Industrial Engineering*, 32(4):949-961.

Teunter, R. (2001). Economic ordering quantities for recoverable item inventory systems, *Naval Research Logistics* 48:484 – 495.

Thierry M.C. (1997). *An Analysis of the Impact of Product Recovery Management on Manufacturing Companies*, Ph.D. thesis, Erasmus University, Rotterdam.

Thierry, M.C., Salomon, M., van Nunen, J.A.E.E., and Van Wassenhove, L.N. (1993). *Strategic Production and Operations Management Issues in Product Recovery Management*, Management Report Series No. 145. Erasmus University/Rotterdam School of Management.

Thierry, M.C., Salomon, M., van Nunen, J., and Van Wassenhove, L. (1995). Strategic issues in product recovery management. *California Management Review* 37:114 – 135.

Toktay, L.B., Wein, L.M., and Zenios, S.A. (2000). Inventory management for remanufacturable products. *Management Science*, 46:1412 – 1426.

van der Laan, E.A. (1997). *The Effects of Remanufacturing on Inventory Control*, Ph.D. thesis, Erasmus University, Rotterdam.

van der Laan, E. and Salomon, M. (1997). Production planning and inventory control with remanufacturing and disposal. *European Journal of Operational Research* 102:264 – 278.

van der Laan, E., Salomon, M., and Dekker, R. (1999). An investigation of lead-time effects in manufacturing/remanufacturing systems under simple PUSH and PULL control strategies. *European Journal of Operational Research* 115:195 – 214.

Veerakamolmal, P., S.M. Gupta, 2000. Optimizing the supply chain in reverse logistics. *Environmentally Conscious Manufacturing*, pp. 167 – 178. Proceedings of SPIE, Volume 4193.

Vicens, E., Alemany, M.E., Andrés, C. , and Guarch, J.J. (2001). A design and application methodology for hierarchical production planning decision support systems in an enterprise integration context. *International Journal of Production Eco-*

nomics, 74:5-20.

Vlachos, D. and Dekker, R. (2003). Return handling options and order quantities for single period products. *European Journal of Operational Research*, 151:38–52.

Vural, A.V. (2004). *A GA Based Meta-Heuristic for the Capacitated Vehicle Routing Problem with Simultaneous Pick Up an Deliveries*, M.Sc. thesis, Sabanci University, Turkey.

Wade, A.C. and Salhi, S. (2002). An investigation into a new class of vehicle routing problems with backhauls. *Omega*, 30:479 – 487.

Whisler, W.D. (1967). A stochastic inventory model for rented equipment. *Management Science*, 13(9):640 – 647.

Chapter 7

MODELS AND METHODS FOR OPERATIONS IN PORT CONTAINER TERMINALS

Kap Hwan Kim

Abstract Because container vessels spend a large portion of transportation time in ports, it is essential to improve the productivity of various handling activities in port container terminals. Also, because port construction requires a large amount of investment, it is important to efficiently utilize the internal resources of container terminals. This chapter introduces various operations in container terminals and decision-making problems that require support by scientific methods. Models and methods in previous researches are reviewed and classified according to their characteristics.

1. Introduction

Container terminals have been playing an important role in global manufacturing and international business as multi-modal interfaces between sea and land transport. The marine container industry has grown dramatically in the last 30 years. In order to increase the benefits of economy of scale, the size of containerships has significantly increased during the last decade. With increasing containerization, the number of container terminals and the competition among them have increased considerably. Issues related to container terminal operations have gained the attention of academic community only recently due to higher competition among container terminals. Many container terminals are attempting to increase their throughput and to decrease the turnaround times of vessels and customers' trucks.

In most existing container terminals, computers are employed to plan and control various handling operations. Because a container terminal is a complicated system with various interrelated components, there are many complicated decisions that operators or planners have to make. Because computer systems have capabilities to maintain a large amount of data and analyze it in a short time, they have been utilized to assist human experts during decision-making processes. This review focuses on the applications of operations research (OR) to these decision-making problems. There have been three similar review papers (Meersmanns and Dekker, 2001; Steenken et al., 2004; Vis and de Koster, 2003) on this topic.

1.1 Operations in port container terminals

The following introduces the operation of port container terminals (Park, 2003). Figures 7.1, 7.2, and 7.3 show quay cranes (QCs), a yard crane (YC), and a straddle carrier (SC), respectively. Container terminals usually have four different types of yard-side equipment: the on-chassis system, the carrier-direct system, the combined system of carrier and yard truck (YT) (straddle-carrier-relay system), and the combined system of yard crane and prime mover (transfer-crane-relay system). According to the different types of yard-side equipment, handing systems can be classified into two groups. One is called "the direct transfer system," which includes the on-chassis system and the carrier direct system, and the other is "the indirect transfer system," which includes the straddle-carrier-relay and the transfer-crane-relay systems. The two groups of systems are explained in more detail below.

In direct transfer systems, no yard cranes (YCs) are used. The same equipment is used to pick up (put down) a container from (into) the marshalling yard, deliver it to (from) the apron, and transfer it to (from) a quay crane (QC). In an on-chassis system which is illustrated in Figure 7.4, every container is stacked on a chassis and a tractor pulls the chassis between the apron and the marshalling yard. In a carrier-direct system which is illustrated in Figure 7.5, containers are stacked in multiple tiers and straddle carriers pick up (put down) containers from (into) the yard and deliver them between the apron and the marshalling yard.

In indirect transfer systems, a yard truck delivers a container between the apron and the marshalling yard. Straddle carriers or yard cranes transfer containers between yard trucks and yard stacks in the marshalling yard. Straddle carriers transfer containers in the straddle-carrier-relay system, while yard cranes do it in the transfer-crane-relay system which is illustrated in Figure 7.6.

In the following explanation on the operation of container terminals, we will assume the transfer-crane-relay system, shown in Figure 7.7, if the handling system is not indicated explicitly.

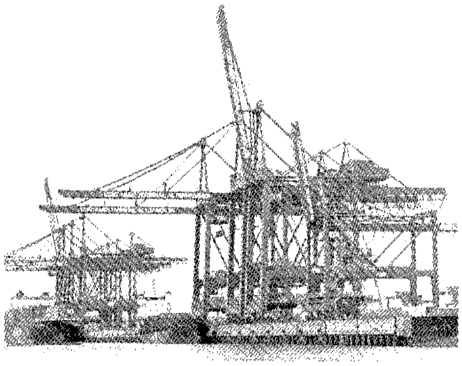


Figure 7.1. An illustration of quay cranes

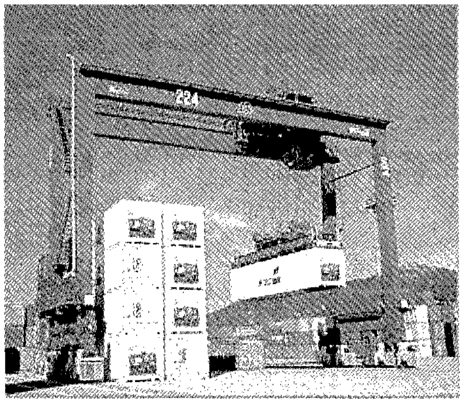


Figure 7.2. An illustration of yard crane

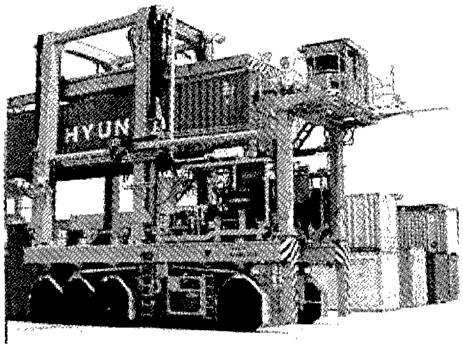


Figure 7.3. An illustration of straddle carrier

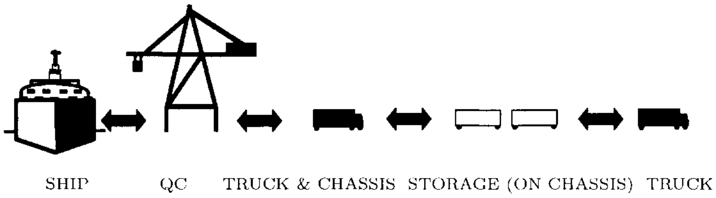


Figure 7.4. Container flows in an on-chassis system

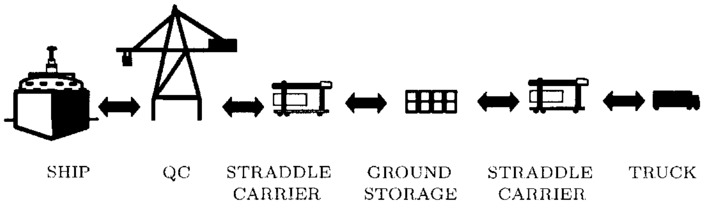


Figure 7.5. Container flows in a carrier-direct system

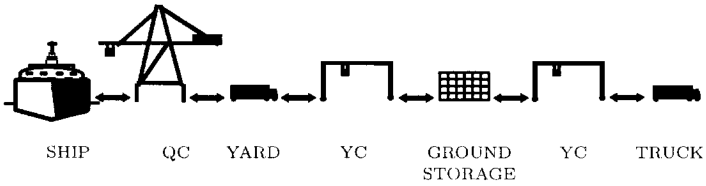


Figure 7.6. Container flows in a transfer-crane-relay system

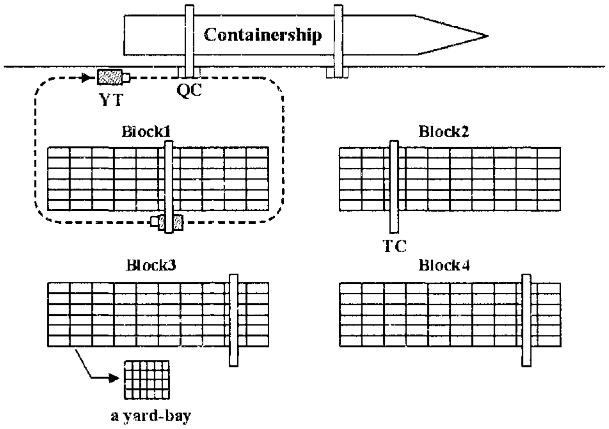


Figure 7.7. An example of a container terminal with a transfer-crane-relay system

The handling operations in container terminals include three types of operations: vessel operations associated with containerships, receiving/delivery operations for outside trucks, and container handling and storage operations in a yard. Vessel operations include the discharging operation, during which containers in a vessel are unloaded from the vessel and stacked in a marshalling yard, and the loading operation, during which containers are handled in the reverse direction of the discharging operation. During the discharging operations, QCs transfer containers from a ship to a prime mover. Then, the prime mover delivers the inbound (import/discharging) container to a yard crane that picks it up and stacks it into a position in a marshalling yard. For the loading operation, the process is carried out in the opposite direction.

During receiving and delivery operations, when a container arrives at a container terminal by an outside truck, the container is inspected at a gate to check whether all documents are ready and check for damages to the container. Also, at the gate, information regarding where an export container is to be stored and where an import container is located is provided to the outside truck. When the outside truck arrives at a transfer point of the yard, yard equipment, which can be a yard crane or straddle carrier, either receives a container from the truck, which is called the "receiving operation," or delivers a container to the truck, which is called the "delivery operation."

1.2 Operation plans in container terminals

Before handling operations in container terminals actually happen, planners in the container terminal usually plan them in advance to maximize the efficiency of the operations. Target resources for the planning process are usually limited in their capacities and thus priorities among handling activities that require the resources must be determined through the planning process. The resources include berths, QCs, yard cranes, other handling equipment, yard spaces, and human operators.

Ship operation planning. The planning process of ship operations consists of berth scheduling, QC scheduling (in practice, called work scheduling), and discharge and load sequencing. During the process of berth scheduling, the berthing time and position of a containership are determined (Figure 7.8), which are represented by the location of the corresponding rectangle on the time-berth space. Through the QC scheduling process, the sequence of ship-bays that each QC will serve and the time schedule for the service are specified (Table 7.1).

For QC scheduling, planners are usually given information such as a stowage plan of the ship — which is illustrated in Figure 7.9 — and

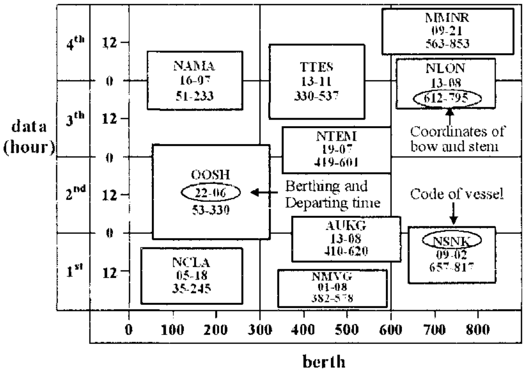


Figure 7.8. An example of a berth schedule (Kim and Moon, 2003)

Table 7.1. An example of a QC work schedule (Kim and Park, 2004)

Quay Crane Work Schedule													
QC 1 (operation time: 09:00~12:00)							**QC 2 (operation time: 09:00~12:00)**						
Operation seq.	Cluster number	Location of task	Type of task	No. of containers	Start time	Finish time	Operation seq.	Cluster number	Location of task	Type of task	No. of containers	Start time	Finish time
1	6	1 Hold*	D**	47	09:00	09:47	**1**	7	3 Deck	D	39	09:00	09:39
2	1	1 Hold	L**	42	09·47	10:29	**2**	9	5 Hold	D	46	09:41	10:26
3	8	3 Hold	D	32	10:31	11:03	**3**	5	5 Hold	L	23	10:26	10:49
4	4	3 Hold	L	8	11:03	11:11	**4**	10	7 Deck	D	24	10:51	11:14
5	3	3 Deck	L	8	11.11	11:19							
6	2	3 Deck	L	16	11:19	11:35							

* This represents " the hold of ship-bay 1"
** D (discharging), L (loading)

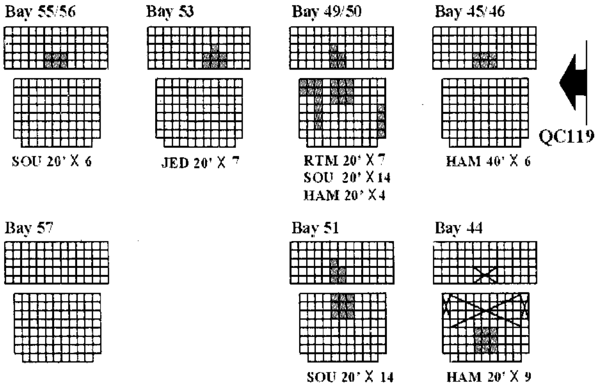


Figure 7.9. A partial example of a stowage plan

Table 7.2. An example of a load sequence list (Kim et al., 2004)

QC number	Seq.	Container number	Location in yard	Location in vessel
101	1	MFU8408374	2C-06-01-03	05-07-01
101	2	DMU2975379	2C-06-01-02	05-08-01
101	3	DMU2979970	2C-06-01-01	05-07-02
101	4	OLU0071308	2C-06-02-03	05-08-02
101	5	MTU4015162	2C-06-02-02	05-07-03
⋮	⋮	⋮	⋮	⋮

the time interval in which each QC is available. The stowage plan in Figure 7.9 consists of four cross-sectional views, each corresponding to a ship-bay. Each small square represents a slot. Shaded squares correspond to slots that containers must be loaded into in this container terminal. The shaded pattern in each slot represents a specific group of containers to be loaded into or picked up from the corresponding slots.

After constructing the QC schedule, the sequence of containers for discharging and loading operations is determined. Table 7.2 illustrates a load sequence list. The fourth column represents the storage location of a container before loading and the fifth column shows the slot in the vessel, which the container should be loaded into.

Yard space planning. One of the important factors that affect the turn-around time of vessels is the method of allocating storage spaces for containers arriving at the marshalling yard. The space allocation is the pre-assignment of storage spaces to the containers of each vessel arriving in the future so that the loading/discharge operation can be performed efficiently. In general, to expedite the loading operation, spaces for containers bound for the same ship should be assigned to locations close to one another. Storage spaces for inbound containers are usually determined in real time at the moment of discharge.

Equipment assignment planning. The equipment assignment process is the allocation of handling tasks to container-handling equipment. Loading/discharging tasks are assigned to one of the QCs, based on the berth schedule and number of loading/discharging tasks for each vessel. Transfer tasks are assigned to yard cranes dynamically, based on real-time information on waiting tasks and the status of each crane.

There are two types of strategies for assigning delivery tasks to prime movers. One is a dedicated strategy and the other is a pooled strategy. In the case of the dedicated strategy, a group of prime movers is assigned to a QC and deliver containers only for that QC. In the pooled strategy, all the prime movers are shared among different QCs and thus any prime mover can deliver containers for any QC, which is a more flexible strategy for utilizing prime movers.

The next section will introduce mathematical models for berth scheduling and QC scheduling. Section 3 discusses models and algorithms for the stowage planning and sequencing. Section 4 introduces previous studies on handling and storage activities in the yard. Section 5 introduces mathematical models and algorithms for real time dispatching various types of equipment in the yard. Some concluding remarks are given in the conclusion section.

2. Berth/quay crane scheduling

The berth is the most important resource in port container terminals because the construction cost is the highest among all cost factors for container terminals. Berth scheduling is the process of determining the time and the position at which each arriving vessel will berth. Quay crane scheduling is the process of determining the vessel that each QC will serve and the time during which the QC will serve the assigned vessel. The berth schedule and the QC schedule are inter-related because the number of QCs to be assigned to a vessel affects the berthing duration of the vessel. Despite this inter-relationship, because of the complexity of the integrated problem, most studies have decomposed the problem into two independent problems except the study by Park and Kim (2003).

A berth is just a structure along the water, thus can be considered a continuous line (continuous berth) which vessels with limited lengths can share with each other. However, most researchers have treated berths as discrete resources (discrete berths) that can be allocated to vessels.

In the following, the mathematical model suggested by Park and Kim (2002) is introduced. The most important objective of berth scheduling is to complete the ship operation within the due time which was pre-specified by a mutual agreement between the ship carrier and the terminal operator. Also, because outbound containers for a vessel may already be stacked in the marshaling yard, there is a most preferable berthing position for a vessel. Thus, Park and Kim minimized the costs resulting from the delayed departures of vessels and the additional handling costs resulting from deviations of the berthing position from the best location.

The following notations were used.

$L =$ the length of the berth.
$l =$ the number of vessels.

$p_i =$ the best berthing location of vessel i. This location is represented by the x-coordinate of the left-most end of the vessel and determined by considering the distribution of containers already arrived or a designated location for a specific vessel. The reference point for x-coordinate is the left-most boundary of the berth.

$a_i =$ the expected arrival time of vessel i.

$b_i =$ the ship operation time required for vessel i. This value includes the required allowance time between the departure of a vessel and the berthing of another vessel.

$d_i =$ the requested departure time of vessel i.

$l_i =$ the length of vessel i. This value includes the required gap between adjacent vessels.

$x_i =$ the berthing position of vessel i (a decision variable).

$y_i =$ the berthing time of vessel i (a decision variable).

$$z_{ij}^x = \begin{cases} 1, & \text{if vessel } i \text{ is located to the left-hand side of vessel } j; \\ 0, & \text{otherwise.} \end{cases}$$

$$z_{ij}^y = \begin{cases} 1, & \text{if vessel } i \text{ is scheduled before vessel } j; \\ 0, & \text{otherwise.} \end{cases}$$

$c_{1i} =$ the additional travel cost (per one grid-width) for delivering containers to vessel i resulting from non-optimal berthing locations.

$c_{2i} =$ the penalty cost (per one grid-length of time) of vessel i resulting from a delayed departure after the requested due time.

Figure 7.10 illustrates the relationships between variables and input data. Given these relationship, the objective function of the berth scheduling problem can be written as follows:

$$\text{Min} \sum_{i=1}^{l} \{c_{1i}|x_i - p_i| + c_{2i}(y_i + b_i - d_i)^+\}, \tag{7.1}$$

where $x^+ = \max\{0, x\}$.

The first term of the objective function comes from the deviation of the berthing position from the best location and the second term is related to the penalty cost from the delay of the departure of vessels after the requested departure time.

Let $|x_i - p_i|$ be α_i^+ when $x_i - p_i \geq 0$ and α_i^- when $x_i - p_i < 0$. And let $(y_i + b_i - d_i)$ be β_i^+ when $y_i + b_i - d_i \geq 0$ and β_i^- otherwise. Then, the berth scheduling problem can be formulated as follows:

$$\text{Min} \sum_{i=1}^{l} \{c_{1i}(\alpha_i^+ + \alpha_i^-) + c_{2i}\beta_i^+\} \tag{7.2}$$

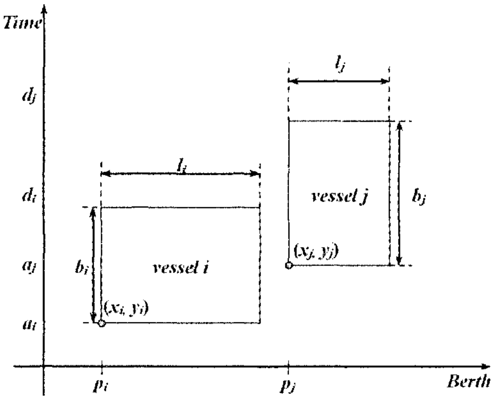


Figure 7.10. An illustrative example of a berth schedule

subject to

$$
\begin{array}{lll}
x_i - p_i = \alpha_i^+ - \alpha_i^- & \text{for all } i, & (7.3) \\
y_i + b_i - d_i = \beta_i^+ - \beta_i^- & \text{for all } i, & (7.4) \\
x_i + l_i \leq L & \text{for all } i, & (7.5) \\
x_i + l_i \leq x_j + M(1 - z_{ij}^x) & \text{for all } i \text{ and } j,\ i \neq j, & (7.6) \\
y_i + b_i \leq y_j + M(1 - z_{ij}^y) & \text{for all } i \text{ and } j,\ i \neq j, & (7.7) \\
z_{ij}^x + z_{ji}^x + z_{ij}^y + z_{ji}^y \geq 1 & \text{for all } i \text{ and } j,\ i < j, & (7.8) \\
y_i \geq a_i & \text{for all } i, & (7.9) \\
\alpha_i^+, \alpha_i^-, \beta_i^+, \beta_i^-, x_i \geq 0 & \text{for all } i, & (7.10) \\
\text{and } z_{ij}^x, z_{ij}^y & 0/1 \text{ integer for all } i \text{ and } j,\ i \neq j. & (7.11)
\end{array}
$$

Constraints (7.3) and (7.4) are related to the definition of α_i^+, α_i^-, β_i^+, and β_i^-. Constraint (7.5) implies that the position of the right-most end of vessel i is restricted by the length of the berth. Constraint (7.6) or (7.7) is effective only when z_{ij}^x or z_{ij}^y equals one. Constraint (7.8) excludes the case that two vessels are in conflict with each other with respect to the berthing time and the berthing position. That is, constraint (7.8) excludes $z_{ij}^x + z_{ji}^x + z_{ij}^y + z_{ji}^y = 0$ in which case the rectangles representing schedules for vessels i and j overlap with each other. Constraint (7.9) implies that a vessel cannot berth before it arrives. Problems (7.2) - (7.11) can be solved by mixed integer programming software. In a Korean container terminal with a berth of 1.2 *km*, the length of the planning horizon is usually one week during which 25 vessels are scheduled on average.

Lai and Shih (1992) proposed four berth allocation rules for discrete berths and compared the performance of the allocation rules by a simulation study. Brown et al. (1994) and Brown et al. (1997) addressed the discrete berth allocation problem which allows berth shifts before the completion of the unloading and loading operation for vessel.

Li et al. (1998) considered the berth-scheduling problem to be a scheduling problem for a single processor (berth) that can simultaneously perform multiple jobs (vessels). Based on the similarity of the problem to the bin-packing problem, they suggested various algorithms based on First-Fit-Decreasing (FFD) heuristics.

Guan et al. (2002) defined the berth-scheduling problem as a scheduling problem in which the processors are arranged along a straight line, and each job (vessel) requires simultaneous processing by multiple consecutive processors (QCs). They proposed a heuristic algorithm for minimizing the total weighted completion time of the vessels. Because this paper firstly defined the berth-scheduling problem in the context of general scheduling problems, it is worthwhile to introduce the definition here. Ships can be considered as jobs $\{J_1, J_2, \ldots, J_n\}$. The berth is represented as a straight line with equally-spaced m QCs which correspond to m parallel processors in the scheduling terms. Each job J_i has a given processing time p_i, a given weight w_i, and a given size s_i, where p_i, w_i, and $s_i \in Z^+$ and $s_i \leq m$. Each job, J_j, has to be processed by s_j consecutive processors simultaneously at any moment in time. The objective is to assign processors to jobs and to schedule the jobs so that the total weighted completion time is minimized.

Imai et al. (1997, 2001), Nishimura et al. (2001), and Imai et al. (2003) addressed discrete berth allocation problems with various objective functions and proposed various solution methods including sub-gradient optimization, heuristic methods, and genetic algorithms.

Although all the above studies considered berths as discrete resources that can be allocated to vessels, some studies have considered the berth as a continuous line that multiple vessels can share with each other at the same time. Thus, when the berth is considered as a continuous line, at a berth of the same length, more vessels can be served simultaneously if they are shorter in length. However, when the berth is considered as a collection of discrete berthing locations, the number of vessels to be served simultaneously is fixed without consideration of their lengths.

For the first time, Lim (1998) proposed an analytical model in which a berth is considered to be a continuous line. He discussed how to minimize the sum of the lengths of vessels that are supposed to berth at the same time by optimally locating their berthing positions. As introduced previously, Park and Kim (2002) solved the continuous-berth-scheduling problem by using the subgradient optimization technique. Kim and Moon (2003) solved the same problem by using the simulated annealing technique. Guan and Cheung (2004) proposed several heuristic search algorithms for the problem suggested by Park and Kim (2002). Heyden and Ottjes (1985) and Suh and Lee (1998) introduced their experiences of software development, related to the berth scheduling problem.

QC allocation and scheduling. Another important resource in container terminals are QCs. In practice, the QC schedule is usually devised by operation planers as a part of a ship operation plan. Studies on the QC scheduling differ from each other in the degree of detail.

Park and Kim (2003) proposed a method for scheduling berths and QCs, simultaneously. Note that as the number of QCs assigned to a vessel increases, the berthing time of the vessel decreases. In their study, the QC schedule specifies only the starting time and the ending time that each QC serves a specific vessel. However, Daganzo (1989) and Peterkofsky and Daganzo (1990) specified the starting time and the ending time that each quay crane serves a specific bay in a vessel. However, neither study considered the interference among QCs and the dynamic arrival of vessels.

Kim and Park (2004) assumed a given service-time window during which each QC is scheduled to serve a vessel, and proposed a scheduling method for QCs to perform unloading and loading operations for the vessel. They considered the detailed movement of QCs and interference between adjacent cranes.

The following introduces the model proposed by Kim and Park (2004). Table 7.1 illustrates a QC schedule. A "task" is defined as a discharging or loading operation for a cluster. It is assumed that once a QC starts to load (or discharge) containers into (from) a cluster of slots, it continues to do so until all the slots in the cluster become filled (empty). Therefore, they considered handling work for a cluster to be a task. When discharging and loading operations must be performed at the same ship-bay, the discharging operation must precede the loading operation. When a discharging operation is performed in a ship-bay, tasks on a deck must be performed before tasks in the hold of the same ship-bay are performed. Also, the loading operation in a hold must precede the loading operation on the deck of the same ship-bay. Thus, there are precedence relationships among clusters, relationships that must be observed during a ship operation. Also, it should be noted that QCs travel on the same track. Thus, certain pairs of tasks cannot be performed simultaneously when the locations of the two clusters corresponding to the tasks are too close to each other, because two adjacent QCs must be apart from each other by at least one ship-bay so that they can simultaneously perform their tasks without interference. Also, if containers for any two tasks must be picked up at or delivered to the same location in a yard, the two tasks may not be performed simultaneously, because doing so will cause interference among yard cranes that transfer containers corresponding to the two tasks. The constraints in the scheduling operations of QCs are shown below.

(1) Each QC can operate after its earliest available time.
(2) QCs are on the same track and thus cannot cross each other.
(3) Some tasks must be performed before others.
(4) There are some tasks that cannot be performed simultaneously.

The following notations are used for a mathematical formulation.

Indices.

i, j: Tasks to be performed. Tasks are ordered in an increasing order of their relative locations in the direction of increasing ship-bay numbers.

k: QCs where $k = 1, \dots, K$. QCs are also ordered in an increasing order of their relative locations in the direction of increasing ship-bay numbers.

Problem data.

p_i: The time required to perform task i.

r_k: The earliest available time of QC k.

l_i: The location of task i (expressed by the ship-bay number).

l_k^0: The starting position of QC k expressed by a ship-bay number.

l_k^T: The final position of QC k expressed by a ship-bay number. This final position may be specified if a job for the next ship is already assigned to QC k.

t_{ij}: The travel time of a QC from location (l_i) of task i to location (l_j) of task j. t_{0j}^k and t_{jT}^k respectively represent the travel time from the initial position (l_k^0) of QC k to location (l_j) of task j, and from location (l_j) of task j to the final destination (l_k^T) of QC k.

M: A sufficiently large constant.

α_1: The weight for the makespan (the maximum completion time).

α_2: The weight for the total completion time.

Sets of indices.

Ω: The set of all tasks.

Ψ: The set of pairs of tasks that cannot be performed simultaneously. When tasks i and j cannot be performed simultaneously, $(i, j) \in \Psi$.

Φ: The set of ordered pairs of tasks between which there is a precedence relationship. When task i must precede task j, $(i,j) \in \Phi$.

Decision variables.

X_{ij}^k: 1, if QC k performs task j immediately after performing task i; 0, otherwise. Tasks 0 and T will be considered to be the initial and final states of each QC, respectively. Thus, when task j is the first task of QC k, $X_{0j}^k = 1$. Also, when task j is the last task of QC k, $X_{jT}^k = 1$.

Y_k: The completion time of QC k.

D_i: The completion time of task i.

Z_{ij}: 1, if task j starts later than the completion time of task i; 0, otherwise.

W: Time at which all tasks are completed.

The QC scheduling problem can be formulated as follows:

$$\text{Minimize } \alpha_1 W + \alpha_2 \sum_{k=1}^{K} Y_k \tag{7.12}$$

subject to

$$Y_k \leq W \qquad \text{for } k = 1, \ldots, K \tag{7.13}$$

$$\sum_{j \in \Omega} X_{0j}^k = 1 \qquad \text{for } k = 1, \ldots, K \tag{7.14}$$

$$\sum_{i \in \Omega} X_{iT}^k = 1 \qquad \text{for } k = 1, \ldots, K \tag{7.15}$$

$$\sum_{k} \sum_{i \in \Omega} X_{ij}^k = 1 \qquad \text{for } j \in \Omega \tag{7.16}$$

$$\sum_{j} X_{ij}^k - \sum_{j} X_{ji}^k = 0 \qquad \text{for } i \in \Omega, \forall\, k = 1, \ldots, K \tag{7.17}$$

$$D_i + t_{ij} + p_j - D_j \leq M(1 - X_{ij}^k) \qquad \text{for } i, j \in \Omega, \forall\, k = 1, \ldots, K \tag{7.18}$$

$$D_i + p_j \leq D_j \qquad \text{for } (i,j) \in \Phi \tag{7.19}$$

$$D_i - D_j + p_j \leq M(1 - Z_{ij}) \quad \text{for } i, j \in \Omega \tag{7.20}$$

$$Z_{ij} + Z_{ji} = 1 \qquad \text{for } (i,j) \in \Psi \tag{7.21}$$

$$\sum_{v=1}^{k} \sum_{u \in \Omega} X_{uj}^v - \sum_{v=1}^{k} \sum_{u \in \Omega} X_{ui}^v \leq M(Z_{ij} + Z_{ji}) \qquad \text{for } i, j \in \Omega,\ l_i < l_j,\ k = 1, \ldots, K \tag{7.22}$$

$$D_j + t_{jT}^k - Y_k \ \leq M(1 - X_{jT}^k)$$

$$\text{for } j \in \Omega,\ k = 1, \ldots, K \tag{7.23}$$

$$r_k - D_j + t_{0j}^k + p_j \leq M(1 - X_{0j}^k)$$
$$\text{for } j \in \Omega,\ k = 1, \ldots, K \tag{7.24}$$

$$X_{ij}^k, Z_{ij} = 0 \text{ or } 1 \qquad \text{for } i, j \in \Omega,\ k = 1, \ldots, K \tag{7.25}$$

$$Y_k, D_i \geq 0 \qquad \text{for } i \in \Omega,\ k = 1, \ldots, K \tag{7.26}$$

In the objective function (7.12), it is assumed that $\alpha_1 >> \alpha_2$, because the minimization of the makespan is considered to be more important than the minimization of the total completion time. This is a valid assumption because a ship can depart only after every QC assigned to it completes all the assigned tasks, and the earliest departure is the primary objective of QC scheduling. However, among possible solutions having the minimum makespan, the schedule with the shortest total completion time must be selected so that more QCs may become available to other ships as soon as possible. Constraint (7.13) evaluates the makespan. Constraints (7.14) and (7.15) respectively select the first and last tasks for each QC. Constraint (7.16) ensures that every task must be completed by exactly one QC. Constraint (7.17) is a flow balance constraint, guaranteeing that tasks are performed in well-defined sequences. Constraint (7.18) simultaneously determines the completion time for each task and eliminates sub-tours. When required, constraint (7.19) denotes that task i should be completed before task j. Constraint (7.20) defines Z_{ij} such that $Z_{ij} = 1$ when the operation for task j starts after the operation for task i is completed; 0, otherwise. Constraint (7.21) guarantees that tasks i and j cannot be performed simultaneously when $(i, j) \in \Psi$. By constraint (7.22), interference among QCs can be avoided. Suppose that tasks i and j are performed simultaneously and $l_i < l_j$. This means that $Z_{ij} + Z_{ji} = 0$. Note that both QCs and tasks are ordered in an increasing order of their relative locations in the direction of increasing ship-bay number. Suppose that, for $k_1 < k_2$, QC k_1 performs tasks j and QC k_2 performs task i. Then, interference between QCs k_1 and k_2 results. However, in such a case, $\sum_{v=1}^{k_1} \sum_{u \in \Omega} X_{uj}^v - \sum_{v=1}^{k_1} \sum_{u \in \Omega} X_{ui}^v = 1$, which cannot be allowed because of constraint (7.22), and $Z_{ij} + Z_{ji} = 0$. The completion time of each QC is defined by constraint (7.23). Constraint (7.24) restricts the earliest starting time of operations by each QC. Kim and Park (2004) proposed a branch and bound algorithm for the optimal solution and a heuristic algorithm based on GRASP to reduce the computational time.

3. Stowage planning and sequencing

A group of containers is defined as a collection of containers of the same size and with the same destination port. Stowage planning determines which block (cluster) of slots in a ship-bay a specific group of containers should be stacked into. During the stowage planning process, rehandles of containers — bound for succeeding ports — in higher tiers for unloading containers — bound for preceding ports — located in lower tiers must be considered. Also, various indices for the stability and strength of the containership could be checked. Stowage planning is usually conducted by planners employed in carrier companies.

Shields (1984), Saginaw II and Perakis (1989), and Wilson and Roach (1999) introduced computer software to aid the construction of stowage plans. Sculli and Hui (1988), Aslidis (1990), Avriel and Penn (1993), Avriel et al. (1998), and Avriel et al. (2000) addressed the problem of stowage planning and attempted to minimize the number of rehandles as the main objective. They proposed various analytical formulations and algorithms to solve the problem. Todd and Sen (1997) proposed a genetic algorithm in order to solve the container-stowage-planning problem according to multiple criteria.

Based on the stowage plan, planners in container terminals determine the sequence of unloading inbound containers and of loading outbound containers. For the outbound containers, in addition to the loading sequence for individual containers, the slot in the vessel into which each outbound container will be stacked must be determined at the same time. When the indirect transfer system is used, the loading sequence of individual containers influences the handling cost in the yard significantly, while, in the direct transfer system, the handling cost in the yard is not significantly affected by the loading sequence. For unloading containers, because determining the discharging sequence is straightforward and determining the stacking locations of containers in the yard is done in real time, more academic researchers have focused on the sequencing problem for loading operations than on that for discharging operations. In loading operations, containers to be loaded into slots in a vessel must satisfy various constraints on the slots pre-specified by a stowage planner. Also, the locations of outbound containers may be scattered over a wide area in a marshaling yard. The time required for loading operations depends on the cycle time of QCs and YCs. Also, the cycle time of a CC depends on the loading sequence of slots, while the cycle time of a TC is affected by the loading sequence of containers in the yard.

Research on load sequencing can be classified into three types according to its problem-solving approach: mathematical programming approaches (Kim and Kim, 1999a,b; Narasimhan and Palekar, 2002), heuristic algorithms (Beliech, 1974; Cojeen and Dyke, 1976; Gifford, 1981; Kim and Kim, 2003) and meta-heuristic approaches (Kim and Kim, 1999c; Kozan and Preston, 1999; Ryu et al., 2001). Research on load sequencing can also be classified by the scope of the problem. Some research has addressed the pickup scheduling problem in which the travel route of each yard crane and the number of containers to be picked up at each yard-bay on the route are determined (Kim and Kim, 1999a,b,c; Narasimhan and Palekar, 2002; Ryu et al., 2001; Kim and Kim, 2003). Other research has focused on the loading sequence of individual containers in the marshaling yard and into slots in the vessel, which is a process that requires more detailed scheduling than does pickup scheduling (Beliech, 1974; Gifford, 1981; Cojeen and Dyke, 1976; Kim et al. 2004; Kozan and Preston, 1999).

Because the problem of load sequencing is highly complicated, most studies have applied heuristic algorithms to solve the problem. Also, the following typical objectives must be pursued, and the following constraints (Kim et al., 2004) must be satisfied, by the loading sequence.

Objectives related to the operation of QCs.

- First fill slots in the same hold.
- First stack containers onto the same tier on deck.
- Stack containers of weights included in the same weight group as specified in the stowage plan.

Objectives related to the operation of TCs.

- Minimize the travel time of TCs.
- Minimize the number of rehandles.
- Pick up containers in locations nearer to the transfer point earlier than those located farther from the transfer point.

Constraints related to the operation of QCs.

- Maintain the precedence relationships (according to work schedules for QCs and the relative positions between slots in a ship-bay) among slots.
- Do not violate the maximum allowed total weight of the stack on deck.
- Do not violate the maximum allowed height of the stack of a hold.
- Load the same type of containers as specified in the stowage plan.

Constraints related to the operation of TCs.

- Maintain such a distance between adjacent TCs that they can transfer containers without interference between each other.

4. Optimizing handling and storage activities in the yard

Related to the operation of the yard, research has been done to minimize handling activities and utilize space efficiently. Chen (1999) introduced practical operational procedures for handling activities in container terminals. Taleb-Ibrahimi et al. (1993) and Castilho and Daganzo (1993) analyzed various handling activities and decision-making problems for allocating spaces to export and import containers. As in the paper by Castilho and Daganzo (1993), Kim (1997) proposed a method for estimating the number of rehandles during the process of retrieving import containers. Castilho and Daganzo (1991) and Holguin-Veras and Jara-Diaz (1999) addressed how to determine the costs of storing freight in temporary storage space in ports.

Space is an important resource in container yards. Also, the storage locations of containers usually affect the productivity of handling activities. Several studies have been completed on locating inbound and outbound containers. Zhang (2000) decomposed the location problem of containers into the space allocation sub-problem, in which the number of incoming containers to be stacked at each storage area (for example, a block) is determined, and the slot determination sub-problem, in which the precise storage location for each arriving container is determined. Zhang et al. (2003) further decomposed the decision-making procedure for the space allocation sub-problem into the following two stages: to determine the total numbers of inbound and outbound containers that can be assigned to each block to balance the number of containers to be handled among blocks in each period, and to allocate the numbers of inbound and outbound containers of each vessel to the blocks in each period to minimize total distance traveled by YTs.

Cao and Uebe (1995), Zhang et al. (2003), Kozan (2000), Kim and Park (2003a), and Kim and Park (2003b) addressed the space allocation sub-problem defined by Zhang (2000). Kim and Kim (1999d) addressed the space allocation problem for import containers in which the amount of space allocated to import containers are determined to accommodate the dynamically changing space requirements of import containers. However, the storage slot was not a decision variable. Duinkerken et al. (2001), Kim et al. (2000), and Preston and Kozan (2001) addressed the slot determination problem.

Because outbound containers arrive at the terminal over a long period and containers with different attributes must not be mixed in the same storage space unit, re-marshaling operations may be necessary to increase the productivity of the loading operation. Kim and Bae

(1998) addressed the problem of remarshaling outbound containers — which have already arrived at the marshaling yard — for more efficient loading operations considering the stowage plan.

The following introduces the space allocation problem and a formulation (Kim and Park, 2003b). In container terminals, there are various flows of containers with different sources and destinations. The examples are export containers coming from gates, feeder vessels (vessels from local ports), or rail yards, and import containers unloaded from vessels, feeder vessels or trains, and transported through gates. A storage activity is defined by a set of containers which start their travel from the same source at the same time and end their travel at the same destination at the same time. Factors related to each storage activity are as follows: the amount of storage space required, the source, the destination, the starting time of the storage activity, and the ending time of the storage activity. For example, an export container unloaded from the rail yard has the rail yard as the source and the berth as the destination. Information regarding all of the above factors can be obtained from delivery schedules or forecasts, and from unloading/loading schedules for vessels.

The space allocation problem can be expressed as a minimum-cost multi-commodity network flow problem in the network $G = (N, A)$, as shown in Figure 7.11. There are three kinds of nodes (N) in the network. First, source nodes, S^i, have only exiting arcs, and correspond to events of initiating storage activity i. Second, terminal nodes, T^i, have only entering arcs, and correspond to events of terminating storage activity i. Intermediate nodes correspond to events of starting or ending storage activities. Intermediate nodes are denoted as i^k, where i is the index of a storage location and k is the index of the period at which a storage or retrieval event occurs. Intermediate nodes have both entering and exiting arcs. Arcs from source nodes to intermediate nodes or from intermediate nodes to terminal nodes have no limits on capacity, but have costs corresponding to the transportation cost from source locations to storage locations, or from storage locations to destinations. Arcs between intermediate nodes have no cost but are limited in their capacities.

Let the kth candidate route from node S^i to node T^i be denoted as R_{ik}. Then, storage activity i can be expressed by one of the following routes:

$$\begin{aligned}
R_{i1}\colon & S^i \rightarrow 1^{h_i} \rightarrow 1^{h_i+1} \rightarrow 1^{h_i+2} \rightarrow \cdots \rightarrow 1^{t_i-1} \rightarrow 1^{t_i} \rightarrow T^i \\
R_{i2}\colon & S^i \rightarrow 2^{h_i} \rightarrow 2^{h_i+1} \rightarrow 2^{h_i+2} \rightarrow \cdots \rightarrow 2^{t_i-1} \rightarrow 2^{t_i} \rightarrow T^i \\
\vdots &
\end{aligned}$$

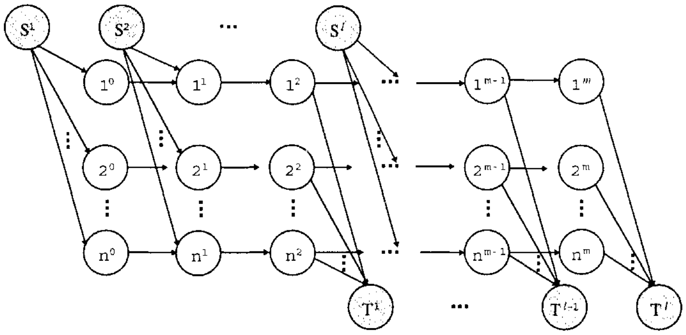


Figure 7.11. A network representation of the space allocation problem

$$R_{in}: S^i \rightarrow n^{h_i} \rightarrow n^{h_i+1} \rightarrow n^{h_i+2} \rightarrow \cdots \rightarrow n^{t_i-1} \rightarrow n^{t_i} \rightarrow T^i.$$

In the above routes, h_i and t_i, respectively, represent the starting and the ending times of storage activity i. Additionally, the following notations are introduced.

$l =$ the total number of storage activities.

$m =$ the number of periods.

$n =$ the number of storage locations.

$c_{ik} =$ the total cost when a container (in units of space) of storage activity i is moved following route k. The total cost is the sum of the transportation cost, the handling cost, and the storage cost. The transportation cost consists of the cost of delivering a container from its source location the storage area on route k, and the cost of delivering it from the storage area to its destination. The handling cost is the cost related to handling activities at the storage area.

$f_{ik} =$ the number of containers (in units of space) of storage activity i that follows route k (a decision variable).

$$a_{ik}^j = \begin{cases} 1, & \text{when arc } j \text{ is included in route } k \text{ for storage activity } i; \\ 0, & \text{otherwise.} \end{cases}$$

$d_i =$ the total space required for storage activity i. For example, d_i of a storage activity in Figure 7.5 represents the total number (in units of space) of containers to be stored.

$CAP_j =$ the capacity of storage location j. This also represents the capacity of arc j. This represents the total amount of space of storage location j. A storage location can be used by multiple different storage activities at the same time only if the capacity of storage location allows.

Thus, the space allocation problem can be formulated as follows:

$$\text{Min} \sum_{i=1}^{l}\sum_{k=1}^{n} c_{ik} f_{ik}, \tag{7.27}$$

subject to

$$\sum_{i=1}^{l}\sum_{k=1}^{n} a_{ik}^{j} f_{ik} \leq CAP_j \quad \text{for all } j \in A \tag{7.28}$$

$$\sum_{k=1}^{n} f_{ik} = d_i \quad \text{for } i = 1, 2, \ldots, l \tag{7.29}$$

$$f_{ik} = \text{a nonnegative integer} \quad \text{for } i = 1, 2, \ldots, l, \text{ and } k = 1, 2, \ldots, n. \tag{7.30}$$

The objective function (7.27) is to minimize the total transportation cost of all the storage activities. Constraint (7.28) limits the total flow of each arc. Actually, constraint (7.28) corresponds to the limitations of the storage space at each storage location. Constraint (7.29) implies that the space requirement of each storage activity must be satisfied.

Formulations (7.27) – (7.30) essentially form the minimum-cost multi-commodity flow problem with an integrality constraint for the flows. Although the problem structure shown in Figure 7.5 is a special case of the general multi-commodity flow problem, it can be shown that the problem is still NP-hard.

Mattfeld and Kopfer (2003) introduced an automated planning and scheduling system for the vehicle trans-shipment operation in Bremerhaven. They described an integrated decision model for determining the number of drivers (manpower) employed for each trans-shipment task and storage locations of vehicles. An integer programming model was proposed to solve the problem.

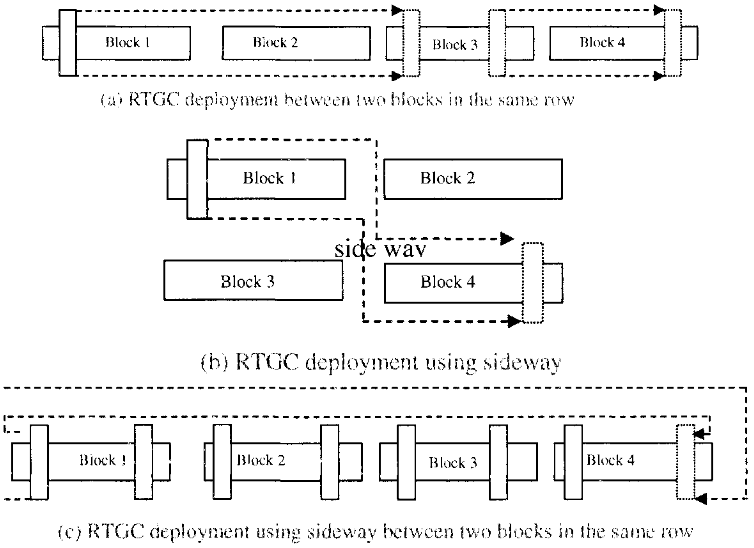
5. Allocating and dispatching yard cranes and prime movers

Yard equipment is also an important resource in container terminals. This equipment includes yard cranes, yard trucks, straddle carriers, or automated guided vehicles in automated container terminals. The main issue in utilizing yard equipment has been the assignment of handling tasks to different pieces of equipment.

Most research has been focused on the problem of dispatching yard cranes and prime movers (yard trucks, automated guided vehicles, straddle carriers). Cheung et al. (2002), Lai and Lam (1994), Lai and Leung

(1996), Kim et al. (2003), Linn and Zhang (2003), Linn et al. (2003), and Zhang et al.!(2002) addressed the problem of dispatching yard cranes to transfer tasks. In the following, a mathematical model proposed by Cheung et al. (2002) is introduced.

Cheung et al. (2002) considered the problem of scheduling the movements of rubber tired gantry cranes (RTGCs) in a container storage yard so as to minimize the total unfinished workload at the end of each time period.



(a) RTGC deployment between two blocks in the same row

(b) RTGC deployment using sideway

(c) RTGC deployment using sideway between two blocks in the same row

Figure 7.12. Trajectories of cranes in a Container Yard (Cheung et al., 2002)

They assumed that a crane can only leave a block at the beginning of a period. Workload can arrive at the beginning of any time period. Due to the physical size of each block and the potential danger of crane collision, the maximum number of cranes that can work simultaneously in one block is limited. Unfinished workload in a time period will be carried over to the next period. They assumed that each crane is at a block at the beginning of period 1 and has to be located at a block at the end of period T, that is, a crane can not be in the middle of an inter block movement at the beginning and the end of the planning horizon.

The following notation is used in their study.

Parameters.

$M =$ number of blocks in the terminal yard,
$T =$ number of time period,

$K =$ maximum number of cranes allowed to work simultaneously in a block,

$\tau_{ij} =$ number of time periods it takes for a crane to travel from block i to block j $(\tau_{ij} \equiv 0)$,

$\omega_{it} =$ estimated workload arriving at block i at the beginning of period t,

$b_i =$ initial number of cranes in block i at the beginning of period 1.

Decision variables.

$x_{ijt} =$ number of cranes leaving block i to travel to block j at the beginning of period (x_{ijt} represents the number of cranes staying in block i during period t),

$S_{it} =$ unfinished workload in block i carried over from the end of period t to the beginning of period $t + 1(S_{i0} \equiv 0)$.Note that in our model, we assume that $S_{i0} = 0$ for $i = 1, \ldots, M$.

The problem can be formulated as an MILP as follows:

$$\text{(P1) Minimize} \sum_{t=1}^{T} \sum -i = 1^M S_{it} \tag{7.31}$$

subject to

$$\sum_{j=1}^{M} x_{ijt} = \sum_{\substack{j=1\\ j\neq 1}}^{M} x_j, i, t - \tau_{ji} + x_{i,i,t-1}, \qquad i = 1, \ldots, M; t = 2, \ldots, T, \tag{7.32}$$

$$\sum_{j=1}^{M} x_{ij1} = b_i, \qquad i = 1, \ldots, M, \tag{7.33}$$

$$\sum_{i=1}^{M} \left(\sum_{\substack{j=1\\ j\neq 1}}^{M} x_{j,i,T-\tau} + x_{iiT} \right) = \sum_{i=1}^{M} b_i \tag{7.34}$$

$$x_{ijt} \leq K, \qquad i = 1, \ldots M; t = 1, \ldots T, \tag{7.35}$$

$$s_{i,t-1} + w_{it} - s_{it} \leq x_{iit}, \qquad i = 1, \ldots, M; t = 1, \ldots, T, \tag{7.36}$$

$$s_{it} \geq 0, \qquad i = 1, \ldots, M; t = 1, \ldots, T, \tag{7.37}$$

$$x_{ijjt} \geq 0, \text{ integer}, \qquad i, j = 1, \ldots, M; t = 1, \ldots, T, \tag{7.38}$$

The objective is to minimize the total unfinished workload at the end of each time period. Constraint (7.32) maintains the conservation of flow of cranes. Constraint (7.33) specifies the number of cranes in each block at the beginning of the first period. Constraint (7.34) requires that the total number of cranes in all the blocks at the end of the last period is the same as that at the beginning of the first period. Constraint (7.35) ensures that at most K cranes can work simultaneously in a block. The left side of constraint (7.36) is the amount of workload finished in block i during period t, where the right side of this constraint is the crane capacity available in block i during that period. They referred to constraint (7.32) – (7.36) and (7.38) as the *crane flow constraints* and to constraint (7.36) as the *workload capacity constraint.* Cheung et al. (2002) proposed a Lagrangean decomposition solution procedure and a successive piecewise-linear approximation method for solving the above mathematical model.

Many studies have been carried out on dispatching prime movers such as yard trucks (Bish, 1999), straddle carriers (Bose et al., 2000), automated guided vehicles (Bish, 2003; Grunow et al., 2004; Kim and Bae, 1999; Kim and Bae, 2004; Vis et al., 2001), and automatic lifting vehicles (van der Meer, 2000). Hartmann (2004) proposed a genetic algorithm to dispatching various handling equipment and manpower in container terminals. Holguin-Veras and Walton (1995) proposed methods for estimating and calibrating the service times of various handling equipment in container terminals for use in a simulation program. Yang et al. (2004) and Vis and Harika (2004) compared the performance of different types of automated vehicles for transporting containers in port container terminals. Evers and Koppers (1996) addressed the traffic control problem and Qiu and Hsu (2000) discussed the routing problem for AGVs in container terminals.

The following are from the mathematical model that Bish (2003) suggested. The objective is (1) to determine a storage location for each unloaded container; (2) to schedule the trip of each container on a vehicle; and (3) to schedule the loading and unloading operations of the QCs so as to minimize the total travel time.

Two ships were considered, sh^- and sh^+, that need unloading and loading, respectively, and that are berthed around the same time so that they can be served by the same set of k vehicles. Let N^- and N^+ denote the set of containers that will be unloaded from ship sh^-, and the set of containers that will be loaded onto ship sh^+, respectively. Also, let c_i be the element in N^-. Associated with each container to be loaded onto a ship is its *current storage location* in the yard area, which is known. Each such container will require a *loaded vehicle trip* from its current storage location to the location of ship sh^+. Let L^+ denote the set of current storage locations of the containers in set N^+. A set of *potential storage location* in the yard area is reserved for the containers of each unloading ship. We are given a set of *potential storage locations* reserved for all containers that will be unloaded from sh^-, and we denote this set as L^-. Each unloaded container will require a *loaded vehicle trip* from the location of ship sh^- to its *selected* storage location. We will make the simplifying assumption that sets L^+ and L^- are disjoints, that is, we assume that no container in set N^- can be stored in a location currently occupied by a container in set N^+. Let $L = L^- \cup L^+$. Also let W_i be the subset of L^- where c_i can be stacked.

In addition to these loaded trips, each vehicle will need to make and empty trip between two loaded trips scheduled right after each other on that vehicle, if the destination of the previous loaded trip and the origin of the next loaded trip are different. Thus, these empty trip times depend on the sequence of loaded trips on each vehicle. Each unloaded container is assigned to exactly one potential storage location and each loaded trip for an unloaded container is matched with a loaded trip for a loading container in a way of minimizing the total travel distance. Let denote loaded trip i by l_i and the travel time of l_i by t_{li}.

The following network is constructed to solve the problem. A supply node, with unit supply, is created for each unloaded container $c_i \in N^-$. And a demand node is created for each location $q \in L^+$. Additionally, two copies of trans-shipment node l'_p and l_p, are made for each potential location $p \in L^-$, in our network. The arc set is given by

$$A = \{(c_i, l'_p) : c_i \in N^-, p \in W_i\} \cup \{(l'_p, l_p) : p \in L^-\} \\ \cup \{(l_p, l_q) : p \in L^-, q \in L^+\},$$

each with unit capacity. For each $(c_i, l'_p) \in A$, the arc cost is the travel time of the corresponding loaded trip for unloading container i, given by t_{lp}. For each $(l'_p, l_p) : p \in L^-$, the cost is zero and for each $(l_p, l_q) : p \in L^-, q \in L^+$, the arc cost is the empty travel time from the destination of the loaded trip l_p to the origin of the loaded trip l_q, denoted by λ_{pq}. For each $(u, v) \in A$, let $X_{uv} = 1$, if arc (u, v) is used in the solution; 0, otherwise. Now we can formulate the problem as a trans-shipment problem, as follows:

$$\text{Minimize} \sum_{c_i \in N^-} \sum_{p \in W_i} t_{l_p} X_{c_i l'_p} + \sum_{p \in L^-} \sum_{q \in L^+} \lambda_{pq} X_{l_p l_q} \qquad (7.39)$$

subject to

$$\sum_{p \in W_i} X_{c_i l'_p} = 1 \qquad \forall c_i \in N^-, \qquad (7.40)$$

$$\sum_{\substack{c_i : c_i \in N^- \\ p \in W_i}} X_{c_i l'_p} = X_{l'_p l_p} \qquad \forall p \in L^- \qquad (7.41)$$

$$X_{l'_p l_p} = \sum_{q \in L^+} X_{l_p l_q} \qquad \forall p \in L^- \qquad (7.42)$$

$$\sum_{p \in L^-} X_{l_p l_q} = 1 \qquad \forall\, q \in L^+, \tag{7.43}$$

$$X_{uv} = 0 \text{ or } 1 \qquad \forall\, (u,v) \in A. \tag{7.44}$$

The first constraint ensures that each unloaded container is assigned to exactly one loaded trip. The second and third constraints are flow-balance constraints. The fourth constraint ensures that each loaded trip is matched with an unloaded trip. The objective is to minimize the total travel-related time.

Because the system of container terminals is highly complicated, many researchers have used the simulation approach to solve various practical problems. Although there are many research papers related to the application of the simulation technique, they were excluded from this review. Also, there are many papers which have addressed issues of estimating the throughput capacity of the berth or of the entire handling system. The main approaches concerned the simulation and queuing theories. The research results can be mainly used in the planning stage of a new container terminal. These papers were also excluded from the list of references.

6. Conclusions

This chapter reviewed previous studies that applied operation research techniques to the operational and design problems of port container terminals. It was shown how operation research techniques can be applied to such planning activities as berth planning, crane scheduling, load sequencing, space allocation, and resource allocation. Also, it revealed that there are many operational problems such as the dispatching of prime movers, yard cranes, automated guided vehicles, and other handling equipment for which operation research can be useful tools. Several mathematical models were introduced.

Recently, much effort has been devoted to automate various operations in container terminals. The effort has been realized in some container terminals such as ECT terminal in Rotterdam, CTA terminal in Hamburg, Thames port in the UK, and others. Automation requires detailed operation orders and decisions for equipment that have been made by human operators in conventional container terminals. Thus, operations researchers now face much more challenging problems realizing the automation of container terminals.

The sizes of containerships are continuously increasing and containerships with capacity larger than 8000 TEU will become popular in the next decade. Thus, the loading and unloading speed of container handling equipment in ports must be increased dramatically for the large-sized vessels to keep their voyage schedules. In addition to developing higher speed equipment, more efficient operational decision-making algorithms must be developed and the computational times of the algorithms must be significantly shortened.

Until now, operational researchers have considered operational problems of container terminals to be isolated from outside logistic nodes (rail yards, feeder ports, inland depots, and so on). However, considering a container terminal is only a node in a much larger logistics network, many new decision-making problems, resulting from integrating functions of outside nodes to those of the container terminal, must be promising issues for future studies.

Acknowledgment This research was supported in part by Pusan National University Research Grant (2003 – 2007).

References

Aslidis, A. (1990). Minimizing of over-stowage in container ship operations. *Operations Research*, 90:457 – 471.

Avriel, M. and Penn, M. (1993). Exact and approximate solutions of the container ship stowage problem. *Computers and Industrial Engineering*, 25:271 – 274.

Avriel, M., Penn, M., Shpirer, N., and Witteboon, S. (1998). Stowage planning for container ships to reduce the number of shifts. *Annals of Operations Research*, 76:55 – 71.

Avriel, M., Penn, M., and Shpirer, N. (2000). Container ship stowage planning: Complexity and connection to the coloring of circle graphs. *Discrete Applied Mathematics*, 103:271 – 279.

Beliech, D.E. Jr. (1974). *A Proposed Method for Efficient Pre-Load Planning for Containerized Cargo Ships.* Master Thesis, Naval Postgraduate School.

Bish, E.K. (1999). *Theoretical Analysis and Practical Algorithms for Operational Problems in Container Terminals.* Ph.D. Thesis, Northwestern University.

Bish, E.K. (2003). A multiple-crane-constrained scheduling problem in a container terminal. *European Journal of Operational Research*, 144:83 – 107.

Bose, J., Reiners, T., Steenken, D., and Voß, S. (2000). Vehicle dispatching at seaport container terminals using evolutionary algorithms. *Proceedings of the 33rd Hawaii International Conference on System Sciences.* IEEE, Piscataway.

Brown, G.G., Cormican, K.J., Lawphongpanich, S., and Widdis, D.B. (1997). Optimizing submarine berthing with a persistence incentive. *Naval Research Logistics*, 44:301 – 318.

Brown, G.G., Lawphongpanich, S., and Thurman, K.P. (1994). Optimizing ship berthing. *Naval Research Logistics*, 41:1 – 15.

Cao, B. and Uebe, G. (1995). Solving transportation problems with nonlinear side constraints with tabu search. *Computers & Operations Research*, 22(6):593 – 603.

de Castilho, B. and Daganzo, C.F. (1991). Optimal pricing policies for temporary storage at ports. *Transportation Research Record*, 1313:66 – 74.

de Castilho, B. and Daganzo, C.F. (1993). Handling strategies for import containers at marine terminals. *Transportation Research B*, 27(2):151 – 166.

Chen, T. (1999). Yard operations in the container terminal — a study in the "unproductive moves". *Maritime Policy & Management*, 26(1):27 – 38.

Cheung, R.K., Li, C.-L., and Lin, W. (2002). Interblock crane deployment in container terminals. *Transportation Science*, 36(1):79 – 93.

Cojeen, H.P. and Dyke, P.V. (1976). The automatic planning and sequencing of containers for containership loading and unloading. In: Pitkin, Roche and Williams (eds.), pp. 415 – 423, *Ship Operation Automation*, North-Holland Publishing Co.

Daganzo, C.F. (1989). The crane scheduling problem. *Transportation Research B*, 23(3):159 – 175.

Duinkerken, M.B., Evers, J.J.M., and Ottjes, J.A. (2001). A simulation model for integrating quay transport and stacking policies on automated container terminals. *Proceedings of the 15th European Simulation Multiconference, ESM2001*, Prague.

Evers, J.J.M. and Koppers, S.A.J. (1996). Automated guided vehicle traffic control at a container terminal. *Transportation Research A*, 30(1):21 – 34.

Grunow, M., Günther, H.-O., and Lehmann, M. (2004). Dispatching multi-load AGVs in highly automated seaport container terminals. *OR Spectrum*, 26:211 – 235.

Guan, Y., Xiao, W.Q., Cheung, R.K., and Li, C.-L. (2002). A multiprocessor task scheduling model for berth allocation: heuristic and worst-case analysis. *Operations Research Letters*, 30(5):343 – 350.

Guan, Y. and Cheung, R.K. (2004). The berth allocation problem: models and solution methods. *OR Spectrum*, 26:75 – 92.

Gifford, L.A. (1981). *Containership Load Planning Heuristic for a Transtainer-Based Container Port.* Unpublished M.Sc. Thesis, Oregon State University.

Hartmann, S. (2004). General framework for scheduling equipment and manpower on container terminals. *OR Spectrum*, 26:51 – 74.

Holguin-Veras, J. and Jara-Diaz, S. (1999). Optimal pricing for priority service and space allocation in container ports. *Transportation Research B*, 33:81 – 106.

Holguin-Veras, J. and Walton, C.M. (1995). *The Calibration of PRIOR, a Computer System for the Simulation of Port Operations Considering Priorities.* Southwest Region University Transportation Center, Center for Transportation Research, The University of Texas, Austin, Texas 78712.

Imai, A., Nagaiwa, K., and Tat, C.W. (1997). Efficient planning of berth allocation for container terminals in Asia. *Journal of Advanced Transportation*, 31(1):75 – 94.

Imai, A., Nishimura, E., and Papadimitriou, S. (2001). The dynamic berth allocation problem for a container port. *Transportation Research B*, 35:401 – 417.

Imai, A., Nishimura, E., and Papadimitriou, S. (2003). Berth allocation with service priority. *Transportation Research B*, 37:437 – 457.

Kim, K.H. (1997). Evaluation of the number of rehandles in container yards. *Computers and Industrial Engineering*, 32(4):701 – 711.

Kim, K.H. and Bae, J.-W. (1998). Remarshaling export containers in port container terminals. *Computers and Industrial Engineering*, 35(3-4):655 – 658.

Kim, K.H. and Bae, J.-W. (1999). A dispatching method for automated guided vehicles to minimize delay of containership operations. *International Journal of Management Science*, 5(1):1 – 25.

Kim, K.H. and Bae, J.-W. (2004). A look-ahead dispatching method for automated guided vehicles in automated port container terminals. Forthcoming in *Transportation Science.*

Kim, K.H., Kang, J.S., and Ryu, K.R. (2004). A beam search algorithm for the load sequencing of outbound containers in port container terminals. *OR Spectrum*, 26:93 – 116.

Kim, K.Y. and Kim, K.H. (1999a). A routing algorithm for a single straddle carrier to load export containers onto a containership. *International Journal of Production Economics*, 59:425 – 433.

Kim, K.Y. and Kim, K.H. (1999b). An optimal routing algorithm for a transfer crane in port container terminals. *Transportation Science*, 33(1):17 – 33.

Kim, K.Y. and Kim, K.H. (1999c). Routing straddle carriers for the loading operation of containers using a beam search algorithm. *Computers and Industrial Engineering*, 36(1):109 – 136.

Kim, K.H. and Kim, H.B. (1999d). Segregating space allocation models for container inventories in port container terminals. *International Journal of Production Economics*, 59:415 – 423.

Kim, K.Y. and Kim, K.H. (2003). Heuristic algorithms for routing yard-side equipment for minimizing loading times in container terminals. *Naval Research Logistics*, 50(5):498 – 514.

Kim, K.H., Lee, K.M., and Hwang, H. (2003). Sequencing delivery and receiving operations for yard cranes in port container terminals. *International Journal of Production Economics*, 84(3):283 – 292.

Kim, K.H. and Moon, K.C. (2003). Berth scheduling by simulated annealing. *Transportation Research B*, 37(6):541 – 560.

Kim, K.H. and Park, K.T. (2003a). A note on a dynamic space-allocation method for outbound containers *European Journal of Operational Research*, 148(1):92 – 101.

Kim, K.H. and Park, K.T. (2003b). Dynamic space allocation for temporary storage. *International Journal of System Science*, 34(1):11 – 20.

Kim, K.H. and Park, Y.-M. (2004). A crane scheduling method for port container terminals. *European Journal of Operational Research*, 156(3):752 – 768.

Kim, K.H. and Park, Y.-M., and Ryu, K.R. (2000). Deriving decision rules to locate export containers in container yards. *European Journal of Operational Research*, 124:89 – 101.

Kozan, E. (2000). Optimizing container transfers at multimodal terminals. *Mathematical and Computer Modeling*, 31:235 – 243.

Kozan, E. and Preston, P. (1999). Genetic algorithms to schedule container transfers at multimodal terminals. *International Transactions in Operational Research*, 6:311 – 329.

Lai, K.K. and Lam, K. (1994). A study of container yard equipment allocation strategy in Hong Kong. *International Journal of Modeling & Simulation*, 14(3):134 – 138.

Lai, K.K. and Leung, J.W. (1996). Analysis of yard crane deployment strategies in a container terminal. 1187 – 1190, *Proceedings of ICC & IE Conference*, Kyungju, Korea.

Lai, K.K. and Shih, K. (1992). A study of container berth allocation. *Journal of Advanced Transportation*, 26(1):45 – 60.

Li, C.-L., Cai, X., and Lee. C.-Y. (1998). Scheduling with multiple-job-on-one-processor pattern. *IIE Transactions*, 30:433 – 445.

Lim, A. (1998). The berth planning problem. *Operation Research Letters*, 22:105 – 110.

Linn, R.J. and Zhang, C.-Q. (2003). A heuristic for dynamic yard crane deployment in a container terminal. *IIE Transactions*, 35:161 – 174.

Linn, R., Liu, J.-Y., Wan, Y.-W., Zhang, C., and Murty, K.G. (2003). Rubber tired gantry crane deployment for container yard operation. *Computers & Industrial*

Engineering, 45:429 – 442.

Mattfeld, D.C. and Kopfer, H. (2003). Terminal operations management in vehicle transshipment. *Transportation Research A*, 37A(5):435 – 452.

Meersmans, P.J.M. and Dekker, R. (2001). *Operations Research Supports Container Handling.* Econometric Institute Report EI 2001-22, Erasmus University.

Narasimhan, A. and Palekar, U.S. (2002). Analysis and algorithms for the transtainer routing problem in container port operations. *Transportation Science*, 36(1):63 – 78.

Nishimura, E., Imai, A., and Papadimitriou, S. (2001). Berth allocation planning in the public berth system by genetic algorithms. *European Journal of Operational Research*, 131:282 – 292.

Park, K.T. and Kim, K.H. (2002). Berth scheduling for container terminals by using a subgradient optimization technique. *Journal of the Operational Research Society*, 53(9):1049 – 1054.

Park, Y.-M. and Kim, K.H. (2003). A scheduling method for berth and quay cranes. *OR Spectrum*, 25:1 – 23.

Park, Y-M. (2003). *Berth and Crane Scheduling of Container Terminals.* Ph.D. Thesis, Pusan National University.

Peterkofsky, R.I. and Daganzo, C.F. (1990). A branch and bound solution method for the crane scheduling problem. *Transportation Research B*, 24(3):159 – 172.

Preston, P. and Kozan, E. (2001). An approach to determine storage locations of containers at seaport terminals. *Computers & Operations Research*, 28:983 – 995.

Qiu, L. and Hsu, W.-J. (2000). Routing AGVs on a mesh-like path topology. 392-397, *Proceedings of the IEEE Inielligent Vehicles Symposium 2000*, Dearborn, USA.

Ryu, K.R., Kim, K.H., Lee, Y.H., and Park, Y.M. (2001). Load sequencing algorithms for container ships by using metaheuristics. *Proceedings of 16th International Conference on Production Research* (CD-ROM), Prague, Czech Republic.

Saginaw II, D.J. and Perakis, A.N. (1989). A decision support system for containership stowage planning, *Marine Technology*, 26(1):47 – 61.

Sculli, D. and Hui, C-F. (1988). Three dimensional stacking of containers. *OMEGA*, 16:585 – 594.

Shields, J.J. (1984). Containership stowage: A computer-aided preplanning system. *Marine Technology*, 21(4):370 – 383.

Steenken, D., Voß, S., and Stahlbock, R. (2004). Container terminal operation and operations research — a classification and literature review. *OR Spectrum*, 26:3 – 49.

Suh, M.S. and Lee, Y.J. (1998). A hierarchical expert system for integrated scheduling of ship berthing, discharging and material transport. *Expert Systems*, 15(4):247 – 255.

Taleb-Ibrahimi, M., Castilho, B., and Daganzo, C.F. (1993). Storage space vs handling work in container terminals. *Transportation Research B*, 27(1):13 – 32.

Todd, D.S. and Sen, P. (1997). A multiple criteria genetic algorithm for containership loading. In: Thomas Back (ed.) *Proceedings of the Seventh International Conference on Genetic Algorithms*, pp. 674 – 681. Morgan Kaufmann Publishers, Inc.

van der Heyden, W.P.A. and Ottjes, J.A. (1985). A decision support system for the planning of the workload on a grain terminal. *Decision Support Systems*, 1:293 – 297.

van der Meer, R. (2000). *Operational Control of Internal Transport.* Erasmus Research Institute of Management, Ph.D. Series Research in Management 1.

Vis, I.F.A., de Koster, R., and Roodbergen, K.J. (2001). Determination of the number of automated guided vehicles required at a semi-automated container terminal. *Journal of the Operational Research Society*, 52(4):409 – 417.

Vis, I.F.A. and Harika, I. (2004). Comparison of vehicle types at an automated container terminal. *OR Spectrum*, 26:117 – 143.

Vis, I.F.A. and de Koster, R. (2003). Transshipment of containers at a container terminal: An overview. *European Journal of Operational Research*, 147:1 – 16.

Wilson, D. and Roach, P.A. (1999). Principles of combinatorial optimization applied to container-ship stowage planning. *Journal of Heuristics*, 5:403 – 418.

Yang, C.H., Choi, Y.S., and Ha, T.Y. (2004). Simulation-based performance evaluation of transport vehicles at automated container terminals. *OR Spectrum*, 26:149 – 170.

Zhang, C. (2000). *Resource Planning in Container Storage Yard.* Ph.D. Thesis, The Hong Kong University of Science and Technology.

Zhang, C., Liu, J., Wan, Y.-W., Murty, K.G., and Linn, R.J. (2003). Storage space allocation in container terminals. *Transportation Research B*, 37:883 – 903.

Zhang, C., Wan, Y.-W., Liu, J., and Linn, R.J. (2002). Dynamic crane deployment in container storage yards. *Transportation Research B*, 36:537 – 555.

Chapter 8

STRATEGIC NETWORK DESIGN FOR MOTOR CARRIERS

James F. Campbell

Abstract This chapter reviews Operations Research models for strategic design of motor carrier networks, including network configuration and terminal location. This includes networks for less-than-truckload (LTL), truckload (TL), and postal motor carriers that serve many origins and destinations in large geographic regions. LTL carriers, as well as postal carriers, use networks with consolidation and break-bulk terminals to combine small shipments into efficient vehicle loads. Some TL carriers use networks with relay terminals where loads can be exchanged to allow drivers to return home more frequently. The chapter reviews research in each area and proposes directions for future research.

1. Introduction

Trucking is the most important mode of land freight transportation in the world. Within the United States, motor carriers account for 81% of the freight bill ($372 billion per year in revenues), 60% of the freight volume (6.7 billion tons per year) and nearly 430 billion miles traveled per year. More broadly, within North America motor carriers account for 64% of the merchandise trade by value (versus 25% for rail) and 32% by weight (versus 17% for rail) (United States Bureau of Transportation Statistics, 2003). Truck transport is even more important within the European Union, where it accounts for 75% of inland freight ton-km (road, rail, inland waterways, and pipelines) and 44.5% of the total freight ton-km (road, rail, short sea shipping, pipelines, and inland waterways) (European Commission 2003).

Motor carrier operations provide an important and rich source of decision problems, and there has been considerable prominent Operations Research (OR) work in a variety of areas. One of the key strategic decisions for motor carriers is the physical network over which the carrier operates. This chapter reviews operations research models for strategic design of motor carrier networks. Our focus is on *strategic* network design (including network configuration and terminal location) and on newer research, rather than on tactical network design, which includes load planning and service network design. Roy (2001) describes strategic planning for motor carriers as including:

(1) "the type and mix of transportation services offered...;

(2) the territory coverage and network configuration, including terminal location; and

(3) the service policy, what service levels are offered to customers in terms of both speed and reliability."

Roy distinguishes this from tactical service network design, which includes selecting routes on which services are offered, determining the sequence of services and terminals used to transport the freight, and the movement of empty trucks and trailers to balance the network.

This chapter considers strategic network design for general freight intercity public (for-hire) motor carriers and for postal motor carriers. The primary business of these carriers is to transport freight owned by others between many origins and destinations dispersed over a large geographic region. General freight carriers are usually classified as truckload (TL) or less-then-truckload (LTL) carriers. TL carriers generally haul full truckloads, usually direct from an origin to a destination. TL carriers may also use networks with relay terminals where loads can be exchanged to allow drivers to return more frequently to their home. LTL carriers use networks with consolidation and break-bulk terminals to combine many small shipments into efficient vehicle loads. Postal (and small parcel) motor carriers are very similar to general freight LTL carriers, but the freight is more specialized, and service constraints may force tight deadlines for delivery (for example, overnight).

The remainder of this chapter is organized as follows. The following section provides some background on motor carrier operations and reviews some relevant transportation network design literature. The next three sections discuss models for strategic network design in LTL trucking, TL trucking, and postal operations. The final section is a conclusion and discussion of directions for future research.

2. Background

Motor carriers have great versatility in being able to carry virtually any type of product, and to visit nearly every address (at least in regions with a well-developed infrastructure). The motor carrier industry can be divided many different ways. Public carriers haul a wide variety of freight for many different shippers, while private carriers haul freight exclusively for their own organization. General freight carriers may haul nearly any product, while specialized carriers may focus on unique products or markets, such as household goods, automobiles and trucks, liquids, hazardous materials, temperature controlled products, express shipments, etc. Public carriers of general freight have developed networks and operations to serve many dispersed origins and destinations. (Private carriers will generally have somewhat different networks designed to serve a few-to-many traffic pattern; for example, linking a few origins, such as manufacturing locations, with many destinations, such as wholesalers, retailers or customers.)

2.1 Operations

We summarize some relevant aspects of trucking operations in this section. See Delorme et al. (1988) and Roy (2001) for more details. Stumpf (1998) provides details on LTL operations in Germany, especially for transporting partial loads, which is common there (though not so much in North America).

LTL firms are the largest part of the motor carrier industry. LTL carriers consolidate many small shipments, each generally between 100 and 10,000 pounds (50 – 4,500 kg.) from many different shippers to make efficient vehicle loads. Trailers may hold 20,000 to 50,000 pounds (9,000 – 23,000 kg.) depending on the freight. LTL carriers typically route shipments via a network consisting of end-of-line terminals and break-bulk terminals. Each end-of-line terminal collects shipments from its local service region using local pickup/delivery trucks. (Shipments may also be delivered to the terminal by the shipper.) Shipments are sorted at the terminal and loaded into line-haul trucks, which carry the shipments to break-bulk terminals for consolidation with other shipments headed in the same direction. Line-haul vehicles then carry the shipments to another break-bulk terminal, where they may be unloaded and sorted again for transport to the end-of-line terminal serving the destination. The freight is then transshipped from the line-haul truck to a local delivery truck for transport to the destination. A typical LTL carrier in the U.S. generally has "an order of magnitude fewer break-bulks than terminals" (Bartholdi et al., 2003), which may mean several hundred end-of-line terminals and a few dozen break-bulks.

In LTL operations the local collection and delivery trucks may be small straight trucks or short tractor-trailer combinations. The local collection and delivery stops may change from day to day and this portion of the operation is generally not included in strategic network design. The linehaul trucks may be long tractor-trailer combinations, with one, two, or sometimes more, trailers. National and local regulations restrict vehicle sizes and weights, though the growth of free trade regions can impose common standards in larger regions.

TL operations are simpler than LTL operations, since consolidation of many small shipments is not required. TL shipments generally fill the trailer, so that the freight may move from the origin to the destination without intermediate handling and sorting. (Some carriers will haul several large loads with a common destination in the same trailer.) In point-to-point operations, a driver hauls the load from the origin to destination. Then, after delivering a load, the driver would like to find a return load originating nearby and destined for the vicinity of his home. Such return loads are rarely available when needed, so efficient routes for drivers may require a sequence of many long-haul trips before returning home. This long-haul nature of the trips and the difficulty in finding backhauls has led to very high turnover rates for drivers (Schwarz, 1992). Annual turnover rates over 100% are common and have been reported up 150%! (Griffin et al., 2000; Hunt, 1998; Road Haulage and Distribution Training Council, 2003). Since most drivers would prefer to return to their home on a regular and frequent basis, some TL carriers have developed networks of relay terminals to allow drivers to exchange loads and operate in more regular delivery lanes or regions, and thereby return home more frequently.

Postal motor carrier operations are quite similar to LTL operations, and these carriers operate networks of consolidation and break-bulk terminals to create efficient loads. Postal carriers may also operate intermodal networks with aircraft to allow fast delivery over longer distances. Our concern is primarily on motor carrier networks, but later in this section we list some relevant research on intermodal or integrated express carriers.

2.2 Freight transportation network design

For many years when motor carrier transportation was regulated, carriers performed a limited amount of strategic planning and network design. Prior to deregulation in the U.S. (via the Motor Carrier Act of 1980) Kallman and Gupta (1979) surveyed 498 motor carriers and found that "few ... planned for longer than a year, and most did so informally." However, in a deregulated environment, the success of any transportation carrier depends on its ability to attract and retain business via competitive rates and quality service. The cost incurred for carrying freight, the rate charged to shippers, and the level of service provided are all affected by the design of the physical network over which the carrier operates.

LTL carriers and postal carriers use a network of terminals to consolidate small shipments into economic truckloads. TL carriers use networks of terminals for different reasons, primarily to allow swapping of trailers so that drivers may return home more frequently. Given that motor carriers generally operate on publicly owned infrastructure, the network to be designed includes nodes representing private or public terminal facilities (for consolidation, break-bulk, sorting, or transshipment) and links representing travel on the roadways. Generally, the demand in motor carrier network design models is for transportation of specified quantities of freight between many origins and destinations. Origins and destinations may represent actual shipment origins and destinations, or the end-of-line terminals to which shipments are collected, and from which shipments are distributed, to the ultimate customers.

While general network design and tactical service network design have drawn considerable attention from operations researchers, much less work has been directed specifically at strategic network design for motor carriers. Our goal in the remainder of this section is to highlight some relevant literature for transportation network design, and to briefly mention the related work on tactical service network design for motor carriers, including load planning.

Crainic (2003) provides a comprehensive review of long-haul (intercity) transportation by both motor and rail carrier. He describes basic problems and solution approaches, and provides a broad perspective for both road and rail transport. Crainic describes strategic (long term) planning as including "design of the physical network and its evolution, the location of major facilities (e.g., terminals), the acquisition of major resources such as motive power units, and the definition of broad service and tariff policies." This is distinguished from tactical network design, which includes: "the design of the service network and may include issues related to the determination of the routes and types of services to operate, service schedules, vehicle and traffic routing, [and] repositioning of the fleet."

Section 13.4 of Crainic (2003) addresses logistics network design. This discusses location-based and network flow-based modeling approaches. Location-based models are used in transportation network design to capture decisions on the terminal locations. For a survey of this work, see Daskin et al. (2005), Daskin and Owen (2003), and Drezner and Hamacher (2002). For the network-flow based approach, Crainic (2003) provides standard arc-based and path-based fixed cost multicommodity capacitated network design formulations. In a multicommodity network flow formulation, the freight for each unique origin-destination pair is viewed as a distinct commodity. Decision variables can represent the flow on each arc or path in the network. Crainic also provides a brief discussion of solution approaches including Lagrangian relaxation, dual ascent, branch and bound, polyhedral approaches, and a variety of heuristics.

Fleischmann (1998) reviews recent literature on freight transportation network design, from the viewpoint of both a manufacturer and of a carrier. He proposes a general model for few-to-many networks and many-to-many networks and describes some solution approaches. He also briefly describes a decision support system (BOSS) for designing LTL networks, which is described in detail later in this chapter in the discussion of Wleck (1998).

2.3 Service network design

Section 13.5 of Crainic (2003) addresses service network design, which includes tactical decisions on the services to be offered (including frequencies and schedules), freight routing, terminal operational policies, and empty balancing strategies. (Kim and Barnhart, 1997, also review transportation service network design.) These problems are usually modeled via fixed cost capacitated multicommodity network design formulations. Crainic subdivides service network design into frequency service network design and dynamic service network design models. Frequency service network design models include transportation or load planning models, which can be used both to determine day-to-day operational policies and "for what-if questions raised ... in strategic planning." Dynamic service network design models are less strategic and "closer to the operational side of things."

Service network design research includes several prominent studies of LTL load planning. Crainic and Roy (1988), Roy and Delorme (1989) and Roy and Crainic (1992) discuss the NETPLAN model for service network design, freight routing and empty balancing. The model is similar to the path formulation of a capacitated multicommodity network design problem, but with a more general cost structure that includes transportation, consolidation, and penalties for capacity violations and missing service standards. The model is tested with data for two Canadian LTL companies with up to 35 terminals and almost 1000 origin-destination pairs.

Powell and Sheffi (1989) describe the APOLLO (Advanced Planner of LTL Operations) interactive DSS, which was implemented at a major LTL carrier (Ryder/PIE). This model focused on determining which direct services should be used between end-of-line terminals and break-bulk terminals, and between two break-bulk terminals. The solution approach is based on local improvement heuristics that add and drop services (links) in the network. Powell and Sheffi (1983) describe the load planning problem and a proposed network optimization model, while Powell and Sheffi (1986) highlight the benefits of having an interactive tool. Powell (1986) reports numerical experiments with 12 break-bulk terminals and over 1000 end-of-line terminals.

Braklow et al. (1992) describe SYSNET, a more comprehensive load planning system developed for one of the largest U.S. LTL carriers (Yellow Freight Systems). This was an extension and enhancement of the work for the APOLLO system. In addition to load planning, SYSNET has been used more strategically to examine questions such as break-bulk locations and capacities, whether to open end-of-lines, and deciding to which break-bulk an end-of-line terminal should be linked. All these strategic issues rely on having a good model for load planning. Bell et al. (2003) report that SYSNET has continued in use at Yellow Freight, in an evolved version, for over a decade.

Hoppe et al. (1999) address strategic load planning using a three stage solution strategy that utilizes a historic load plan to eliminate unlikely direct services, followed by a network construction phase based on the dual ascent approach of Balakrishnan et al. (1989), and then an add/drop heuristic. Numerical results are presented using real-world data sets from three different motor carriers with 48 to 92 terminals. These results demonstrate the value of having a historic load plan as a starting point — and the high quality of the historic load plans!

Dynamic service network design models include multiple time periods and use a space-time network to model schedules. Farvolden and Powell (1994) present a dynamic service network design model for general LTL transportation with 15 terminals and 18 time periods. Farvolden et al. (1993) use primal partitioning and decomposition to solve problems motivated by LTL trucking with 18 time periods and up to 30 terminals. Equi et al. (1997) provide a dynamic service network design model for transporting wood from cutting areas to ports.

In addition to the research on service network design for motor carriers, other applications of operations research to trucking include: LTL terminal layout and scheduling (Bartholdi and Gue, 2000, 2004; Gue, 1999), assigning drivers to loads for TL carriers (Powell et al., 1988), location and size of public terminals in congested areas in Japan (Taniguchi et al., 1999); fixed charge network design (Lamar and Sheffi, 1987; Lamar et al., 1990), and a large literature on freight routing (for example, see Akyilmaz, 1994; Crainic and Roy, 1992, Leung et al., 1990, and Lin, 2001).

Truck transportation is an important part of many multimodal systems. Research on air-ground multimodal network design for express package and postal delivery systems includes Barnhart and Schneur (1996), Cheung et al.!(2001), Grunert and Sebastian (2000), Grunert et al. (1999) and Kim et al. (1999). These models use trucks for collection and delivery and short-haul transportation, and aircraft for longer distance transport. For a review of intermodal rail-truck freight transport literature, see Bontekoning et al. (2004).

A final area of relevant literature is continuous approximation models for many-to-many transportation. This work reflects a somewhat higher level of planning than network design and provides analytical cost expressions to help determine the appropriate number of transshipments and terminals. Rather than treating input as discrete shipments between origins and destinations, it models demand as a continuous density function over a service region. For a review of relevant work on many-to-many transportation with transshipments, see Daganzo (1987, 1999), Hall (2003), and Langevin et al. (1996).

The following three sections of this paper review strategic network design models for LTL motor carriers, TL motor carriers and postal motor carriers.

3. Less-than-truckload network design

This section describes research on strategic network design for less-than-truckload (LTL) motor carriers. To keep a consistent set of notation and terminology we will refer to end-of-line terminals as "terminals" and break-bulk and consolidation terminals as "break-bulks." In various papers the end-of-line terminals are refereed to as depots, terminals, end-of-lines, satellite terminals, and branch offices; and the break-bulk terminals are referred to as hubs, operations centers, and sorting centers.

Haresamudra et al. (1995) describe BBNET (Breakbulk Network Software), an interactive decision support system for LTL network design. The primary focus is on finding break-bulk locations to minimize total transportation and handling costs. The software seeks to find a "near optimal design without the use of complicated mathematical programming alternatives." The package is developed in Turbo C as an extension of the HUBNET system developed for TL network design. (See the following section for details on HUBNET.)

The model includes transportation and handling costs based on input transportation and handling rates ($/lb/mile and $/lb, respectively). It assumes that adequate labor and real estate exist for the break-bulks, and that the capital requirements for different sites do not vary drastically. The required input data includes the origin, destination, and weight for each shipment, transportation and handling cost rates, handling times (min/lb), and average speed (mph). The user can specify up to 60 break-bulk locations, and the links between break-bulks, and can add or remove them interactively. Each origin and destination can either be assigned to the nearest break-bulk or the user can assign two degree × two degree latitude/longitude cells to a particular break-bulk. In addition, the user can specify the maximum number of break-bulks where a shipment is handled for each 500-mile trip increment.

BBNET determines routes based on the specified assignment of terminals to break-bulks using shortest paths through the links between break-bulks. It then calculates various performance measures. No algorithm is presented for locating break-bulks, but the authors suggest placing break-bulks in regions of high "freight density" (measured as freight flow in and out of each region) to reduce transportation cost and increase consolidation opportunities.

BBNET is validated with data from ABF Freight Systems, Inc. Numerical results are presented using disguised data for two months (average and high volume) with 10 break-bulks. The report states that the software is installed at ABF Freight Systems, Inc. where it "is being validated and verified for continued use."

Wleck (1998) describes an interactive DSS called "BOSS" used for design of LTL motor carrier networks in Europe, including location of terminals and break-bulks. Sparked by deregulation of motor carriers in Germany in the early 1990s, one strategy for small and medium size (regional) carriers was to join together to offer nationwide service in Germany and beyond. (This is very similar to the situation in the U.S. following deregulation a decade earlier.)

BOSS is used to address strategic questions, such as the number and location of terminals and break-bulks to minimize costs for facilities, transportation, and handling while meeting time standards. It uses approximations of transportation costs to allow quick evaluation of solutions. The model in BOSS assumes single assignment of customers to terminals and uses a specified maximum distance between customers and a terminal. The goal is to provide 24-hour transport between all customers. Regions served by terminals are compact and non-overlapping.

The solution method is designed to use various heuristics that can produce solutions in a "very limited computation time." It first finds an initial solution based on opening terminals that are close to the largest aggregated demands. Additional terminals may then be opened to ensure all customers are within a specified maximum distance of a terminal.

Initially a single break-bulk is opened to minimize the total distance to the initial set of terminals. Then, several stochastic and deterministic local neighborhood search heuristics are considered to find a set of improved terminals and break-bulks. Solutions near the best found are explored in more detail through a "1-opt" type exchange procedure. The solution algorithms are implemented in an interactive decision support system called BOSS.

Numerical results compare various heuristics with three sets of data for German motor carriers ranging from 35 thousand to 123 thousand customers, with up to 100 potential terminal locations and up to 50 potential break-bulk locations. Wleck also describes an application of BOSS with multiple cooperating carriers to evaluate questions concerning closure of a terminal and changing the number of terminals.

Results showed that all the heuristics performed similarly in terms of cost, but no lower bounds are available to evaluate the solution quality. Wleck states that four German carriers are using BOSS and that the algorithms "perform well in real life applications." He also argues that an interactive DSS is valuable for strategic network design to better allow many different scenarios to be examined in light of uncertain data, and to gain better insight into the sensitivity of costs and network structures to the various parameters.

Nagy and Salhi (1998) present a hub location-type model for many-to-many distribution with multistop collection and delivery tours. They include two types of vehicles: access vehicles for local delivery/collection, and linehaul vehicles for transportation between break-bulks. Each vehicle type has a volume capacity and a maximum distance/time per route. They seek to determine the number and location of break-bulks, and the local collection and delivery tours (by access vehicles) from break-bulks to origins/destinations, to minimize cost while satisfying demand and vehicle capacities. Costs include the fixed facility costs for break-bulks and transportation costs, which differ by vehicle type. The network includes direct links between all break-bulks, by assumption, so the routing of shipments between break-bulks is implicitly determined by the break-bulk locations. (The lowest cost path is a direct arc.)

Nagy and Salhi present a large integer linear programming formulation (an extension of location routing problem LR1 from Laporte, 1989), but do not solve it. Instead they present a decomposition approach where break-bulk locations are determined by an add/drop heuristic with tabu search. Routing for collection and delivery is based on the multi-depot vehicle routing heuristic in Salhi and Sari (1997). They report heuristic solutions with 249 customers and 10 break-bulks.

Other models for hub location and network design are relevant to LTL trucking, though most hub location research has been more focused on airline networks. For a recent review of hub location and network design, see Campbell et al. (2002). In a motor carrier context, hubs are break-bulks, and origins and destinations are end-of-line terminals. Early hub location models assumed two types of vehicles (often aircraft), where larger more efficient vehicles traveled between hubs and less efficient vehicles provided collection and delivery between the origins/destinations and the hubs. Hub location models have been examined for networks with single allocation (each terminals sends and receives all freight via one hub), multiple allocation (terminals may send and receive via more than one hub), arc and node capacities, and flow dependent costs. While most hub location research has focused on air networks, O'Kelly and Lao (1991) developed models for an intermodal (air-truck) two-hub express delivery network to determine where truck transportation should be used.

Tansel and Kara (2002) design a cargo delivery network that minimizes the delivery time of the latest item. This is formulated as an extension of the minimax hub location model that minimizes the arrival time of the last item (Kara and Tansel, 2001). Freight shipments follow a 3-leg route from the origin terminal (a branch office of the delivery firm) to the first break-bulk to a second break-bulk, then to destination terminal. Customers may drop-off and pick-up items at a terminal, or there may be local collection and delivery routes from the terminal. This local collection and delivery is not included in the model.

Terminals may be visited on routes with stopovers, and the model includes three types of route segments: "main lines" between terminals and break-bulks; "feeder lines" that visit several terminals and end at main lines; and "express lines" that connect two break-bulks. Main lines link one or more larger cities to a break-bulk and are served by large trucks. Feeder lines connect one or more smaller cities to the main line and are served by smaller trucks. Express lines are direct links between two break-bulks. Thus, a shipment may travel on a multistop feeder line from its origin terminal to another terminal on a main line, then on a main line to a break-bulk, then on an express line between two break-bulks, then again on a main line to the destination terminal (or to a feeder line that visits the destination terminal). Main lines and feeder lines may make multiple stops at terminals, but express lines make no intermediate stops. The problem is to determine the locations of break-bulks, the allocation of terminals to break-bulks, and the route structure between terminals and break-bulks with multiple stopovers and feeders, so as to minimize the arrival time of the latest arriving cargo at destinations. The total time includes travel times and waiting times at hubs.

To model the complex transportation with stopovers and feeder lines, there are three types of vehicles: express (type 0), main (type 1) and feeder (type 2) vehicles. A trip is defined as the path traversed from the node where an empty vehicle is initially loaded to the node where the vehicle is completely emptied. We now present the formulation.

Let $N = \{1, 2, \ldots, n\}$ be the set of nodes that serve as origins and destinations of freight. N is partitioned into two subsets where N_1 is all terminals that can be on a main line, and N_2 is all terminals that can be on feeder lines. Nodes in N_1 are handled by main line trucks; nodes in N_2 may be handled by feeder line trucks or main line trucks (if the node is on the route of a main line truck). All nodes in N_1 are potential break-bulks. Let A be the set of arcs in the transportation network that connect the nodes. Arcs for feeder lines, main lines and express lines are selected from A.

The model includes four sets of binary variables. Break-bulk location variables (y) indicate, for each potential break-bulk location, whether or not a break-bulk is established. Trip type variables (Z) indicate whether trips are with feeder line vehicles or main line vehicles. Trip arc variables (X) indicate which arcs are traversed on a trip. Service type variables (u) indicate whether each node is serviced by a main line truck or a feeder line truck. Thus:

$y_i =$ 1 if node i is a break-bulk, and 0 otherwise, where $i \in N_1$,

$Z^1_{ij} =$ 1 if a main line trip takes place between i and j with a type 1 vehicle, and 0 otherwise, where $i, j \in N_1$,

$Z^2_{ij} =$ 1 if a feeder line trip takes place between i and j with a type 2 vehicle, and 0 otherwise, where $i \in N_2$ and $j \in N_1$,

$X^{kl}_{ij} =$ 1 if the trip between i and j includes arc (k, l), and 0 otherwise, and

$u^r_{ij} =$ 1 if node r is served by a main line or feeder truck operating between nodes i and j, and 0 otherwise, where $r \in N$.

Define the following parameters:

$p =$ the number of break-bulks to be located,

$q_1 =$ the number of main line vehicles available,

$q_2 =$ the number of feeder line vehicles available,

$t_{kl} =$ the time to traverse arc (k,l) by a main line or feeder vehicle,

$r_i =$ the time that freight is ready at origin node i,

$\alpha =$ a scale factor to reflect reduced travel times on express lines: $\alpha \leq 1$,

$\delta =$ the time spent loading or unloading at each stop, and

$\gamma =$ the maximum allowable time for a feeder line trip.

Intermediate parameters calculated in the formulation are:

$A_j =$ the arrival time of a vehicle at node j,
$T_{hj} =$ the total trip time from h to j,
$D_h =$ the departure time of a main line vehicle from break-bulk h,
$\hat{D}_h =$ the latest time at which all incoming freight by main line trucks is available at node h.

The latest arrival time at a destination is denoted by Ω, which is given by the maximum of the A_j values. The formulation is:

Minimize Ω

Subject to

$$Z^1_{ij} \leq y_j \qquad \text{for all } i, j \in N_1 \tag{8.1}$$

$$Z^2_{ij} \leq 1 - y_j \qquad \text{for all } i \in N_2,\ j \in N_1 \tag{8.2}$$

$$\sum_{i,j \in N_1} Z^1_{ij} \leq q_1 \tag{8.3}$$

$$\sum_{\substack{i \in N_2 \\ j \in N_1}} Z^2_{ij} \leq q_2 \tag{8.4}$$

$$\sum_{j \in N_1} y_j = p \tag{8.5}$$

$$\sum_{\substack{i \in N, j \in N_1 \\ i \neq j, j \neq r}} u^r_{ij} = \begin{cases} 1 & \text{if } r \in N_2 \\ -1 & \text{if } r \in N_1 \end{cases} \tag{8.6}$$

$$u^r_{ij} \leq Z^1_{ij} \qquad \text{for all } i, j, r \in N_1 \tag{8.7}$$

$$u^r_{ij} \leq Z^t_{ij} \qquad \text{for all } i \in N_t,\ j \in N_1,\ r \in N_2,\ t = 1, 2 \tag{8.8}$$

$$\sum_{k:(k,l) \in A} X^{kl}_{ij} - \sum_{k:(l,k) \in A} X^{lk}_{ij} = \begin{cases} Z^t_{ij} & \text{if } l = j \\ 0 & \text{if } l \neq i, j \\ -Z^t_{ij} & \text{if } l = i \end{cases}$$
$$\text{for all } i \in N_t,\ j \in N_1,\ t = 1, 2 \tag{8.9}$$

$$X^{kl}_{ij} \leq Z^t_{ij} \qquad \text{for all } (k, l) \in A,\ i \in N_t,\ j \in N_1,\ t = 1, 2 \tag{8.10}$$

$$\sum_{a:(a,r) \in A} X^{ar}_{ij} - \sum_{b:(r,b) \in A} X^{rb}_{ij} \geq u^r_{ij}$$
$$\text{for all } i \in N,\ j \in N_1,\ r \in N \tag{8.11}$$

$$A_j = (D_h + T_{hj}) Z^1_{jh} \qquad \text{for all } j, h \in N_1 \tag{8.12}$$

$$\sum_{(k,l) \in A} t_{kl} X^{kl}_{jh} + \delta \sum_{r \neq h, j} u^r_{hj} = T_{hj}$$

$$\text{for all } j \in N_1,\ h \in N_1 \tag{8.13}$$

$$\hat{D}_h \geq (r_i + T_{ih}) Z_{ih}^1 \qquad \text{for all } i \in N_1,\ h \in N_1 \tag{8.14}$$

$$D_h \geq (\hat{D}_l + \alpha t_{lh}) y_h \qquad \text{for all } h, l \in N_1 \tag{8.15}$$

$$2\left[\sum_{(k,l)\in A} t_{kl} X_{ij}^{kl} + \delta \sum_{r \neq i} u_{ij}^r \right] \leq \gamma Z_{ij}^2$$

$$\text{for all } i \in N_2,\ j \in N_1 \tag{8.16}$$

$$\Omega \geq A_j \qquad \text{for all } j \in N_1 \tag{8.17}$$

$$X_{ij}^{kl} \in \{0,1\} \qquad \text{for all } i \in N,\ j \in N_1,\ (k,l) \in A \tag{8.18}$$

$$u_{ij}^r \in \{0,1\} \qquad \text{for all } i, r \in N,\ j \in N_1 \tag{8.19}$$

$$Z_{ij}^t \in \{0,1\} \qquad \text{for all } i \in N_t,\ t = 1,2,\ j \in N_1 \tag{8.20}$$

$$\hat{D}_h, D_h \geq 0 \qquad \text{for all } h \in N_1 \tag{8.21}$$

$$A_j, T_{jh} \geq 0 \qquad \text{for all } j, h \in N_1 \tag{8.22}$$

Constraint (8.1) forces main line trips to end at break-bulks, and constraint (8.2) forces feeder line trips to end at non-break-bulk nodes. Constraints (8.3) and (8.4) enforce the limits on the availability of vehicles. Also, it is assumed that the number of express vehicles available allows for a direct trip between each pair of break-bulks, so the number of express vehicles is at least $p(p-1)/2$. Constraint (8.5) requires that p break-bulks be established. Constraint (8.6) ensures that all non-break-bulk nodes r are assigned to a truck, and that no break-bulk nodes are assigned to main line or feeder trucks. Constraints (8.7) and (8.8) ensure that a trip is established between nodes i and j, whenever any node r is assigned to it. Constraint (8.9) is the flow conservation equation and constraint (8.10) assures that if arcs are assigned to a trip, then the trip must exist. Constraint (8.11) assures that for every node visited by a trip, there is some arc in or out of the node. Constraints (8.12) and (8.13) establish the arrival times of vehicles, based on departure times, travel times, and loading/unloading times. Constraints (8.14) and (8.15) are nonlinear constraints that establish the departure times for trucks at each node. Constraint (8.16) enforces the maximum time limit for feeder lines and constraint (8.17) sets the latest arrival time at the end of a main line trip. Constraints (8.18) – (8.22) limit the values of decision variables and intermediate parameters appropriately.

This model includes two different service types (main line and feeder), three different vehicle and trip types (main line, feeder and express), as well as time limits for feeder line trips. It assumes trucks capacities are not an issue, though main line and feeder line capacities are discussed and constraints are provided. The formulation includes several nonlinear constraints (8.12), (8.14) and (8.15) for which the authors provide linear forms.

This formulation is not solved in Tansel and Kara (2002). However, when the feeder component is removed and the number of main line vehicles is unrestricted (remove constraint 3), then this reduces to the latest arrival hub location problem (Kara and Tansel, 2001). In this case, the set N_2 is null and there are two types of vehicles: express vehicles operating directly between two break-bulks, and main line vehicles operating directly between a break-bulk and a terminal. Kara and Tan (2003) present some solutions for ground transportation of parcels in Turkey. (Note that air transportation is not needed in Turkey to provide a high level of service due to the small size of the country, and the good infrastructure for ground transport.) Results show that four well-located break-bulks (instead of the 25 currently used) can reduce the latest delivery time by almost two hours, as well as the number of vehicles required and the fuel consumed.

Bartholdi and Dave (2002) and Bartholdi et al. (2003) report on the development of a network design tool for LTL carriers. Bartholdi and Dave (2002) describe a "visual, user friendly tool, NetworkDesigner®, that generates the hub-and-spoke distribution system." No details on the model or solution algorithm are provided, but the report mentions the use of "custom heuristics based on problem structure ... implemented within the commercial MIP solver." This was developed to redesign the network at RPS (now FedEx Ground) and the report states that it generates "robust solutions that compare favorably with solutions generated by a commercial model being used by FedEx Ground." Though no details are provided, some of the questions that can be addressed with the tool "include break-bulk location."

Bartholdi et al. (2003) provide details on a model to assign terminals to break-bulks and route LTL freight through the network. This model explicitly includes the use of a truck (tractor) pulling two 28 foot "pups" between break-bulks. It assumes these pups can be used for local collection and delivery, though it does not model local collection and delivery. Because the total daily volume between each pair of terminals is less than a full truckload (2 pups), break-bulks are used to consolidate shipments. Each terminal is assigned to one hub and the basic decision is: To which break-bulk should each terminal be assigned? These assignments need to be determined before, or concurrently with, the break-bulk locations. This paper focuses on the assignment question; break-bulk location is not explicitly discussed.

The solution approach used a greedy heuristic for initial assignment of terminals to break-bulks, then an improvement heuristic to consider skipping an intermediate break-bulk — or skipping sorting at a break-bulk. The greedy heuristic seeks to minimize the approximate transportation and sorting costs, by assigning terminals in decreasing order of "freight intensity" (the sum of the freight in and freight out), where each assignment must satisfy certain business rules (e.g., driving hours, sorting time and capacity).

Because of consolidation at break-bulks, most trailers between break-bulks are fully loaded. This results in all paths visiting either one or two break-bulks for sorting. The improvement heuristic considers direct paths, as well as paths via break-bulks, but without the sorting at a break-bulk. Trailers sent on direct paths (not sorted at a break-bulk) must utilize at least a minimum percentage of capacity (75% for FedEx Ground). For example, if one trailer (one "pup") at an origin terminal can be filled for a specific destination terminal, then that trailer need not be opened and sorted at any intermediate break-bulks. It might travel direct from the origin terminal to the destination terminal, or via one or two break-bulks with another pup, that is opened and sorted.

The model does not allow multiple stops at terminals on route to/from a break-bulk, but it does permit shipments between an origin and destination (o-d) to be split over multiple routes. (There is not a unique path for an o-d pair.) Origins and destinations are terminals, and the model defines a freight flow variable for each possible path type for each origin-destination pair. It also includes variables for the flow of trailers and trucks, where a tractor can pull two trailers. A lengthy Integer Linear Programming (ILP) formulation is provided that minimizes transportation costs plus sorting costs at break-bulks. Small instances were solved using CPLEX 7.5, but it was "difficult to solve even small problems with 3 break-bulks and 30 terminals." (The FedEx Ground network had 388 terminals and 24 break-bulks!)

To find solutions for problems of realistic size in reasonable time, they partitioned the problem based on break-bulk pairs and associated terminals (termed a "dyad"), and then solved the routing problem for each dyad. Each dyad consisted of two break-bulks and a number of terminals (usually 20 – 40). Thus, any freight between two terminals assigned to the two different break-bulks shows up in exactly one such dyad. (This was about 89% of the freight in the data set used.) However, any freight between terminals assigned to the same break-bulk shows up in many such dyads — and might be routed differently in different dyads! In the computational experiments, these different routings occurred rarely.

For each dyad they determined the freight routing using parallel computing on a cluster of commodity computers (up to 128 Pentium processors). With 24 break-bulks, there are 276 dyads ($276 = 24 \times 23/2$). They report that a "typical run" took 6 hours and used a total processing time of 457 hours. However, 17% of the dyads (46/276) were not solved within 10% of optimality, and no integer solution at all was found for 10 dyads ($10/276 = 3.6\%$).

Typical results showed that about 34% of packages are double sorted at two break-bulks (vs. 89% in the initial solution), 34% are routed via two break-bulks, but sorted only at one break-bulk (usually the 2nd one visited), 22% of packages are routed via a single break-bulk where they are also sorted, and about 9% of packages are not sorted at any break-bulk (though they may be routed via two, one or zero break-bulks). One interesting finding was that more trailers were routed direct from the origin terminal to the break-bulk of the destination (bypassing the break-bulk of the origin terminal), than the other way around.

4. Truckload network design

This section describes research on strategic network design for truckload (TL) motor carriers. TL carriers may operate without a network by dispatching a driver sequentially on a long tour of point-to-point trips. For efficiency, the carrier would like to find a sequence of trips that minimizes the empty miles traveled from the destination of one trip to the origin of the subsequent trip. Such tours may take a driver away from home for 14–21 days (Taylor et al., 1999), and lengthy tours have led to high turnover rates among TL drivers. Hunt (1998) reports driver turnover rates for TL carriers as high as 200%, in contrast to rates often less than 10% for LTL carriers. High rates of driver turnover both increase training costs (estimated at $3000 to $5000 per driver in the U.S.) and accident rates (Hunt, 1998).

Some TL carriers have developed relay networks, where terminals serve as relay point at which drivers can exchange loads (trailers). A relay network can produce much shorter driver tour lengths, and can help increase efficiency by allowing the load to continue moving with another driver while the first driver rests. Disadvantages of relay networks include the extra distance that might be traveled via the terminals, and the added time for swapping loads.

To keep a consistent set of notation and terminology we will refer to relay terminals as "terminals". In various papers these are refereed to as relay points, hubs, and transshipment points. Note that terminals for TL networks do not involve the loading, unloading and sorting functions of break-bulks in LTL networks. Terminals serve primarily as places for drivers to swap trailers. Thus, it is possible for these terminals to be simple facilities such as public rest areas or private truck stops.

Only a few authors have addressed TL network design. Meinert and Taylor (1999) summarize a number of studies that have been carried out over the past decade using data for the largest U.S. TL carrier. In general, this work uses simulation models to explore different strategic and operational concerns. The earliest work on relay networks for TL carriers involves the HUBNET interactive simulation tool developed for J.B. Hunt, Inc. Taha and Taylor (1994) provide an overview of this work, including results of preliminary testing. This paper also highlights the differing motivations for hub-and-spoke-like networks in LTL and TL trucking. They identify a key tradeoff for TL networks as whether or not the added circuity to travel via hubs is offset by the decrease in costs associated with reduced driver turnover.

HUBNET is a simulation system to evaluate relay terminal networks for TL trucking. It provides interactive tools to help the user construct a network, and then it simulates TL operations. HUBNET assume three types of drivers: local drivers for collection and delivery between the terminals and the shipment origins/destinations, lane drivers between terminals, and non-network drivers for loads that would exceed the maximum circuity if sent via the network. (Note that not all loads are sent via the network.) Local drivers are based at a terminal, and lane drivers travel along the network to one terminal before returning to their home terminal. Rather than treat each demand point individually, HUBNET divides the U.S. into sixty-five two degree × two degree latitude and longitude geographic regions. It calculates freight density and load imbalances for each region to assist in the network design.

HUBNET is designed to address three problems: location of terminals, determination of which terminals to connect with direct routes, and determination of the geographic service area for each terminal. Explicit solution algorithms for these problems are not provided but it states that the solutions "use load volume and geographical distance considerations to suggest initial hub (terminal), spoke, and area layouts, but allow for significant user interaction ... "

Three important factors for finding terminal locations are identified: (1) locate in or near "high volume geographical regions"; (2) place hubs at "almost equal distances across the service area," so that drivers can be "assigned to runs equal to some fractional or complete multiple of a shift duration" to "maximize driving time while returning home much more often"; and (3) the location of existing terminals. The final suggestion is that "perhaps a hybrid of each of the three above considerations should drive hub (terminal) location."

Direct links between hubs are selected based on distance and load volume. For the most part, nodes that are less than a one shift drive apart are connected via direct routes, but there are exceptions if intermediate nodes are on the direct path, and for very high volume nodes. The size and shape of service regions are based on load volume, proximity to other hubs, roadways and geography.

HUBNET provides an initial solution superimposed on a map of the U.S. The user can then add or remove terminals and direct links, and re-allocate service regions to terminals based on the two degree × two degree latitude/longitude cells. Then primary function of HUBNET is to simulate TL operations with the interactively designed network, and to generate performance measures to compare the relay network and point-to-point operations. Thus, the input to the simulation phase is an order history (demand), a relay network with up to 60 nodes, the service area of each terminal (specified by two degree × two degree regions), the number of drivers available, the percentages for each of three types of drivers, and the maximum allowable circuity (as an excess mileage percentage). HUBNET assigns drivers to terminals based on local demand and uses shortest paths for travel between terminals. For more details on the software, see Taha et al. (1996), which describes the local module for intra-hub area driver assignments and load assignment; and the freight lane module for inter-hub transportation.

Results are reported for two networks to serve the U.S.: one with 24 terminals and one with 32 terminals. Results show the average tour length can be "drastically reduced" from about 18 days with the current point-to-point operations to 2 days or less with a relay network. However, the circuity increases from 3.5% to 15% with a network, and the "first dispatch empty miles" also increase from 5.6% to about 15% with the network.

Taylor et al. (1995) describes the use of HUBNET to evaluate terminal location methodologies, the number of terminals, and a policy that restricts drivers to a particular traffic lane. They compare three terminal location methodologies: "distance-based," "flow-based," and "hybrid-based." For distance-based location, terminals are located one day apart — but no method is provided. For flow-based location, terminals are placed in regions "characterized by low imbalance between originating and destinating loads". (As earlier, this is based on a partition of the U.S. into 65 grid cells based on latitude and longitude.) To do this, the authors provide the following small IP that "minimizes the total freight imbalance of a user-specified number of selected regions".

Let $X_j = 1$ if a terminal is assigned to grid j; and 0 otherwise. Let C_j be the load imbalance for grid j (equal to total loads in − total loads out) and A_{ij} is 1 if grid i is contiguous to grid j; and 0 otherwise. Let p be the desired number of terminals.

$$\text{Minimize } Z = \sum_{j=1}^{65} |C_j| X_j$$

Subject to:

$$\sum_{j=1}^{65} A_{ij} X_j \geq 1 \qquad \text{for all } i \tag{8.23}$$

$$\sum_{j=1}^{65} X_j = p \tag{8.24}$$

$$X_j = \{0, 1\} \qquad \text{for all } j$$

The objective minimizes the sum of absolute values of freight imbalances. Constraint (8.23) ensures that each grid cell either contains a terminal or is adjacent to a cell with a terminal. Constraint (8.24) forces the number of hubs to be the desired value (24 or 32 were used in the computational results).

The third terminal location methodology, hybrid-based, is a combination of "heuristics, expert judgment, and the location of existing terminal locations for J.B. Hunt Transport, Inc."

HUBNET is used to simulate operations with either 24 or 32 terminals, whose locations are derived from the three location methodologies, and with policies that allow drivers to travel to one or two terminals from home, before transferring loads. Five primary performance measures are calculated including lane and local driver tour length, average miles per driver per day, first dispatch empty miles, and average circuity as a function of trip miles.

The best results were with 32 terminals, hybrid-based locations, and a policy that restricts drivers to travel between two adjacent terminals. In comparison to the current point-to-point method of operations, this reduces average tour length by 90% (to about 2 days), and total miles per driver by 14.6%. However, circuity and first dispatch empty miles increase. Further testing showed that the best scenario is when 53% of the loads are moved via the network (vs. point-to-point). This suggests that limited implementation of a relay terminal network could be worthwhile, to allow some shipments to travel direct while others use the network. As for the best method for network design, in this paper expert judgment was preferred. The authors state that J.B. Hunt is experimenting with a zone delivery system where drivers return home at least one day each week. They also say that

> "real-time optimization technology is used to identify ... beneficial load switches, along with recommended switch points and times, based on current positions and final destinations. In a sense, this allows a ... network to be implemented with a nearly infinite number of hubs since truck stops, rest areas, and existing terminal yards are used as switch points."

Because the research with HUBNET indicated that a partial relay network was most promising, Taylor and co-authors followed up with a series of papers addressing different alternatives for implementing a limited network. One key theme in these works is the desire to restrict drivers to lanes between terminals or to geographic zones. These approaches will reduce tour length for the driver (and hence turnover), but will generally increase the total distance traveled, as routes are longer than point-to-point.

Taylor et al. (1999) examines a region in the southeastern U.S. and compares seven alternatives:

(1) point-to-point routes.
(2) a southeast zone with 6 zone perimeter terminals "in locations that provide access to major highways and existing freight corridors."
(3) one "key lane" in and out of the southeast region.
(4) another "key lane" in and out of the southeast region.
(5) two "key lanes" in and out of the southeast region.
(6) one "key terminal" in the center of the southeast region.
(7) a "hybrid model" with 1 central terminal and 6 perimeter terminals.

The point-to-point scenario is the default condition where drivers haul a sequence of TL moves from origin to destination. This produces long tours where drivers are on the road for 2 – 3 weeks at a time. The zone scenario allows drivers to stay within the southeast region by exchanging loads at the perimeter terminals. In the three "key lane" scenarios a percentage of loads are sent via drivers shuttling back and forth along a high traffic corridor between Atlanta, Georgia (near the center of the southeast region), and another city providing good access to points outside the southeast. The "key terminal" scenario uses a single terminal in Atlanta, rather than the 6 perimeter terminals to exchange loads. The "hybrid model" combines the "key terminal" and zone model.

For each scenario, the authors simulate one week of operations and collect performance measures for drivers, carriers and customers with four key metrics and eleven secondary metrics. The key carrier performance metrics are percentage circuity (actual miles compared to point-to-point miles) and first dispatch empty miles (average number of empty miles from dispatch to load pickup). The key driver performance metric is the average number of miles per driver per day. (This affects driver pay and turnover.) The key customer performance metric is the percentage of loads that are delivered late.

Results showed that different scenarios were preferred by the different stakeholders (drivers, carrier and customers), but the zone model appeared to provide the best overall solution in this region when considering the driver, carrier and customer objectives together. The authors also considered having more or fewer terminals and concluded that 4-6 terminals in the southeast region seem to "offer the best compromise solutions relative to all four of the key metrics." They then considered the northeastern region of the U.S. and found similar results; and even stronger evidence for a zone scenario in some cases, due to the more isolated nature of the northeast relative to the rest of the U.S.

Finally, they reported that J.B. Hunt is using a key lane approach in the eastern U.S. and a zone system in the northeast U.S., and that these have reduced turnover rates from 53% for the general point-to-point drivers to 22% for those with regular routes or zones. Meinert and Taylor (1999) mention consideration of a national network of zones, though they provide no details.

Taylor and Meinert (2000) further examine a zone strategy from the perspective of the customer, the carrier, and the driver. They seek strategies that can improve quality for driver (job quality), customer (on time pickup and delivery) and carrier (lower turnover). They provide an experimental design to evaluate how the number of terminals, the length of haul from zone centroid to the terminal and back, and the distribution of freight within the zone affect a zone-based network. They develop a simulation model (in SIMNET II) for an idealized rectangular two-zone system. They consider $1-4$ terminals evenly spaced along a 500-mile boundary, average hauls of 400, 600 or 800 miles, and a uniform and concentrated demand distribution.

The model generates demand patterns, simulates operations (using the U.S. Department of Transportation driver work rules), and calculates performance measures for the driver, customer and carrier. It includes rules to determine whether the load is sent direct or via a terminal. (Generally, longer trips and those with less circuity are sent via a terminal.) Results shows that the zone model reduces flow times and can improve on-time service. Taylor et al. (2001) also consider zone dispatching. Simulation results indicate that multi-zone dispatching works best when zone boundaries are configured to minimize, to the extent possible, the freight imbalance between zones

Hunt (1998) provided a different model for designing a relay network for TL motor carriers. The general approach involves first routing freight flows over a roadway network and then locating relay terminals at intervals along the network. Drivers relay loads between adjacent terminals. (Hunt also provides some interesting historical background on ancient relay networks of the Persians, Romans, and Chinese, as well as the Pony Express system in the U.S. from the mid-1800s.)

The solution approach is a four step process, using the underlying U.S. interstate highway system as the physical road network. The first step routes freight flows across the physical network. The second step is to create the relay network by locating relay points on the physical network. The third step is to route the commodities across the relay network, and the final step is to assign drivers to relay points. We will focus on the first three steps below.

The input includes the demand (origin, destination, and time window), the physical network (e.g., U.S. interstate highway system) and the desired, minimum and maximum distances between relay points. Hunt considered several methods of routing freight across the network, each of which may produce a different relay network. The simplest method is an independent shortest path algorithm that ignores interactions and backhaul opportunities. Several other methods presented try to create improved (lower cost) routings by accommodating backhauls. These include solving IP formulations, using a dependent shortest path algorithm that routes and re-routes commodities to try to improve backhauls, a shortest path tree algorithm, and a linear programming relaxation.

The second step of locating relay points uses the freight routes from the first step as input. The problem is then to determine the smallest set of relay points (i.e., fewest) along routes, such that the travel distance between adjacent relay points is between the specified minimum and maximum distances and the travel distance between the end points (origins and destinations) and the closest relay point is less than half the specified maximum distance. (Hunt suggests that for full-day driving the minimum distance be 300 miles and the maximum distance be 500 miles.)

Hunt describes two algorithms for locating relay points: the "Spring Algorithm," that tries to iteratively improve a feasible solution; and a greedy algorithm that iteratively adds relay points one at a time. We present the Spring Algorithm first.

The idea behind the Spring algorithm is inspired by the forces of attraction from a stretched spring and repulsion from a compressed spring. For example, two terminals that are closer together than the minimum distance would experience a repulsive force, while two terminals farther apart than the maximum distance would experience an attractive force. The algorithm begins by creating a feasible solution that places terminals along each route the desired distance apart. It then calculates "spring forces" between all adjacent terminals and between terminals and adjacent origins/destinations. These may be attractive or repulsive depending on the spacing between terminals. It then calculates "gravitational forces" between pairs of terminals on different routes but nearby (e.g., within 200 miles). Finally it combines spring forces and gravitational forces for all terminals and calculates new positions for terminals as the projection of the sum of forces along the route. Then any terminals in close proximity (within a specified distance of each other) are combined into a single terminal.

Hunt also considered a greedy algorithm based on identifying feasible terminal "windows" along the roadways for each route, that take into account the minimum and maximum distances between terminals. Because routes may overlap, several windows may overlap. The greedy approach selects terminal locations that "cover" the most uncovered sets of windows one at a time until all are covered.

The Spring algorithm and the greedy algorithm were implemented using an object oriented design in Java and C++. The methods were tested on small problems using data for the southeast U.S., where the interstate system included 251 nodes and 329 edges. Demand was based on test data from the U.S. Postal Service that contained up to 50 origin-destination pairs. Any origin-destination pairs less than 250 miles apart were treated separately outside the relay terminal network. Results showed that the Spring algorithm consistently produced fewer relay points that the greedy algorithm (ranging from 17 to 63 for various problems), but did take more cpu time. However, this may not be an important factor for strategic network design.

Once the relay network is established, then the freight must be routed via the relay terminals. This is similar to the first step and the same solution approaches as in step one are used, but now all freight must be routed via at least one relay terminal. From the resulting freight flows, the traffic on individual "legs" between adjacent relay points and between origins/destinations and relay points can be calculated. This is then used to determine the number of drivers to assign to each terminal, where each driver travels no farther than the adjacent relay terminal.

Results showed that different initial freight routings did produce different relay networks, but the "majority of loads required less than 25 extra miles for travel via the relay networks (vs. direct point-to-point routes), and the majority of loads had equal or improved service times. However, in some cases the maximum excess miles was very large (over

20%). Thus, some loads should probably not be sent through relay network.

Hunt suggests (as did the earlier work with HUBNET) that TL carriers might operate partial relay networks, where some loads are sent through relay terminals and others are sent point-to-point. Hunt also mentions some problems and areas for future research in implementing the Spring Algorithm, including some caused by network structures that prevent the algorithm from escaping local optima.

5. Postal network design

This section describes research on strategic network design for postal motor carriers. While there is much research on integrated or intermodal parcel and express carriers that combine air and truck, it is beyond the scope of this chapter.

Donaldson et al. (1999) provide a model to design the network for first class mail transport in the U.S. The origins and destinations of shipments are 148 area distribution centers (ADCs) that serve local post offices. Any origin-destination pairs over 1800 miles apart must be served by air and are not considered. Also, local mail within the metropolitan area of an ADC is not considered. Mail may be sent direct from an origin to destination if demand is sufficient, or via one crossdocking center (i.e., transshipment terminal). Service levels are specified so that mail for origins and destinations less than 600 miles apart should be delivered in two days; and mail for origins and destinations less than 1800 miles apart should be delivered in three days.

The fundamental problem is to locate crossdocking centers to minimize the total transportation cost. The authors formulated an IP to calculate transportation costs for a given set of origins, destinations and crossdocks. They solved the IP for various specified sets of crossdocks to find the "best" set. The formulation is presented below using the following variables and parameters:

$I = \{i\}$ is the set of origin nodes,
$J = \{j\}$ is the set of destination nodes,
$K = \{k\}$ is the set of crossdock nodes,
$x_{ij}^k =$ flow on the path from origin i to destination j through crossdock k,
$x_{ij} =$ direct flow from origin i to destination j,
$R_{ij} =$ number of trucks on link (i,j) from origin i to destination j,
$O_{ik} =$ number of trucks on link (i,k) from origin i to crossdock k,
$D_{kj} =$ number of trucks on link (k,j) from crossdock k to destination j,
$c_{ij} =$ cost of sending a truck from i to j,

$C =$ truck capacity,
$S_{ij} =$ flow (demand) from origin i to destination j.

Any full truckloads from origin i to destination j will be shipped direct and are not included in the formulation. Thus, S_{ij} includes only partial truck loads. The formulation is:

$$\text{Minimize } \sum_{i,j} R_{ij}c_{ij} + \sum_{i,k} O_{ik}c_{ik} + \sum_{k,j} D_{kj}c_{kj}$$

Subject to:

$$\sum_k x_{ij}^k + x_{ij} = S_{ij} \qquad \text{for all } i \in I,\ j \in J \qquad (8.25)$$

$$\sum_j x_{ij}^k \leq CO_{ik} \qquad \text{for all } i \in I,\ k \in K \qquad (8.26)$$

$$\sum_i x_{ij}^k \leq CD_{kj} \qquad \text{for all } k \in K,\ j \in J \qquad (8.27)$$

$$x_{ij} \leq R_{ij}S_{ij} \qquad \text{for all } i \in I,\ j \in J \qquad (8.28)$$

$$x_{ij} \geq 0 \qquad \text{for all } i \in I,\ j \in J \qquad (8.29)$$

$$x_{ij}^k \geq 0 \qquad \text{for all } i \in I,\ j \in J,\ k \in K \qquad (8.30)$$

$$R_{ij} = \{0, 1\} \qquad \text{for all } i \in I,\ j \in J \qquad (8.31)$$

$$D_{jk}, O_{ik} \text{ nonnegative integers} \qquad \text{for all } i \in I,\ j \in J,\ k \in K \qquad (8.32)$$

The objective minimizes total transportation cost. Constraint (8.25) ensures that all destinations are satisfied either via a direct link or a crossdock. Constraints (8.26) and (8.27) establish the number of trucks to carry the flows through crossdocks. Constraint (8.28) establishes the number of direct trucks. Constraints (8.29) – (8.32) restrict the variables to be nonnegative and integer, as appropriate.

Before solving this IP, there is preprocessing to generate only the feasible direct links and paths through crossdocks, based on travel times, handling times at crossdocks, and specified service levels. Several solution approaches were tried, including branch and bound, Bender's cuts, and a relaxation heuristic. Only this last approach was efficient enough for the real-world problems considered.

The relaxation heuristic relaxes the integrality constraints on the links from origins to crossdocks or from crossdocks to destinations This allows the problem to be decomposed by either origin or destination, and though it does not guarantee optimality, it did produce small gaps on the problems considered. The solution procedure is to iteratively solve single commodity problems from one origin to all destinations, where the truck variables are integer on direct links, and on links from origins to crossdocks, but not on links from crossdocks to destinations. These single origin solutions are then combined by adding the flows on common links. This may produce fractional truck variables on links from crossdocks to destinations, and this provides a lower bound. Fractional values are rounded up to provide an integer solution and an upper bound. Any crossdock-to-destination links whose flow is below a specified threshold are then eliminated. The procedure stops when the gap between the upper and lower bounds is small enough. The authors considered relaxing the integrality constraints between either the origin and crossdock or between the crossdock and destination, but the results were similar.

Results were provided for two situations. The first considered where to locate a single crossdock center to serve the southeast and mid-Atlantic U.S. There were 36 origins/destinations (ADCs) in eleven states, and three possible crossdock locations were considered. The problem was solved for each crossdock location using the relaxation heuristic and all three gaps were about 4%. The best heuristic solution was within 0.1% of the optimal solution (found by branch and bound).

The second analysis considered where to locate crossdocks for the entire U.S. For this analysis there were 148 origins/destinations (ADCs). Nineteen different sets of crossdocks were considered, ranging from a single crossdock in one of five different cities to a set of 22 crossdocks. The relaxation heuristic solved all problems in reasonable time. The gaps depended on the candidate crossdocks sets and ranged from 1% for one crossdock to almost 20% for the twenty-two crossdock problem. Though transportation costs decreased with larger numbers of crossdock centers, the results showed little improvement with more than 4 or 5 crossdocks. Note that while realistic problems of continental scale could be solved, the crossdock locations were inputs to the model, and the best set of crossdocks examined is not necessarily the best set that exists.

Ernst and Krishnamoorthy (1996, 1999) developed hub location models for the design of motor carrier postal networks that focus specifically on the locations of mail consolidation and sorting centers. Ernst and Krishnamoorthy (1996) introduce the use of hub location models in postal network design for Australia Post. The model includes the collection of mail from postcode districts (origins) to a mail sorting center (hub), the transfer of mail between sorting centers (hubs), and then distribution from a sorting center (hub) to the destination postal district (destination). Because each of these three components may involve a different type of transportation (e.g., size of motor vehicle), there are separate cost coefficients for each type of transport.

Ernst and Krishnamoorthy (1996) provides an efficient formulation for the single allocation p-hub median problem (Campbell, 1996), which restricts each origin and destination to a single sorting center. The problem is formulated on a complete graph $G = \{V, E\}$ with node set $V = \{v_1, v_2, \dots, v_M\}$, where nodes correspond to origins and destinations (i.e., postal districts) and potential hub locations. Let d_{ij} be the distance from node i to node j and let W_{ij} be the volume of mail to be transported from i to j. Distances are assumed to satisfy the triangle inequality. Let p be the number of hubs to locate. Mail travels from the origin to a hub, possibly to a second hub, and then to the destination. Three cost parameters χ, α, and δ are the unit cost for transportation from an origin to a hub (collection), between two hubs (transfer), and from a hub to a destination (distribution), respectively. Generally, shipments are consolidated at the hubs to exploit the economies of scale, so $\alpha < \chi$ and $\alpha < \delta$. For the Australia Post application, the respective values are: $\chi = 3$, $\alpha = 0.75$, and $\delta = 2$.

The decision variables are:

$Z_{ik} = 1$ if node i is allocated to a hub at node k, and 0 otherwise, and
$Y^i_{kl} =$ flow from hub k to hub l that originates at node i.

Thus, a hub is located at node k if $Z_{kk} = 1$. The total flow originating at origin i is:

$$O_i = \sum_{j \in V} W_{ij} \quad \text{for all } i \in V,$$

and the total flow destined for destination i is:

$$D_i = \sum_{j \in V} W_{ji} \quad \text{for all } i \in V.$$

The formulation to minimize total transportation costs is:

$$\text{Minimize} \sum_{i \in V} \sum_{k \in V} d_{ik} Z_{ik} (\chi O_i + \delta D_i) + \sum_{i \in V} \sum_{k \in V} \sum_{l \in V} \alpha d_{kl} Y^i_{kl}$$

Subject to:

$$\sum_{k \in V} Z_{kk} = p \qquad (8.33)$$

$$\sum_{k \in V} Z_{ik} = 1 \qquad \text{for all } i \in V, \qquad (8.34)$$

$$\sum_{j \in V} W_{ij} Z_{jk} + \sum_{l \in V} Y^i_{kl} = \sum_{l \in V} Y^i_{lk} + O_i Z_{ik} \qquad \text{for all } i, k \in V, \qquad (8.35)$$

$$Z_{ik} \le Z_{kk} \qquad \text{for all } i, k \in V, \qquad (8.36)$$

$$Y^i_{kl} \ge 0 \qquad \text{for all } i, k, l \in V, \qquad (8.37)$$

$$Z_{ik} \in \{0, 1\} \qquad \text{for all } i, k \in V, \qquad (8.38)$$

The objective is to minimize the total cost for collection, distribution and transfer. Constraint (8.33) ensures that exactly p hubs are selected, and constraints (8.34) ensure that each origin/destination is allocated to a single hub. Constraints (8.35) enforce flow conservation at the hubs. Constraints (8.36) ensure that every allocation establishes a hub. Constraints (8.37) and (8.38) restrict the variables appropriately.

The authors solve this formulation using branch and bound with an upper bound based on simulated annealing for the Australia Post data set, based on postal operations around Sydney, Australia. They find optimal solutions for problems with up to 50 origins/destinations and five hubs. For larger problems with up to 200 origins and destinations and 20 hubs, they find near-optimal solutions (within 1%) using a simulated annealing-based heuristic.

Ernst and Krishnamoorthy (1999) extended the single allocation hub location problem to include capacities on the flow being sorted at the hubs (not the total flow through the hub). They also replaced the specification of exactly p hubs, by a fixed cost for hubs in the objective, so that the model would determine both the number and locations of hubs. Capacities are specified by a parameter Γ_k and the following constraints are added to the formulation:

$$\sum_{i \in V} O_i Z_{ik} \leq \Gamma_k Z_{kk} \quad \text{for all } k \in V.$$

Although the capacitated hub location problems are more difficult to solve than the corresponding uncapacitated hub median problems, the authors provide optimal solutions for problems with up to 50 origins and destinations, using two levels of fixed costs for hubs and two levels of capacity. Generally, tightening the capacities increases the cpu times, as well as the optimal number of hubs.

There has been considerable subsequent research on a wide range of hub location problems, including those with single and multiple allocation, node and arc capacities, flow thresholds, flow-based cost functions, and more general network structures. In general, hub location problems explicitly model different vehicle types (and costs) to reflect consolidation activities, and address strategic network design including location of consolidation or sorting centers and selection of network links. These types of networks are common for a variety of transportation systems, including LTL motor carriers and postal system. Although much of the hub location research is relevant to the design of LTL and postal networks, and much of it uses the Australia Post data set for testing, this literature is generally more algorithmic and theoretical, rather than being applied explicitly to design motor carriers networks. (The recent work by Kara and Tansel discussed in the section on LTL network design is an exception.) See Campbell et al. (2002) for a recent review of hub location research.

6. Conclusion

Motor carriers provide a vital function in modern societies, and their importance is likely to grow with the increasing customer requirements for better service and reduced cycle times. There is a vast amount of operations research work on tactical planning (e.g., load planning) and operational planning (e.g., vehicle routing) for motor carriers, but somewhat less attention on strategic planning, including strategic network design. This chapter provides a survey of relevant published work on strategic network design for less-than-truckload, truckload and postal motor carriers. The focus has been on research with a strong link to *motor carrier* network design, not on general network design or on primarily algorithmic advancements. While the published research conveys a range of models and solution techniques to address different problems, many for major motor carrier firms, there may well be other significant models and results currently in use by carriers, but not described in the literature.

Motor carriers are large complex organizations that must serve a varying demand over a large geographic area in a very competitive and dynamic environment, often with tight service constraints. These environmental pressures generate a need for future research in a variety of areas. One area for future work is to better address the merger of motor carrier networks. This may result from standardization of local and national transportation regulations as international trade rules are liberalized — and from the increasing concentration in the industry through mergers and acquisitions. A second area for future work is strategic network design for time definite trucking, in which motor carriers provide "guaranteed" service of one, two, three, ... days between specified origins and destinations. This market allows motor carriers to exploit their cost advantage over air carriers for deferred airfreight. A third area is in application of hub location research to less-than-truckload network design. There is a great deal of research on optimal hub location and network design that is more theoretically oriented, than practically oriented, and more tuned to air transportation, than ground transportation. Extending this work with applications for LTL carriers could be quite beneficial. Finally, given the current size of large motor carriers, and trends for them to become even larger, research is needed to help solve larger problems of practical size.

References

Akyilmaz, M.O. (1994). An algorithmic framework for routeing LTL shipments. *Journal of the Operational Research Society*, 45:529 – 538.

Balakrishnan, A., Magnanti, T.L., and Wong, R.T. (1989). A dual-ascent procedure for large-scale uncapacitated network design. *Operations Research*, 37:716 – 740.

Barnhart, C. and Schneur, R.R. (1996). Air network design for express shipment service. *Operations Research*, 44:852 – 863.

Bartholdi, J. and Dave, D. (2002). Caliber technology: Design of a less-then-truckload network. *Leaders in Logistics Report*, The Logistics Institute, Georgia Institute of Technology, Atlanta, Georgia.

Bartholdi, J., Dave, D., and Lee, E. (2003). Routing freight on very large less-then-truckload networks. Unpublished report.

Bartholdi, J. and Gue, K. (2000). Reducing labor costs in an LTL crossdocking terminal. *Operations Research*, 48:823 – 832.

Bartholdi, J. and Gue, K. (2004). The best shape for a crossdock. Forthcoming in *Transportation Science*.

Bell, P., Anderson, C., and Kaiser, S. (2003). Strategic operations research and the Edelman Prize finalist applications. *Operations Research*, 51:7 – 31.

Braklow, J., Graham, W., Hassler, S., Peck, K., and Powell, W.B. (1992). Interactive optimization improves service and performance for Yellow Freight Systems. *Interfaces*, 22(1):147 – 172.

Bontekoning, Y.M., Macharis, C., and Trip, J.J. (2004). Is a new applied transportation research field emerging? — A review of intermodal rail-truck freight transport literature. *Transportation Research A*, 38:1 – 34.

Campbell, J.F. (1996). Hub location and the *p*-hub median problem. *Operations Research*, 44:923 – 935.

Campbell, J.F., Ernst, A.T., and Krishnamoorthy, M. (2002). Hub location problems. In: Z. Drezner and H. Hamacher (eds.), *Facility Location: Applications and Theory*, pp. 373 – 407. Springer-Verlag, Berlin, Germany.

Cheung, W., Leung, L., and Wong, Y.M. (2001). Strategic service network design for DHL Hong Kong. *Interfaces*, 31(4):1 – 14.

Crainic, T.G. (2003). Long haul freight transportation. In: R. Hall (ed.), *Handbook of Transportation Science*, 2nd edition, pp. 451 – 516. Kluwer Academic Publishers, Boston, Massachusetts.

Crainic, T. G. and Roy, J. (1988). O.R. tools for tactical freight transportation planning. *European Journal of Operational Research*, 33:290 – 297.

Crainic, T. G. and Roy, J. (1992). Design of regular intercity driver routes for the LTL motor carrier industry. *Transportation Science*, 26:280 – 295.

Daganzo, C.F. (1987). The break-bulk role of terminals in many-to-many logistic networks. *Operations Research*, 35:543 – 555.

Daganzo, C.F. (1999). *Logistics Systems Analysis*, 3rd edition. Springer-Verlag, Berlin.

Daskin, M. and Owen, S. (2003). Location models in transportation. In: R. Hall (ed.), *Handbook of Transportation Science*, 2nd edition, pp. 321 – 372. Kluwer Academic Publishers, Boston, Massachusetts.

Daskin, M., Snyder, L.V., and Berger, R.T. (2005). Facility location in supply chain design. In: A. Langevin and D. Riopel (eds.), *Logistics Systems: Design and Optimization*, Kluwer Academic Publishers, Boston, Massachusetts. Forthcoming.

Delorme, L., Roy. J., and Rousseau, J-M. (1988). Motor carrier operations planning models: A state of the art. In: L. Bianco and A.L. Bella (eds.), *Freight Transport*

Planning and Logistics, pp. 510 – 545. Springer-Verlag, Berlin.

Donaldson, H., Johnson, E., Ratliff, H.D., and Zhang, M. (1999). Schedule driven cross-docking networks. *The Logistics Institute Research Report.* Georgia Institute of Technology, Atlanta, Georgia.

Drezner, Z., and Hamacher, H. (eds.) (2002). *Facility Location: Applications and Theory*, Springer-Verlag, Berlin.

Equi, L., Gallo, G., Marziale, S., and Weintraub, A. (1997). A combined transportation and scheduling problem. *European Journal of Operational Research*, 97:94 – 104.

Ernst, A.T. and Krishnamoorthy, M. (1996). Efficient algorithms for the uncapacitated single allocation p-hub median problem. *Location Science*, 4:139 – 154.

Ernst, A.T. and Krishnamoorthy, M. (1999). Solution algorithms for the capacitated single allocation hub location problem. *Annals of Operations Research*, 86:141 – 159.

European Commission. (2003). Panorama of transport: Statistical overview of transport in the European Union. Office for Official Publications of the European Communities, Luxembourg.

Farvolden, J. and Powell, W.B. (1994). Subgradient methods for the service network design problem. *Transportation Science*, 28:256 – 272.

Farvolden, J. M., Powell, W.B., and Lustig, I.J. (1993). A primal partitioning solution for the arc-chain formulation of a multicommodity network flow problem. *Operations Research*, 41:669 – 693.

Fleischmann, B. (1998). Design of freight traffic networks. In: B. Fleischmann, J.A.E.E. Nunen, M.G. Speranza, and P. Stähly (eds.), *Advances in Distribution Logistics*, *Lecture Notes in Economics and Mathematical Systems*, 460:55 – 81, Springer-Verlag, Berlin.

Griffin G., Kalnbach, L., Lantz, B., and Rodriguez, J. (2000). Driver retention strategy: The role of a career path. Upper Great Plains Transportation Institute, North Dakota State University, Fargo, North Dakota.

Grunert, T. and Sebastian, H.-J. (2000). Planning models for long-haul operations of postal and express shipment companies. *European Journal of Operational Research*, 122:289 – 309.

Grunert, T., Sebastian, H-J., and Tharigen, M. (1999). The design of a letter-mail transportation network by intelligent techniques. In: R. Sprague, (ed.), *Proceedings of the 32nd Hawaii International Conference on System Sciences*, January 1999, Computer Society Press.

Gue, K. (1999). The effects of trailer scheduling on the layout of freight terminals. *Transportation Science*, 33:419 – 428.

Hall, R.W. (2003). Supply chains. In: R.W. Hall (ed.), *Handbook of Transportation Science*, 2nd edition. Kluwer Academic Publishers, Boston, Massachusetts.

Haresamudra, B., Taylor, G.D., and Taha, H. (1995). An interactive approach to locate terminals for LTL trucking operations. *MBTC Final Report 1037*, Mack Blackwell National Rural Transportation Center, University of Arkansas, Fayetteville, Arkansas.

Hoppe, B., Klampfl, E.Z., McZeal, C., and Rich, J. (1999). Strategic load planning for less-than-truckload trucking. *Center for Research on Parallel Computation Working Paper CRPC-TR99812-S*, Rice University, Houston, Texas.

Hunt, G. (1998). *Transportation Relay Network Design.* Ph.D. Thesis, Georgia Institute of Technology, Atlanta, Georgia.

Kallman, E.A. and Gupta R.C. (1979). Top management commitment to strategic planning: An empirical study. *Managerial Planning*, May/June:34 – 38.

Kara, B. and Tan, P. (2003). The latest arrival hub location problem. Powerpoint Presentation, Bilkent University, Ankara, Turkey.

Kara, B. and Tansel, B. (2001). The latest arrival hub location problem. *Management Science*, 47:1408 – 1420.

Kim, D. and Barnhart, C. (1997). Transportation service network design: Models and algorithms. Submitted to 7th International Workshop on Computer-Aided Scheduling of Public Transport, MIT, Cambridge, Massachusetts.

Kim, D., Barnhart, C., Ware, K., and Reinhardt, G. (1999). Multimodal express package delivery: A service network design application. *Transportation Science*, 33:391 – 407.

Lamar, B.W. and Sheffi, Y. (1987). An implicit enumeration method for LTL network design. Transportation Research Record, 1120:1 – 11.

Lamar, B., Sheffi, Y., and Powell, W.B. (1990). A capacity improvement lower bound for fixed charge network design problems. *Operations Research*, 38:704 – 710.

Langevin, A., Mbaraga, P., and Campbell, J.F. (1996). Continuous approximation models in freight distribution: An overview. *Transportation Research B*, 30:163 – 188.

Laporte, G. (1989). A survey of algorithms for location-routing problems. *Investigacion Operativa*, 1:93 – 123.

Leung, J.M.Y., Magnanti, T., and Singhal, V. (1990). Routing in point-to-point delivery systems: Formulations and solution heuristics. *Transportation Science*, 24:245 – 260.

Lin, C-C. (2001). The freight routing problem of time-definite freight delivery common carriers. *Transportation Research B*, 35:525 – 547.

Meinert, T.S. and Taylor, G.D. (1999). Summary of route regularization alternatives: A historical perspective. In: G.D. Taylor, E.M. Malstrom, J.A. Watson and K.G. Stanley (eds.), *Proceedings of the 1999 Industrial Engineering Research Conference*, Phoenix, Arizona.

Nagy, G. and Salhi, S. (1998). The many-to-many location routing problem. *Sociedad de Estadistica e investigacion Operativa Top*, 6:261 – 275.

O'Kelly, M.E. and Lao, Y. (1991). Mode choice in a hub-and-spoke network: A zero-one linear programming approach. *Geographical Analysis*, 23:283 – 297.

Powell, W.B. (1986). A local improvement heuristic for the design of less-than-truckload motor carrier networks. *Transportation Science*, 20:246 – 257.

Powell,W.B. and Sheffi, Y. (1983). The load planning problem of motor carriers: Problem description and a proposed solution approach. *Transportation Research A*, 17:471 – 480.

Powell, W.B. and Sheffi, Y. (1986). Interactive optimization for motor carrier load planning. *Journal of Business Logistics*, 7(2):64 – 90.

Powell, W.B. and Sheffi, Y. (1989). Design and implementation of an interactive optimization system for the network design in the motor carrier industry. *Operations Research*, 37:12 – 29.

Powell, W.B., Sheffi, Y., Nickerson, K.S., Butterbaugh, K., and Atherton, S. (1988). Maximizing profits for North American Van Lines' truckload division: A new framework for pricing and operations. *Interfaces*, 18(1):21 – 41.

Road Haulage and Distribution Training Council. (2003). Driver retention: Research with managers and LGV drivers in Scotland. A Report for the Scottish Road Haulage Modernisation Fund.

Roy, J. (2001). Recent trends in logistics and the need for real-time decision tools in the trucking industry. *CRG working paper 10-2001*, Centre de Recherche en

Gestion, UQAM, Montreal, Quebec, Canada.

Roy, J., and Crainic, T.G. (1992). Improving intercity freight routing with a tactical planning model. *Interfaces*, 22(3):31 – 44.

Roy, J. and Delorme, L. (1989). NETPLAN: A network optimization model for tactical planning in the less-than-truckload motor carrier industry. *INFOR*, 27:22 – 35.

Salhi, S. and Sari, M. (1997). Models for the multi-depot vehicle fleet mix problem. *European Journal of Operational Research*, 103:95 – 112.

Schwarz, M. (1992). *J.B. Hunt: The Long Haul to Success.* University of Arkansas Press, Fayetteville, Arkansas.

Stumpf, P. (1998). Vehicle routing and scheduling for truck haulage. In: B. Fleischmann, J.A.E.E. Nunen, M.G. Speranza, and P. Stähly (eds.), *Advances in Distribution Logistics, Lecture Notes in Economics and Mathematical Systems*, 460:341 – 371, Springer-Verlag, Berlin.

Taha, T. and Taylor, G.D. (1994). An integrated modeling framework for evaluating hub-and-spoke networks in truckload trucking. *Logistics and Transportation Review*, 30(2):141 – 166.

Taha, T., Taylor, G.D., and Taha, H. (1996). A simulation-based software system for evaluating hub-and-spoke transportation networks. *Simulation Practice and Theory*, 3:327 – 346.

Taniguchi, E., Noritake, M., Yamada, T., and Izumitani, T. (1999). Optimal size and location planning of public logistics terminals. *Transportation Research E*, 35:207 – 222.

Tansel, B. and Kara, B. (2002). The latest arrival hub location problem for cargo delivery systems with feeders and stopovers. *IEOR-2002-06*, Dept. of Industrial Engineering, Bilkent Univeristy, Bilkent, Turkey.

Taylor, G.D. and Meinert, T.S. (2000). Improving the quality of operations in truckload trucking. *IIE Transactions*, 32:551 – 562.

Taylor, G.D., Harit, S., English, J., and Whicker, G. (1995). Hub and spoke networks in truckload trucking: Configuration, testing and operational concerns. *Logistics and Transportation Review*, 31:209 – 237.

Taylor, G.D., Meinert, T.S., Killian, R.C., and Whicker, G.L. (1999). Development and analysis of alternative dispatching methods in truckload trucking. *Transportation Research E*, 35:191 – 205.

Taylor, G.D., Whicker, G.L., and Usher, J.S. (2001). Multi-zone dispatching in truckload trucking. *Transportation Research E*, 37:375 – 390.

United States Bureau of Transportation Statistics. (2003). U.S. International Trade and Freight Transportation Trends. BTS03-02, Washington, D.C., February.

Wleck, H. (1998). Local search heuristics for the design of freight carrier networks. In: B. Fleischmann, J.A.E.E. Nunen, M.G. Speranza, and P. Stähly (eds.), *Advances in Distribution Logistics*, pp. 265 – 285. Volume 460 of *Lecture Notes in Economics and Mathematical Systems.* Springer-Verlag, Berlin.

Chapter 9

NEW HEURISTICS FOR THE VEHICLE ROUTING PROBLEM

Jean-François Cordeau
Michel Gendreau
Alain Hertz
Gilbert Laporte
Jean-Sylvain Sormany

Abstract This chapter reviews some of the best metaheuristics proposed in recent years for the Vehicle Routing Problem. These are based on local search, on population search and on learning mechanisms. Comparative computational results are provided on a set of 34 benchmark instances.

1. Introduction

The classical *Vehicle Routing Problem* (VRP) is defined on an undirected graph $G = (V, E)$ where $V = \{v_0, v_1, \dots, v_n\}$ is a vertex set and $E = \{(v_i, v_j) : v_i, v_j \in V, i < j\}$ is an edge set. Vertex v_0 is a depot at which are based m identical vehicles of capacity Q, while the remaining vertices represent customers. A non-negative *cost*, *distance* or *travel time* matrix $C = (c_{ij})$ is defined on E. Each customer has a non-negative demand q_i and a non-negative service time s_i. The VRP consists of designing a set of m vehicle routes (i) of least total cost, (ii) each starting and ending at the depot, and such that (iii) each customer is visited exactly once by a vehicle, (iv) the total demand of any route does not exceed Q, and (v) the total duration of any route does not exceed a preset bound D.

The VRP is a hard combinatorial problem. Exact algorithms (see, e.g., Naddef and Rinaldi, 2002; Baldacci et al., 2004) can only solve relatively small instances and their computational times are highly variable. To this day, heuristics remain the only reliable approach for the solution of practical instances. In contrast to exact algorithms, heuristics are better suited to the solution of VRP variants involving side constraints such as time windows (Cordeau et al., 2002a), pickups and deliveries (Desaulniers et al., 2002), periodic visits (Cordeau et al., 1997), etc.

In recent years several powerful heuristics have been proposed for the VRP and its variants, based on local search, population search and learning mechanisms principles. Local search includes descent algorithms (Ergun et al., 2003), simulated annealing (Osman, 1993), deterministic annealing (Golden et al., 1998; Li et al., 2005), tabu search (Osman, 1993; Taillard, 1993; Gendreau et al., 1994; Xu and Kelly, 1996; Rego and Roucairol, 1996; Rego, 1998; Barbarosoğlu and Ögür, 1999; Cordeau et al., 2001). The two best known types of population search heuristics are evolutionary algorithms (Prins, 2004; Berger and Barkaoui, 2004; Mester and Bräysy, 2005) and adaptive memory procedures (Rochat and Taillard, 1995; Tarantilis and Kiranoudis, 2002). Examples of learning mechanisms are neural networks (Ghaziri, 1991, 1996; Matsuyama, 1991; Schumann and Retzko, 1995) and ant algorithms Reimann et al. (2004).

The field of VRP heuristics is very active, as witnessed by the large number of recent articles listed in the previous paragraph. This chapter summarizes some of the most important new developments in the area of VRP heuristics and presents comparative computational results.

Several surveys have recently been published on VRP heuristics (Laporte and Semet, 2002; Gendreau et al., 2002; Cordeau et al., 2002a; Cordeau and Laporte, 2004). This chapter focuses on recent material not covered by these surveys. In the following section we provide a general classification scheme for VRP heuristics. We then provide in Section 3 a description of nine recent heuristics, and computational results in Section 4. The conclusion follows.

2. Classification of VRP heuristics

Providing classification schemes in the area of combinatorial optimization can be a daunting task because of the large number of fields and descriptors needed to account for the diversity and intricacy of the concepts involved in the various algorithms — the devil is in the details. By and large broad classification systems that concentrate on the essential ideas can be quite instructive.

At a macro-level, VRP heuristics combine some of the following four components: (1) *construction* of an initial solution; (2) *improvement* procedures; (3) *population* mechanisms; and (4) *learning* mechanisms.

2.1 Constructive heuristics

The ideas behind most constructive heuristics are well known and well documented (Laporte and Semet, 2002). These include the Clarke and Wright (1964) savings concept, the sweep mechanism (Gillett and Miller, 1974), and cluster-first route-second methods (Fisher and Jaikumar, 1981), and route-first cluster-second methods (Beasley, 1983).

2.2 Improvement heuristics

Most constructive procedures are followed by an improvement phase. In the simplest case, a post-optimization procedure designed for the *Traveling Salesman Problem* (TSP) is applied to individual routes: r-opt exchanges (Lin, 1965), Or-opt exchanges (Or, 1976), 2-opt* exchanges (Potvin and Rousseau, 1995), 4-opt* exchanges (Renaud et al., 1996), and the more involved unstringing and stringing (US) mechanism (Gendreau et al., 1992). Exchanges often involve two vehicle routes, such as chain exchanges (Fahrion and Wrede, 1990) and the λ-interchange mechanisms (Osman, 1993), the string cross, string exchange and string relocation schemes of van Breedam (1994). Finally, more complicated operations involve several routes: cyclic exchanges (Thompson and Psaraftis, 1993), edge exchange schemes (Kindervater and Savelsbergh, 1997), ejection chains (Xu and Kelly, 1996; Rego and Roucairol, 1996; Rego, 1998), and very large neighbourhoods in which a sequence of moves is determined through the solution of an auxiliary network flow optimization problem (Ergun et al., 2003).

Most classical improvement mechanisms work in a descent mode until a local optimum is reached. In metaheuristics (e.g., simulated annealing, deterministic annealing, tabu search) the same mechanisms are embedded within sophisticated neighbourhood search structures which allow for intermediate deteriorating solutions and even infeasible solutions (e.g., Gendreau et al., 1994). In variable neighbourhood search (VNS), introduced by Mladenović and Hansen (1997), the neighbourhood structure is allowed to vary during the search; this concept can be coupled with descent methods or with tabu search, for example. Figure 9.1 depicts a number of ways to design heuristics consisting of a construction phase followed by an improvement phase.

2.3 Population mechanisms

Combination of solutions is the basic mechanism of population search which includes a large number of variants known as genetic algorithms (e.g., Reeves, 2003), and memetic algorithms (e.g., Moscato and Cotta,

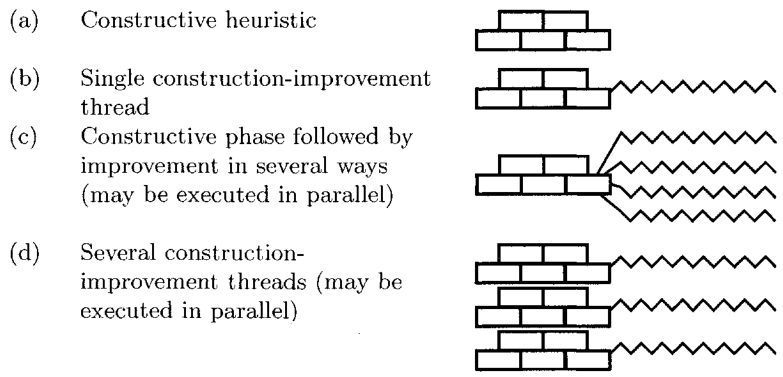

(a) Constructive heuristic
(b) Single construction-improvement thread
(c) Constructive phase followed by improvement in several ways (may be executed in parallel)
(d) Several construction-improvement threads (may be executed in parallel)

Figure 9.1. Graphical representation of several construction and improvement heuristics (▭: construction; 〰: improvement)

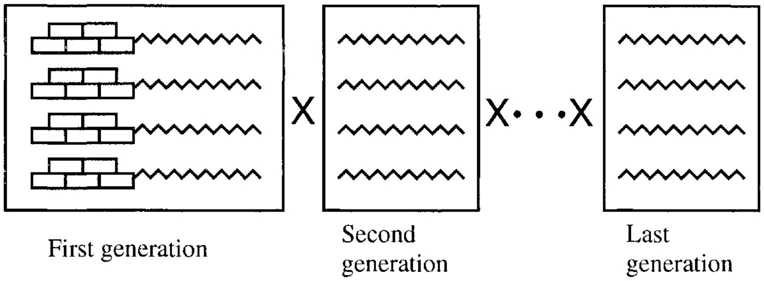
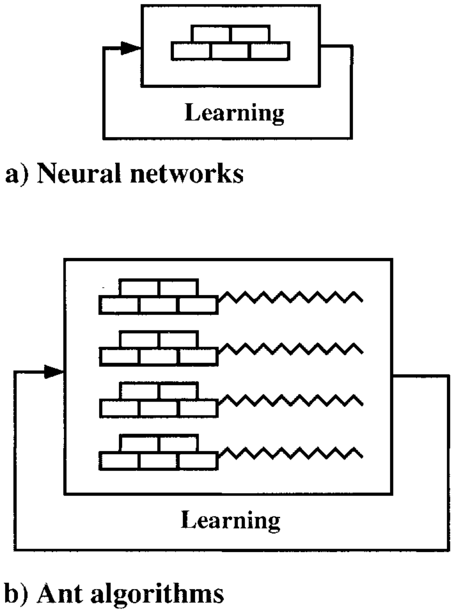


Figure 9.2. Depiction of a population algorithm obtained by combining (X) some elements of a generation to obtain the next generation

2003). Classical genetic algorithms operate on a population of encoded solutions called chromosomes. At each iteration (generation) the following operations are applied k times: select two parent chromosomes; generate two offspring from these parents using a crossover operator; apply a random mutation to each offspring with a small probability; remove the $2k$ worst elements of the population and replace them with the $2k$ offspring. Several ways of performing crossovers have been proposed for sequencing problems (e.g., Potvin, 1996; Bean, 1994; Drezner, 2003).

The idea of combining solutions to generate new ones is central to the adaptive memory procedure put forward by Rochat and Taillard (1995) for the solution of the VRP. These authors extract vehicle routes from several good solutions and use them as a basis for the construction of offspring. A variant, proposed by Tarantilis and Kiranoudis (2002), initiates offspring from chains of vertices extracted from parent solutions. Figure 9.2 depicts a population mechanism.



a) Neural networks

b) Ant algorithms

Figure 9.3. Depiction of two learning mechanisms

2.4 Learning mechanisms

Two main learning mechanisms have been used for the design of VRP heuristics. Neural networks operate on a set of deformable templates which are essentially rings that are candidates to become feasible vehicle routes. Rings compete for vertices through a random mechanism in which the probability of assigning a vertex to a ring evolves through a learning process. It is fair to say that neural networks cannot yet compete with most other VRP heuristics. Ant algorithms are also derived from a learning paradigm. They are derived from an analogy with ants which lay pheromone on their trail as they forage for food. With time paths leading to the best food sources are more frequented and are marked with a larger amount of pheromone. In construction or improvement heuristics for the VRP elementary moves leading to better solution can be assigned a higher probability of being selected. An algorithm based on such a learning feedback mechanism will be outlined in Section 3. Figure 9.3 depicts two learning mechanisms. The learning feedback loop enables the process to restart with different rules or parameter settings.

3. Some recent VRP heuristics

We now summarize nine recent VRP heuristics. The first four are based on local search, the next four are population based, while the last one is an ant algorithm.

3.1 The Toth and Vigo granular tabu search algorithm

The granular tabu search (GTS) algorithm put forward by Toth and Vigo (2003) *a priori* removes from the graph edges that are unlikely to appear in an optimal VRP solution, with the aim of curtailing computation time. The idea was first implemented in conjunction with a tabu search method but the principle is general and could be beneficial to other type of algorithms. Specifically, Toth and Vigo recommend retaining only the edges incident to the depot and all edges whose length does not exceed a given *granularity threshold* $\nu = \beta\bar{c}$, where $\bar{c}$ is the average edge cost in a good feasible solution obtained by a fast heuristic, and β is a sparsification parameter typically chosen in the interval $[1.0, 2.0]$. With this choice of β, the percentage of edges remaining in the reduced graph tends to lie between 10% and 20% of the original number. In practice β is dynamically decreased to provide an intensification effect, or increased to diversify the search. Toth and Vigo have implemented GTS in conjunction with some features included in the tabu search algorithms of Taillard (1993) and of Gendreau et al. (1994). Neighbour solutions were obtained by performing intra-route and inter-route edge exchanges.

3.2 The Li, Golden and Wasil heuristic

The search heuristic developed by Li et al. (2005) combines the record-to-record (RTR) principle first put forward by Dueck (1993) with a variable-length neighbour list whose principle is similar to GTS (Toth and Vigo, 2003). The algorithm is called VRTR for variable-length neighbourhood list record-to-record travel. Only a proportion α of the 40 shortest edges incident to each vertex are retained. The value of α varies throughout the algorithm.

The RTR search is applied different times from three initial solutions generated with the Clarke and Wright (1964) algorithm with savings s_{ij} defined as $c_{i0} + c_{0j} - \lambda c_{ij}$, where $\lambda = 0.6$, 1.4 and 1.6. Neighbour solutions are obtained by means of intra-route and inter-route 2-opt moves. During the search, deteriorating solutions are accepted as long as their solution value does not exceed 1.01 times that of the best known solution. When the value of the incumbent has not improved for a number of iterations a final attempt is made to improve the best known solution by means of a perturbation technique. This is done by reinserting some of its vertices in different positions and restarting the search process.

3.3 The unified tabu search algorithm

The unified tabu search algorithm (UTSA) was originally put forward by Cordeau et al. (1997) as a unified tool to solve periodic and multi-depot VRPs. It has been extended to the site dependent VRP (Cordeau et al., 2001), and to the time-windows version of these problems (Cordeau et al., 2001, 2004). It possesses some of the features of Taburoute (Gendreau et al., 1994), namely the consideration of intermediate infeasible solutions through the use of a generalized objective function containing self-adjusting coefficients, and the use of continuous diversification. Neighbour solutions are obtained by moving a vertex from its route between two of its closest neighbours in another route, by means of a generalized insertion (or GENI) (see Gendreau et al., 1992). Contrary to Taburoute, UTSA uses only one initial solution and fixed tabu durations. The tabu mechanism works with an attribute set $B(x)$ associated with solution x, defined as $B(x) = \{(i, k)$: vertex v_i is visited by vehicle k in solution $x\}$. The neighbourhood mechanism removes an attribute (i, k) from $B(x)$ and replaces it with (i, k'), where $k' \neq k$; attribute (i, k) is then declared tabu. Recently a new diversification phase was introduced into UTSA. Whenever the value of the best known solution has not improved for a number of iterations, the depot is moved to the first vertex of a randomly selected route and temporarily remains in this location. This computational device is a form of data perturbation, a principle put forward by Codenetti et al. (1996). On benchmark test problems the implementation of this simple device has helped reduce the average deviation from the best known solution values from 0.69% to 0.56% without any increase in computing time (see Table 9.1).

3.4 Very large neighbourhood search

Very large neighbourhood search (VLNS) attempts, at every iteration, to identify an improving solution by exploring a neighbourhood whose size is very large with respect to the input data. It was applied to the VRP by Ergun et al. (2003). The heuristic developed by these authors is a descent mechanism that operates on several routes at once, not unlike cyclic transfers (Thompson and Psaraftis, 1993) and ejection chains (Rego and Roucairol, 1996) which act on a set of h routes $r_1, \dots, r_h$ by moving vertices from route r_ℓ to route $r_{\ell+1 (\text{mod } h)}$, $\ell = 1, \dots, h$. Neighbour solutions are defined by means of 2-opt moves, vertex swaps between routes, and vertex insertions in different routes. In order to determine the best sequence of moves at a given iteration, a shortest path problem is solved on an auxiliary graph, called improvement graph. The main advantage of this type of approach is that it allows a broad search to be performed by acting on several moves at once. Its disadvantage lies in the computational effort required at each iteration to determine the best compounded move.

3.5 The evolutionary algorithm of Prins

The heuristic put forward by Prins (2004) combines the two main features of evolutionary search: crossover and mutation operations. Improvements are obtained by means of a local search procedure applied to a candidate solution. Hence this algorithm is best described as a memetic algorithm (Moscato and Cotta, 2003). The moves include vertex and edge reinsertions, vertex swaps, combined vertex and edge swaps, and edge swaps. The search procedure ends at the first improving move. This algorithm operates on solutions represented as ordered sequences of customers: all vertices except the depot first appear on a cycle, without trip delimiters, as if a single vehicle traveled all routes in succession, in a cyclic manner, without going through the depot. An optimal partition of this cycle is then determined by solving a shortest path problem on an auxiliary graph, like what is done in route-first cluster-second algorithms (Beasley, 1983). This procedure is also applied after each mutation. Crossovers are performed as follows. Two cutting locations i and j are determined in parent # 1 and the corresponding string is placed in positions $i, \ldots, j$ of offspring # 1, which is then completed by sweeping parent # 2 circularly from position $j + 1$. A second offspring is created by reversing the roles of the two parents.

3.6 The Bone Route adaptive memory algorithm of Tarantilis and Kiranoudis

Tarantilis and Kiranoudis (2002) have developed a rather effective adaptive memory procedure for the VRP. In a first phase a solution is obtained by means of the Paessens (1988) constructive procedure, which is an enhancement of the Clarke and Wright (1964) algorithm, followed by a tabu search procedure in which neighbours are defined by 2-opt moves, vertex swaps between routes, and vertex reinsertions in the same route or in a different route. The adaptive memory procedure does not initiate new solutions by combining full vehicle routes, as did Rochat and

Taillard (1995) but route segments, called *bones*, extracted from good quality routes.

3.7 The AGES algorithm of Mester and Bräysy

The active guided evolution strategies (AGES) algorithm of Mester and Bräysy (2005) was initially applied to the VRP with time windows but results have recently been obtained for the classical VRP (Mester, 2004). AGES combines guided local search (Voudouris, 1997) with evolution strategies (Rechenberg, 1973) into an iterative two-stage procedure. Contrary to standard evolutionary search heuristics, AGES uses a deterministic rule for parent selection and the search is driven by a high mutation rate. AGES does not recombine parents, but it creates a single offspring from a single parent through a mutation procedure, and the offspring replaces the parent if it has a better fitness value. Guided local search operates like simple memory based metaheuristics such as simulated annealing and tabu search. It penalizes some solution features that are unlikely to appear in an optimal solution (like long edges) and also uses a frequency weight. This search mechanism therefore combines the basic principle of granular tabu search with continuous diversification. Neighbour solutions are defined by vertex swaps and interchanges, and by 2-opt moves (Potvin and Rousseau, 1995). The search procedure uses very large neighbourhoods (Shaw, 1998). A restart mechanism applied to the best solution encountered is reported to play a significant role in reaching high quality solutions.

3.8 The hybrid genetic algorithm of Berger and Barkaoui

The hybrid genetic algorithm of Berger and Barkaoui (2004) combines evolutionary search and local search. It is best described as a memetic algorithm. Its originality lies in the use of two populations. New offspring are created in each population whose size is kept constant by replacing the worst elements by the best ones. A migration operation is then applied by swapping the best elements of each population. Crossovers are performed by creating one offspring from two parents as follows: a number of routes are extracted from parent # 1, yielding a partial solution which is then completed by inserting in some of the routes vertices of parent # 2 selected according to a proximity criterion (their closeness to the centroid of a route), or by creating new routes. The insertion mechanism I1 (Solomon, 1987) is modified to include a random choice of the cost function parameters. Solutions are improved by performing a large scale neighbourhood search (Shaw, 1998) that combines three insertion mechanisms, and by applying a route improvement scheme. The first insertion mechanism is based on I1: vertices are first ranked according to a function that combines their best reinsertion cost and the number of feasible insertions, and the highest ranked vertex is then reinserted according to a cheapest insertion cost criterion. The second mechanism uses a reject function as in Liu and Shen (1999) that compares for a given vertex the cost of an insertion opportunity to the best insertion cost achievable in the neighbourhood of that vertex. The third mechanism operates on two routes at a time by performing moves or swaps involving up to two vertices (as in the Osman, 1993, λ-interchange mechanism), and by implementing the first improving move. Finally, an attempt to improve each route is made by removing in turn each vertex and reinserting it by means of I1.

3.9 The *D*-ants savings based heuristic of Reimann, Doerner and Hartl

The D-ants heuristic of Reimann et al. (2004) repeatedly applies two phases until a stopping criterion is reached. The first phase iterates between a savings based procedure for generation of a pool of good solutions and an improvement mechanism applied to each of these solutions. A learning mechanism guides the creation of each new generation. Solutions are generated by means of a savings algorithm. Instead of using the classical Clarke and Wright (1964) saving $s_{ij} = c_{i0} + c_{0j} - c_{ij}$, the authors use an attractiveness value χ_{ij} equal to $\tau_{ij}^{\alpha} s_{ij}^{\beta}$, where τ_{ij}^{α} contains information on how good combining i and j turned out to be in previous iterations, and α, β are user-controlled weights. The combination of vertices v_i and v_j occurs with probability p_{ij} defined as $\chi_{ij}/(\sum_{(h,\ell)\in\Omega_k} \chi_{h\ell})$, where Ω_k is the set of the feasible (i, j) combinations yielding the k best savings. The authors use 2-opt in the improvement phase. In the second phase, the best solution identified in the first phase is decomposed into several subproblems, each of which is optimized by means of the procedure used in the first phase.

4. Comparative computational results

We now present computational results for the various heuristics described in Section 3. Statistics for the 14 Christofides et al. (1979) instances ($50 \leq n \leq 199$) and for the 20 Golden et al. (1998) instances ($200 \leq n \leq 480$) are reported in Tables 9.1 and 9.2, respectively. Unless otherwise indicated, solution values correspond to a single run with a given parameter setting. Best solution values are in boldface.

The column headings are as follows:

n: number of customers;
Value: best solution value produced by the heuristic;
%: percentage deviation from the best known value;
Minutes: computation time in minutes;
Best : best known solution value.

Our first observation relates to the accuracy level reached by these algorithms. On the CMT instances, the average percentage deviation form the best known solution value always lies between 0.03 and 0.64. The worst performers (GTS and UTSA) are two tabu search algorithms, while the best algorithms combine population search and local search (e.g., AGES, Bone Route and the Prins algorithm). This observation is consistent with the results obtained with the first generation of tabu search heuristics (Cordeau and Laporte, 2004) for which the average deviation was typically higher. Results obtained for the larger instances (Table 9.2) point in the same direction but they must be interpreted with more care because these instances have not been as extensively studied as the first ones.

Computation times are provided for information but, as usual these are hard to interpret because of the different computers employed. The Dongarra (2004) study which is frequently updated throws some light on relative computer speeds. Irrespective of this, it appears that the AGES heuristic of Mester and Bräysy runs rather fast and comes out as the overall winner when both accuracy and computing time are taken into account.

Two additional performance criteria stated by Cordeau et al. (2002b) are simplicity and flexibility. Simplicity relates to ease of understanding and coding of an algorithm. According to this criterion the Li et al. (2005) heuristic is probably the best: it is based on a rather simple mechanism and requires a relatively small amount of coding. The Prins algorithm, UTSA and GTS also possess simple structures which should make them easier to reproduce. At the other extreme, VLNS and AGES are probably the most complicated of all algorithms used in the comparisons. Flexibility measures the capacity of adapting an algorithm to effectively deal with additional constraints. There exists abundant documented evidence that UTSA can be applied to a host of VRP extensions (Cordeau et al., 2001, 2004). Also, some generic principles like GST and VLNS apply to several contexts and should score high on the flexibility criterion, although they have not to our knowledge yet been applied to VRP extensions such as the VRP with time windows (VRPTW). On the other hand, Prins and Mester and Bräysy report results on the VRPTW.

Table 9.1. Computational results for the Christofides et al. (1979) instances

			GTS Toth and Vigo (2003)			Li et al. (2005)		USTA Cordeau et al. (2001)			VLNS Ergun et al. (2003)			Prins (2004)		
Instance	n	Type[a]	Value	%	Minutes[b]	Value[c]	%	Value[d]	%	Minutes[e]	Value[f]	%	Minutes[g]	Value	%	Minutes[h]
1	50	C	**524.61**	0.00	0.81	**524.61**	0.00	**524.61**	0.00	2.32	**524.61**	0.00	23.13	**524.61**	0.00	0.01
2	75	C	838.60	0.40	2.21	836.18	0.11	835.28	0.00	14.78	835.43	0.02	33.93	**835.26**	0.00	0.77
3	100	C	828.56	0.29	2.39	827.39	0.15	**826.14**	0.00	11.67	827.46	0.16	21.30	**826.14**	0.00	0.46
4	150	C	1033.21	0.47	4.51	1045.36	1.65	1032.68	0.41	26.66	1036.24	0.76	24.45	1031.63	0.31	5.50
5	199	C	1318.25	2.09	7.50	1303.47	0.94	1315.76	1.90	57.68	1307.33	1.24	57.25	1300.23	0.69	19.10
6	50	C, D	**555.43**	0.00	0.86			**555.43**	0.00	3.03	**555.43**	0.00	3.50	**555.43**	0.00	0.01
7	75	C, D	920.72	1.21	2.75			**909.68**	0.00	7.41	910.04	0.04	36.53	912.30	0.29	1.42
8	100	C, D	869.48	0.41	2.90			865.95	0.00	10.93	**865.94**	0.00	12.43	**865.94**	0.00	0.37
9	150	C, D	1173.12	0.91	5.67			1167.85	0.46	51.66	1164.88	0.20	42.47	1164.25	0.15	7.25
10	199	C, D	1435.74	2.86	9.11			1416.84	1.50	106.28	1404.36	0.61	28.32	1420.20	1.74	26.83
11	120	C	1042.87	0.07	3.18	**1042.11**	0.00	1073.47	3.01	11.67	**1042.11**	0.00	69.13	**1042.11**	0.00	0.30
12	100	C	**819.56**	0.00	1.10	**819.56**	0.00	**819.56**	0.00	9.02	**819.56**	0.00	5.98	**819.56**	0.00	0.05
13	120	C, D	1545.51	0.28	9.34			1549.25	0.53	21.00	1544.99	0.25	39.73	1542.97	0.12	10.44
14	100	C, D	**866.37**	0.00	1.41			**866.37**	0.00	10.53	**866.37**	0.00	6.55	**866.37**	0.00	0.09
Average				0.64	3.84		0.41		0.56	24.62		0.23	28.91		0.24	5.19

[a] C: Capacity restrictions; D: Route length restrictions.

[b] Pentium (200 MHz).

[c] Best variant ($\alpha = 0.4$).

[d] Results of recent computational experiments (see Section 3.3); the average % deviation in Cordeau et al. (2001) is 0.69.

[e] Pentium IV (2GHz).

[f] Best of five runs.

[g] Time for reaching the best value for the first time (Pentium III, 733 MHz).

[h] GHz PC (75 MFlops).

Table 9.1 (continued).

			Bone Route (Tarantilis and Kiranoudis, 2002)			AGES best (Mester and Bräysy, 2005)			AGES fast (Mester and Bräysy, 2005)			Berger and Barkaoui (2004)			
Instance	n	Type[a]	Value	%	Minutes[i]	Value[j]	%	Minutes[k]	Value[j]	%	Minutes[k]	Value	%	Minutes[l]	Best
1	50	C	**524.61**	0.00	0.11	**524.61**	0.00	0.01	**524.61**	0.00	0.01	**524.61**	0.00	2.00	**524.61**
2	75	C	**835.26**	0.00	4.56	**835.26**	0.00	0.26	**835.26**	0.00	0.26	**835.26**	0.00	14.33	**835.26**
3	100	C	**826.14**	0.00	7.66	**826.14**	0.00	0.05	**826.14**	0.00	0.05	827.39	0.15	27.90	**826.14**
4	150	C	1030.88	0.24	9.13	**1028.42**	0.00	0.47	**1028.42**	0.00	0.47	1036.16	0.75	48.98	**1028.42**
5	199	C	1314.11	1.77	16.97	**1291.29**	0.00	101.93	1294.25	0.23	0.50	1324.06	2.54	55.41	**1291.29**
6	50	C, D	**555.43**	0.00	0.10	**555.43**	0.00	0.02	**555.43**	0.00	0.02	**555.43**	0.00	2.33	**555.43**
7	75	C, D	**909.68**	0.00	0.92	**909.68**	0.00	0.43	**909.68**	0.00	0.43	**909.68**	0.00	10.50	**909.68**
8	100	C, D	**865.94**	0.00	4.28	**865.94**	0.00	0.44	**865.94**	0.00	0.44	868.32	0.27	5.05	**865.94**
9	150	C, D	1163.19	0.06	5.83	**1162.55**	0.00	1.22	1164.54	0.17	0.50	1169.15	0.57	17.88	**1162.55**
10	199	C, D	1408.82	0.93	14.32	1401.12	0.41	2.45	1404.67	0.42	0.45	1418.79	1.64	43.86	**1395.85**
11	120	C	**1042.11**	0.00	0.21	**1042.11**	0.00	0.05	**1042.11**	0.00	0.05	1043.11	0.10	22.43	**1042.11**
12	100	C	**819.56**	0.00	0.10	**819.56**	0.00	0.01	**819.56**	0.00	0.01	**819.56**	0.00	7.21	**819.56**
13	120	C, D	1544.01	0.19	8.75	**1541.14**	0.00	0.63	1543.26	0.14	0.47	1553.12	0.78	34.91	**1541.14**
14	100	C, D	**866.37**	0.00	0.10	**866.37**	0.00	0.08	**866.37**	0.00	0.08	**866.37**	0.00	4.73	**866.37**
Average				0.23	5.22		0.03	7.72		0.07	0.27		0.49	21.25	

[i] Pentium II (400 MHz).

[j] For C instances, see Mester and Bräysy (2005). Otherwise, see Mester (2004)

[k] Pentium IV (2 GHz).

[l] Pentium (400 MHz).

Table 9.2. Computational results for the Golden et al. (1998) instances

			GTS			Li et al. (2005)		USTA			VLNS			Prins (2004)		
			Toth and Vigo (2003)					Cordeau et al. (2001)			Ergun et al. (2003)					
Instance	n	Type[a]	Value	%	Minutes[b]	Value[c]	%	Value[d]	%	Minutes[e]	Value[f]	%	Minutes[g]	Value	%	Minutes[h]
1	240	C	5736.15	1.93	4.98	5666.42	0.69	5681.97	0.97	10.29	5741.79	2.03	134.95	5646.63	0.34	32.42
2	320	C	8553.03	1.24	8.28	8469.32	0.25	8657.36	2.48	35.39	8917.41	5.56	150.83	**8447.92**	0.00	77.92
3	400	C	11402.75	3.32	12.94	11145.80	0.99	11037.40	0.01	55.39	12106.64	9.70	15.67	**11036.22**	0.00	120.83
4	480	C	14910.62	9.44	15.13	13758.08	0.98	13740.60	0.85	83.19	15316.69	12.42	106.50	**13624.52**	0.00	187.60
5	200	C	6697.53	3.66	2.38	6478.09	0.26	6756.44	4.57	5.13	6570.28	1.69	15.50	**6460.98**	0.00	1.04
6	280	C	8963.32	6.54	4.65	8539.61	1.51	8537.17	1.48	18.64	8836.25	5.03	81.98	**8412.80**	0.00	9.97
7	360	C	10547.44	3.45	11.66	10289.72	0.92	10267.40	0.70	25.60	11116.68	9.03	85.00	10195.59	0.00	39.05
8	440	C	12036.24	3.20	11.08	11920.52	2.20	11869.50	1.77	71.44	12634.17	8.32	33.95	11828.78	1.42	88.30
9	255	C,D	593.35	1.71	11.67	588.25	0.83	587.39	0.69	37.26	587.89	0.77	49.20	591.54	1.40	14.32
10	323	C,D	751.66	1.30	15.83	749.49	1.01	752.76	1.45	51.11	749.85	1.05	125.05	751.41	1.26	36.58
11	399	C,D	936.04	1.92	33.12	925.91	0.81	929.07	1.16	41.54	932.74	1.56	171.05	933.04	1.59	78.50
12	483	C,D	1147.14	3.61	42.90	1128.03	1.88	1119.52	1.11	157.01	1134.63	2.48	388.62	1133.79	2.40	30.87
13	252	C,D	868.80	1.13	11.43	865.20	0.71	875.88	1.95	34.83	870.90	1.37	235.13	875.16	1.87	15.30
14	320	C,D	1096.18	1.38	14.51	1097.78	1.52	1102.03	1.92	21.56	1097.11	1.46	31.17	1086.24	0.46	34.07
15	396	C,D	1369.44	1.80	18.45	1361.41	1.20	1363.76	1.38	57.64	1367.15	1.63	65.30	1367.37	1.65	110.48
16	480	C,D	1652.32	1.83	23.07	1635.58	0.79	1647.06	1.50	129.50	1643.00	1.25	31.58	1650.94	1.74	130.97
17	240	C,D	711.07	0.46	14.29	711.74	0.56	710.93	0.44	18.03	716.46	1.22	223.62	710.42	0.37	5.86
18	300	C,D	1016.83	1.81	21.45	1010.32	1.16	1014.62	1.59	67.11	1023.32	2.46	299.23	1014.80	1.61	39.33
19	360	C,D	1400.96	2.49	30.06	1382.59	1.15	1383.79	1.24	66.21	1404.84	2.78	393.03	1376.49	0.70	74.25
20	420	C,D	1915.83	5.20	43.05	1850.92	1.63	1854.24	1.82	135.29	1883.33	3.41	121.62	1846.55	1.39	210.42
Average				2.87	17.55		1.05		1.45	56.11		3.76	137.95		0.91	66.90

[a] C: Capacity restrictions; D: Route length restrictions.

[b] Pentium (200 MHz).

[c] Best variant ($\alpha = 0.01$).

[d] Results of recent computational experiments (see Section 3.3).

[e] Pentium IV (2GHz).

[f] Best of two runs.

[g] Time for reaching the best value for the first time (Pentium III, 733 MHz).

[h] GHz PC (75 MFlops).

Table 9.2 (continued).

			Bone Route (Tarantilis and Kiranoudis, 2002)			AGES best (Mester and Bräysy, 2005)			AGES fast (Mester and Bräysy, 2005)			D-Ants (Reimann et al., 2004)			
Instance	n	Type[a]	Value	%	Minutes[i]	Value[j]	%	Minutes[k]	Value[j]	%	Minutes[k]	Value[l]	%	Minutes[m]	Best
1	240	C	5676.97	0.88	27.86	**5627.54**	0.00	8.73	5644.00	0.30	0.70	5644.02	0.29	62.52	**5627.54**
2	320	C	8512.64	0.77	55.62	**8447.92**	0.00	46.66	8468.00	0.24	0.20	8449.12	0.01	57.67	**8447.92**
3	400	C	11199.72	1.48	59.21	**11036.22**	0.00	40.55	11146.00	0.99	0.70	**11036.22**	0.00	21.92	**11036.22**
4	480	C	13637.53	0.10	47.63	**13624.52**	0.00	470.00	13704.52	0.59	2.50	13699.11	0.55	119.12	**13624.52**
5	200	C	**6460.98**	0.00	11.34	**6460.98**	0.00	0.17	6466.00	0.08	0.50	**6460.98**	0.00	0.87	**6460.98**
6	280	C	8429.28	0.20	12.54	**8412.88**	0.00	75.22	8539.61	1.51	0.10	8412.90	0.00	5.72	**8412.80**
7	360	C	10216.50	0.21	42.50	**10195.56**	0.00	2.55	10240.42	0.44	0.85	10195.59	0.00	14.03	**10195.56**
8	440	C	11936.16	2.34	79.69	**11663.55**	0.00	34.30	11918.75	2.19	0.27	11828.78	1.42	35.30	**11663.55**
9	255	C,D				**583.39**	0.00	8.33	588.25	0.83	0.80	586.87	0.60	21.52	**583.39**
10	323	C,D				**742.03**	0.00	6.00	752.92	1.39	0.43	750.77	1.25	17.48	**742.03**
11	399	C,D				**918.45**	0.00	110.00	925.94	0.82	1.10	927.27	0.96	96.88	**918.45**
12	483	C,D				**1107.19**	0.00	600.00	1128.67	1.94	1.50	1140.87	3.04	61.38	**1107.19**
13	252	C,D				**859.11**	0.00	10.25	865.20	0.71	0.18	865.07	0.69	87.20	**859.11**
14	320	C,D				**1081.31**	0.00	1.22	1097.68	1.51	0.28	1093.77	1.15	25.85	**1081.31**
15	396	C,D				**1345.23**	0.00	7.17	1354.76	0.71	0.26	1358.21	0.96	23.80	**1345.23**
16	480	C,D				**1622.69**	0.00	20.00	1634.99	0.76	1.15	1635.16	0.77	39.90	**1622.69**
17	240	C,D				**707.79**	0.00	0.75	710.22	0.34	0.16	708.76	0.14	68.50	**707.79**
18	300	C,D				**998.73**	0.00	2.50	1009.53	1.08	0.18	998.83	0.01	42.73	**998.73**
19	360	C,D				**1366.86**	0.00	6.00	1381.88	1.10	0.25	1367.20	0.02	112.80	**1366.86**
20	420	C,D				**1821.15**	0.00	8.40	1840.57	1.03	0.55	1822.94	0.10	71.42	**1821.15**
Average				0.74	42.05		0.00	72.94		0.93	0.63		0.60	49.33	

[i] Pentium II (400 MHz).

[j] For C instances, see Mester and Bräysy (2005). Otherwise, see Mester (2004)

[k] Pentium IV (2 GHz).

[l] Best value obtained in several experiments.

[m] Pentium (900 MHz).

5. Conclusion

In recent years several new metaheuristics have been put forward for the solution of the VRP. These combine a variety of principles including tabu search, population search and learning mechanisms. The best methods combine population search and local search, thus providing at the same time breadth and depth in the solution space exploration. All algorithms described in this study are highly accurate and some are also quite fast. What is now needed is a greater emphasis on simplicity and flexibility.

Acknowledgments This work was partially funded by the Canadian Natural Sciences and Engineering Research Council under grants 227837-00, OGP00338816, 105384-02 and OGP00039682 This support is gratefully acknowledged. Thanks are due to Pierre Sormany for a useful suggestion.

References

Baldacci, R., Hadjiconstantinou, E.A. and Mingozzi, A. (2004). An exact algorithm for the capacitated vehicle routing problem based on a two-commodity network flow formulation. *Operations Research*, 52(5):723 - 738.

Barbarosoğlu, G. and Ögür, D. (1999). A tabu search algorithm for the vehicle routing problem. *Computers & Operations Research*, 26:255 - 270.

Bean, J.C. (1994). Genetic algorithms and random keys for the sequencing and optimization. *ORSA Journal on Computing*, 6:154 - 160.

Beasley, J.E. (1983). Route-first cluster-second methods for vehicle routing. *Omega*, 11:403 - 408.

Berger, J. and Barkaoui, M. (2004). A new hybrid genetic algorithm for the capacitated vehicle routing problem. *Journal of the Operational Research Society*, 54:1254 - 1262.

Christofides, N., Mingozzi, A., and Toth, P. (1979). The vehicle routing problem. In: N. Christofides, A. Mingozzi, and P. Toth (eds.), *Combinatorial Optimization*, pp. 315 - 338. Wiley, Chichester.

Clarke, G. and Wright, J.W. (1964) Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research*, 12:568 - 581.

Codenetti, B., Manzini, G., Margara, L., and Resta, G. (1996). Perturbation: An efficient technique for the solution of very large instances of the Euclidean TSP. *INFORMS Journal on Computing*, 8:125 - 133.

Cordeau, J.-F., Desaulniers, G., Desrosiers, J., Solomon, M.M., Soumis, F. (2002a). VRP with time windows. In: P. Toth and D. Vigo (eds.), *The Vehicle Routing Problem*, pp. 157 - 193. SIAM Monographs on Discrete Mathematics and Applications, Philadelphia.

Cordeau, J.-F., Gendreau, M., and Laporte, G. (1997) A tabu search heuristic for periodic and multi-depot vehicle routing problems. *Networks*, 30:105 - 119.

Cordeau, J.-F., Gendreau, M., Laporte, G., Potvin, J.-Y., and Semet, F. (2002b). A guide to vehicle routing heuristics. *Journal of the Operational Research Society*,

53:512 – 522.

Cordeau, J.-F. and Laporte, G. (2004). Tabu search heuristics for the vehicle routing problem. In: C. Rego and B. Alidaee (eds.), *Metaheuristic Optimization via Memory and Evolution: Tabu Search and Scatter Search*, pp. 145 – 163. Kluwer, Boston.

Cordeau, J.-F., Laporte, G., and Mercier, A. (2001). A unified tabu search heuristic for vehicle routing problems with time windows. *Journal of the Operational Research Society*, 52:928 – 936.

Cordeau, J.-F., Laporte, G., and Mercier, A. (2004). An improved tabu search algorithm for the handling of route duration constraints in vehicle routing problems with time windows. *Journal of the Operational Research Society*, 55:542 – 546.

Desaulniers, G., Desrosiers, J., Erdmann, A., Solomon, M.M., and Soumis, F. (2002). VRP with pickup and delivery. In: P. Toth and D. Vigo (eds.), *The Vehicle Routing Problem*, pages 225 – 242. SIAM Monographs on Discrete Mathematics and Applications, Philadelphia.

Dongarra, J.J. (2004). Performance of various computers using standard linear equations software. Technical Report CS – 89 – 85. Computer Science Department, University of Tennessee.

Drezner, Z. (2003). A new genetic algorithm for the quadratic assignment problem. *INFORMS Journal on Computing*, 15:320 – 330.

Dueck, G. (1993). New optimization heuristics: The great deluge algorithm and the record-to-record travel. *Journal of Computational Physics*, 104:86 – 92.

Ergun, Ö., Orlin, J.B., and Steele-Feldman, A. (2003). Creating Very Large Scale Neighborhoods Out of Smaller Ones by Compounding Moves: A Study on the Vehicle Routing Problem. Working paper, Massachusetts Institute of Technology.

Fahrion, R. and Wrede, M. (1990). On a principle of chain-exchange for vehicle-routeing problems (1-VRP). *Journal of the Operational Research Society*, 41:821 – 827.

Fisher, M.L., and Jaikumar, R. (1981). A generalized assignment heuristic for the vehicle routing problem. *Networks*, 11:109 – 124.

Gendreau, M., Hertz, A., and Laporte, G. (1992). New insertion and postoptimization procedures for the traveling salesman problem. *Operations Research*, 40:1086 – 1094.

Gendreau, M., Hertz, A., and Laporte, G. (1994). A tabu search heuristic for the vehicle routing problem. *Management Science*, 40:1276 – 1290.

Gendreau, M., Laporte, G., and Potvin, J.-Y. (2002). Metaheuristics for the VRP. In: P. Toth and D. Vigo (eds.), *The Vehicle Routing Problem*, pages 129 – 154. SIAM Monographs on Discrete Mathematics and Applications, Philadelphia.

Ghaziri, H. (1991). Solving routing problems by a self-organizing map. In: T. Kohonen, K. Makisara, O. Simula, and J. Kangas (eds.), *Artificial Neural Networks*, pp. 829 – 834. North-Holland, Amsterdam.

Ghaziri, H. (1996). Supervision in the self-organizing feature map: Application to the vehicle routing problem. In: I.H. Osman and J.P. Kelly (eds.), *Meta-Heuristics: Theory and Applications*, pp. 651 – 660. Kluwer, Boston.

Gillett, B.E. and Miller, L.R. (1974). A heuristic algorithm for the vehicle dispatch problem. *Operations Research*, 22:340 – 349.

Golden, B.L., Wasil, E.A., Kelly, J.P., and Chao, I-M. (1998). Metaheuristics in vehicle routing. In: T.G. Crainic and G. Laporte (eds.), *Fleet Management and Logistics*, pp. 33 – 56. Kluwer, Boston.

Kinderwater, G.A.P. and Savelsbergh, M.W.P. (1997). Vehicle routing: Handling edge exchanges. In: E.H.L. Aarts and J.K. Lenstra (eds.), *Local Search in Combinatorial*

Optimization, pages 337 – 360. Wiley, Chichester.

Laporte, G. and Semet, F. (2002). Classical heuristics for the capacitated VRP. In: P. Toth and D. Vigo (eds.), *The Vehicle Routing Problem*, pp. 109 – 128. SIAM Monographs on Discrete Mathematics and Applications, Philadelphia.

Li, F., Golden, B.L., and Wasil, E.A. (2005). Very large-scale vehicle routing: New test problems, algorithms, and results. *Computers & Operations Research*, 32:1165 – 1179.

Lin, S. (1965). Computer solution of the traveling salesman problem. *Bell System Technical Journal*, 44:2245 – 2269.

Liu, F.-H., and Shen, S.-Y. (1999). A route-neighbourhood-based metaheuristic for vehicle routing problem with time windows. *European Journal of Operational Research*, 118:485 – 504.

Matsuyama, Y. (1991). Self-organization via competition, cooperation and categorization applied to extended vehicle routing problems. In: *Proceedings of the International Joint Conference on Neural Networks*, pp. 385 – 390. Seatle, WA.

Mester, D. (2004). Private communication.

Mester, D. and Bräysy, O. (2004). Active guided evolution strategies for the large scale vehicle routing problems with time windows. *Computers & Operations Research*, 32:1593 – 1614.

Mladenović, N. and Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, 24:1097 – 1100.

Moscato, P. and Cotta, C. (2003). A gentle introduction to memetic algorithms. In: F. Glover and G.A. Kochenberger (eds.), *Handbook of Metaheuristics*, pp. 105 – 144. Kluwer, Boston.

Naddef, D. and Rinaldi, G. (2002). Branch-and-cut algorithms for the capacitated VRP. In: P. Toth and D. Vigo (eds.), *The Vehicle Routing Problem*, pp. 53 – 84. SIAM Monographs on Discrete Mathematics and Applications, Philadelphia.

Or, I. (1976). *Traveling Salesman-Type Combinatorial Problems and their Relation to the Logistics of Regional Blood Banking.* Ph.D. Thesis, Northwestern University, Evanston, IL.

Osman, I.H. (1993). Metastrategy simulated annealing and tabu search algorithms for the vehicle routing problem. *Annals of Operations Research*, 41:421 – 451.

Paessens, H. (1988). The savings algorithm for the vehicle routing problem. *European Journal of Operational Research*, 34:336 – 344.

Potvin, J.-Y. (1996). Genetic algorithms for the traveling salesman problem. *Annals of Operations Research*, 63:339 – 370.

Potvin, J.-Y. and Rousseau, J.-M. (1995). An exchange heuristic for routing problems with time windows. *Journal of the Operational Research Society*, 46:1433 – 3446.

Prins, C. (2004). A simple and effective evolutionary algorithm for the vehicle routing problem. *Computers & Operations Research* 31:1985 – 2002.

Rechenberg, I. (1973). *Evolutionsstrategie.* Fromman-Holzboog, Stuttgart.

Reeves, F. (2003). Genetic algorithms. In: F. Glover and G.A. Kochenberger (eds.), *Handbook of Metaheuristics*, pp. 55 – 82. Kluwer, Boston.

Rego, C. (1998). A subpath ejection method for the vehicle routing problem. *Management Science*, 44:1447 – 1459.

Rego, C. and Roucairol, C. (1996). A parallel tabu search algorithm using ejection chains for the vehicle routing problem. In: *Meta-Heuristics: Theory and Applications*, pp. 661 – 675, Kluwer, Boston.

Reimann, M., Doerner, K., and Hartl, R.F. (2004). D-Ants: Savings based ants divide and conquer the vehicle routing problem. *Computers & Operations Research*,

31:563 – 591.

Renaud, J., Boctor, F.F., and Laporte, G. (1996). A fast composite heuristic for the symmetric traveling salesman problem. *INFORMS Journal on Computing*, 8:134 – 143.

Rochat, Y. and Taillard, É.D. (1995). Probabilistic diversification and intensification in local search for vehicle routing. *Journal of Heuristics*, 1:147 – 167.

Shaw, P. (1998). Using constraint programming and local search methods to solve vehicle routing problems. In: M. Maher and J.-F. Puget (eds.), *Principles and Practice of Constraint Programming*, pp. 417 – 431. *Lecture Notes in Computer Science*, Springer-Verlag, New York.

Schumann, M. and Retzko, R. (1995). Self-organizing maps for vehicle routing problems — minimizing an explicit cost function. In: F. Fogelman-Soulie (ed.), *Proceedings of the International Conference on Artificial Neural Networks*, pp. 401 – 406. Paris.

Solomon, M.M. (1987). Algorithms for the vehicle routing and scheduling problems with time windows. *Operations Research*, 35:254 – 265.

Taillard, É.D. (1993). Parallel iterative search methods for vehicle routing problem. *Networks*, 23:661 – 673.

Tarantilis, C.-D. and Kiranoudis, C.T. (2002). Bone Route: An adaptive memory-based method for effective fleet management. *Annals of Operations Research*, 115:227 – 241.

Thompson, P.M. and Psaraftis, H.N. (1993). Cyclic transfer algorithms for multivehicle routing and scheduling problems. *Operations Research*, 41:935 – 946.

Toth, P. and Vigo, D. (2003). The granular tabu search and its application to the vehicle routing problem. *INFORMS Journal on Computing*, 15:333 – 346.

van Breedam, A. (1994). *An Analysis of the Behavior of Heuristics for the Vehicle Routing Problem for a Selection of Problems with Vehicle-Related, Customer-Related, and Time-Related Constraints.* Ph.D. Dissertation, University of Antwerp.

Voudouris, C. (1997). *Guided Local Search for Combinatorial Problems.* Dissertation, University of Essex.

Xu, J. and Kelly, J.P. (1996). A network flow-based tabu search heuristic for the vehicle routing problem. *Transportation Science*, 30:379 – 393.

Chapter 10

ROUTING PROPANE DELIVERIES

Moshe Dror

Abstract This chapter is about solving the problem of propane deliveries. It is commonly viewed as a representative problem of a much larger family of hard problems of considerable practical significance. This problem has been on the "front burner" of the logistics academic and practitioners community for over twenty years. In this chapter I attempt to describe the practices of a propane distribution company and to summarize the literature on the more general topic of inventory routing. It is one person's point of view and I apologize *ex ante* for my unavoidable biases.

1. Introduction

The outline of this chapter is as follows:

- Personal experience, the basic problem and its variants,
- Real-life examples, starting with Bell et al. (1983),
- The initial analysis of Federgruen and Zipkin (1984),
- The propane distribution as in Dror (1983), and extensions,
- The new wave: Kleywegt et al. (2002, 2004) and Adelman (2003a,b, 2004),
- Summary

1.1 Personal experience

One day in the Fall of 1995, I had the opportunity to spend a day in the "passenger seat" of a propane delivery truck observing a propane delivery operation in a rural area of upstate New York 'first-hand.' More than 10 years before that, in the Spring of 1983, I completed a Ph.D. thesis on this very topic. However, I was only given the chance to observe a "real-life" delivery operation for one day in 1995. While writing my thesis, I did visit the headquarters of a large propane distribution company 3 – 4 times; I talked to it's managers, collected data from the operational offices of one district, but did not drive a delivery truck until Fall of 1995. Talking to the driver of the delivery truck for the whole day, visiting the different customers, and observing the operations first hand was a new experience. On that day we visited customers who had almost empty tanks and customers whose tanks were full, and a number of customers in between the two extremes. After leaving the depot, before starting on their routes, the propane delivery drivers gather (unofficially) at a local diner for a morning coffee and a short chat. They discuss their respective routes and the customers on their routes. Past experiences, road conditions, and advice are freely exchanged. Only after that early gathering do the drivers go ahead and deliver propane to their customers. I believe that this kind of first-hand viewing of an operation is very valuable for understanding and subsequent analysis of logistics operations. I learned a lot that day.

1.2 The basic problem of propane distribution

The basic propane truck delivery operations are usually conducted in rural areas. Propane (and similar liquid or gas products) is used to heat individual houses and other facilities, which are not connected by a delivery (rigid-pipe) network. In densely populated cities economies of scale dictate a different mode of operation via a connected delivery (pipeline) network which provides the commodity (propane) on demand to its customers just as electricity and water is delivered. Rural propane customers are dispersed in a certain geographical area serviceable from a central facility (a depot) located in their area and are serviced from this single depot by delivery trucks (special tank-trucks filled with propane). The delivery trucks are usually of a few (2 – 3) fixed capacity types. Each customer has a propane tank located on his property (next to his house). The tank usually belongs to the propane company servicing the customer on a long-term basis. The customer tanks come in a number (5 to 8) of different sizes. A propane service contract requires the company to maintain a sufficient level of propane in the customer's tank at any time. Once the propane is delivered to a customer, a payment invoice is issued requiring the customer to pay within a week or so. In essence, once the propane is delivered it belongs to the customer. In this setting the subsequent inventory holding costs for the propane are incurred by the customer. It is quite common to hear customers complain about the company filling their tanks to capacity just before summer — a period of very low propane consumption. Even in sparse rural areas there might be pockets of a few locally concentrated users connected by a rigid pipe network to a single propane storage facility. In this case, each customer is equipped with a metering device and billed periodically for his/her consumption. Thus, the propane in the tank incurs a holding cost absorbed by the company. From an academic perspective one might want to view propane distribution as what is called in business practice as *vendor managed inventory* replenishment (VMI). In this chapter we do not try to link propane distribution, modelling, and the literature, with the more general area of VMI.

In a medium size propane service district (2,000 – 5,000 customers, though 10,000 customer districts can also be found) a company uses (owns) 3 to 7 delivery trucks. Each morning, the truck drivers are given a list of customers (their location, tank sizes, names, if necessary special individual characteristics, etc.) and simple dispatching instructions and off they go on their delivery routes. At times, they are unknowingly assigned a customer whose tank is still full and subsequently a service visit is wasted. Presently, without the more prevalent information technology, the exact demand (tank emptiness) becomes known only when the driver checks the tank's propane level on his arrival. Truck dispatching is based on partial information, experience, and a demand forecast which are all used to generate daily dispatching lists. Each day (5 days in every regular week) a preselected subset of customers is replenished. The basic refill policy is to fill-up customers' tanks to their capacity on each replenishment (service) visit. Since the individual customers' consumption rates are only estimated (they are viewed as random variables), this gives rise to stock-out events which necessitate "emergency" deliveries, which represent propane deliveries in response to customer's request because of an empty tank. These emergency deliveries have to be performed 7 days a week, weekends and holidays included, and are quite costly. In one Pennsylvania district of slightly over 2,000 customers, from which initial data was collected for the 1983 study (Dror, 1983), there were about 100 stock-out related deliveries in a period of three months. From the propane company point of view, it is pertinent to operate efficiently with a long time view perspective. That is, the company would like to design the logistic operations which minimize its long run cost of delivering propane to a given set of customers (a district). It is quite common for such US companies to own (operate) 300 – 500 districts. In some larger districts, propane companies may own separate propane storage facilities called satellite depots (see Bard et al., 1998). These satellite depot facilities serve as intermittent refill points for the delivery trucks enabling the trucks to extend their delivery routes without returning to the home depot for refill. It is a quite complex routing setting with many parameters which are only partially known at the beginning of the work day and vary over time. Recently, electronic automatic measuring and reporting devices have been introduced into the propane tanks. These devices relay tank propane level information to the district office. Thus, wasteful visits to customers which do not require certain minimal volume propane delivery could be avoided. I do not know how common this automated reporting technology is. Even if this technology is presently quite common, it only helps in deciding who not to service on a given day but does not eliminate many of the difficulties inherent in efficient propane distribution planning. Key question: What does the district know about the status and evolution of a system (the customers, the delivery trucks, and the depots), and when do they know it?

2. The industrial gases inventory case

A motivating example from Adelman (2003), attributed to Bell et al. (1983), is very useful for introducing some basic difficulties regarding the inventory routing decisions. This example with deterministic daily demands is restated in Figure 10.1 (where the internode distances are in miles) and in the corresponding table below.

Customer	Tank Capacity	Daily Demand
A	5000 gallons	1000 gallons per day
B	3000	3000
C	2000	2000
D	4000	1500

Assuming vehicles of 5000 gallon capacity, a simple inventory routing solution is to replenish customers A and B together every day, and cus-

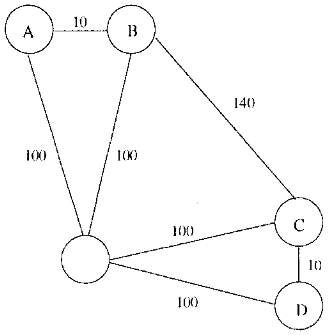


Figure 10.1. A simple example with 4 customers

tomers C and D together every day. The daily "cost" of this solution is 420 miles and uses two vehicles.

An improved routing solution consists of a cycling replenishment pattern which repeats every two days. On the first day use only one vehicle and deliver 3000 gallons to customer B and 2000 gallons to customer C, travelling 340 miles. On the second day two vehicles are used. Vehicle 1 delivers 2000 gallons to customer A and 3000 to customer B. Vehicle 2 delivers 2000 to customer C and 3000 to customer D. Each vehicle travels on that day 210 miles. Thus, the average daily distance travelled over two day period is 380 miles. This is 10% lower than the first solution. It is interesting to note that even though this solution has been known since 1983, Adelman (2003a) is the first to derive it analytically and prove its optimality!

The above example illustrates that finding an optimal replenishment solution even in a simple (deterministic) setting can be quite difficult. However, a number of successful distribution solutions have been developed over the years. One of the first success stories is that of industrial gases distribution systems developed by Marshall L. Fisher (Fisher et al., 1982) and his associates and is described below in some detail.

2.1 Industrial gases delivery system

The inventory management of industrial gases introduces real-life logistics issues very similar to the inventory management for propane deliveries. We repeat the operational description from Bell et al. (1983).

In the industrial gases case the main products are oxygen, nitrogen, hydrogen, argon, and carbon monoxide. Essentially, liquid oxygen and nitrogen are manufactured in highly automated plants. The plants serve as supply depots where liquified gases are stored at a temperature less than -320°F. The liquified gases are distributed in cryogenic bulk tankers to industrial users and hospitals. Storage tanks at customer sites are monitored by the supplier under long-term contracts. Similarly to propane, the supplier of liquid gases delivers the product at his discretion with the guarantee of continuous availability. In 1982, one company — the Air Products corporation, employed 340 trucks which travelled over 22 million miles a year. Distribution efficiency is the main competitive tool differentiating among the producers since the manufacturing costs among different companies are about the same and lower distribution costs allow lower pricing or higher profit margins. The decisions taken in distribution operations set the customers' tanks inventory levels, by determining how much to deliver, how to combine the different loads on a truck and how to route the truck. That is, inventory management at customer locations is integrated with vehicle routing. A single liquid gas plant distribution problem may involve several hundred customers and about 20 trucks. Complicating factors include estimating customer usage rates which vary considerably over time. Inventory must be maintained above a specified safety-stock level and customers are not open for delivery every day of the week or during every hour of the day (this varies across customers). Trucks also differ in their capacity and operating costs which might even change by state boundaries because of different state laws from one state to another.

There are numerous other driving cost related characteristics which need to be accounted for in real-life dispatching. We will not go into this here and the more specific details of the system are described in Bell et al. (1983). As we will see in the case of propane delivery, liquid gas delivery also requires careful forecasting of the rate at which each customer is consuming its product and the calculation of the "best" time to deliver, in terms of cost and delivery feasibility. What is usually known in terms of consumption rates is the inventory levels which are recorded before each delivery. In the case of liquid gases some customers are contacted (telephoned) from time to time to establish exact inventory levels to facilitate forecasting and dispatching. However, when deciding on vehicle routing sequences, it is comforting to note that feasible routes contain between two to four customers only. That is, even when dispatching 10 to 30 trucks daily, efficient routes are not very difficult to construct. In fact, the system described in Bell et al. (1983), is designed to produce a distribution schedule for the next two to five days. To select the delivery routes, first, a set of possible routes is generated with the sequencing order of customer stops. However, the delivery amount is not specified in the route generation stage. Since the number of customers on a route is small, "the number of technically feasible routes is not unreasonably large." A large mixed-integer programming model is solved each time which selects the routes from the set of externally generated potential routes, and determines the vehicle, the time each route starts, and the amount to be delivered to each customer on the route. We do not repeat here the mathematical formulation of the route selection model. However, we note that it incorporates parameters which represent the effect of short term delivery decisions on the events beyond the horizon of the model which is two to five days. Otherwise, a short term solution would "paint" the long term efficiency objective into a bad corner. One main difference between delivery of industrial liquid gases and propane is that in the propane case the policy is to fill the customer's tank to capacity on each service visit. In addition, delivery routes usually have between 4 – 12 customers making the route construction scheme more difficult.

In the next section we focus on one of the first academic attempts to model inventory distribution.

3. The initial analysis

The first, more "mathematical" analysis of the inventory routing problem is contained in the paper by Federgruen and Zipkin (1984). In that paper Federgruen and Zipkin examine:
"the combined problem of allocating scarce resources available at some central depot among several locations (or "customers"), each experiencing a random demand pattern, while deciding which deliveries are to be made by each set of vehicles and in what order."

It sounds very promising and is viewed as an extension of the standard vehicle routing problem where the *"deliveries serve to replenish the inventories to levels that appropriately balance inventory carrying and shortage costs, but thereby incur transportation costs as well."*

Essentially, the problem is examined from the point of view of inventory management in multiple locations with the added complication of routing — constructing delivery routes for a fleet of capacitated vehicles. The inventory status information for each location is assumed to be available at the beginning of the day and delivery quantities together with routes for each vehicle are then computed. The deliveries are executed and then the actual demand is observed with its resulting subsequent holding and shortage penalties. There is no requirement of visiting all the customers. As the authors state *"ours is the first attempt to integrate the allocation and routing problems into a single model."* The importance of such integration and analysis is very nicely motivated by Herron (1979). Federgruen and Zipkin (1984) present a very direct model which views each customer's inventory from the perspective of the newsvendor problem. That is, there are zero delivery costs to a customer and the deliveries to the customer(s) are driven by the shortage costs. For completeness we restate the mathematical formulation below slightly changing the original notation.

paragraphConstants

NV = number of vehicles
$\hat{n} =$ number of locations, with 0 indicating the depot location
$Q_v =$ capacity of vehicle v
$c_{ij} =$ cost of direct travel from location i to location j
$F_i(\cdot) =$ cumulative distribution function of the one period demand in location i, assumed strictly increasing
$h_i^+ =$ inventory carrying cost per unit in location i
$h_i^- =$ shortage cost per unit in location i

$I_i =$ initial inventory at location i
$A =$ total amount of product available at the central depot.

Variables.

$y_{ik} = 1$, if delivery point i is assigned to route (vehicle) k, and is 0 otherwise.

$x_{ijk} = 1$, if vehicle k travels directly from location i to location j, and is 0 otherwise.

q_i is the amount delivered to location i. Note that in the spirit of classical VRP formulations, at most one vehicle visits any given location.

Just like in the newsvendor inventory model, the inventory cost function $C_i(\cdot)$ and its derivative $C_i'(\cdot)$, for $i = 1, \ldots, \hat{n}$, are given by

$$C_i(q_i) = \int_{I_i+q_i}^{\infty} h_i^-(\xi - I_i - q_i)dF_i(\xi) + \int_0^{I_i+q_i} h_i^+(I_i + q_i - \xi)dF_i(\xi)$$

$$C_i'(q_i) = (h_i^+ + h_i^-)F_i(I_i + q_i) - h_i^-$$

Now the mathematical formulation expressing a single period cost minimization is stated as follows:

$$\min \sum_{i,j,k} c_{ij}x_{ijk} + \sum_i C_i(q_i) \tag{10.1}$$

subject to the following constraints

$$\sum_i q_i y_{ik} \leq Q_k, \qquad k = 1, \ldots, \text{NV}; \tag{10.2}$$

$$\sum_i q_i \leq A; \tag{10.3}$$

$$\sum_{k=1}^{\text{NV}} y_{0k} = \text{NV}; \tag{10.4}$$

$$\sum_{k=0}^{\text{NV}} y_{ik} = 1, \qquad i = 1, \ldots, \hat{n}; \tag{10.5}$$

$$\sum_i x_{ijk} = y_{jk}, \qquad j = 0, \ldots, \hat{n}; k = 1, \ldots, \text{NV}; \tag{10.6}$$

$$\sum_j x_{ijk} = y_{ik}, \qquad i = 0, \ldots, \hat{n}; k = 1, \ldots, \text{NV}; \tag{10.7}$$

$$\sum_{(i,j)\in S\times S} x_{ijk} \leq |S| - 1, \qquad S \subseteq \{1, \ldots, \hat{n}\}, 2 \leq |S| \leq \hat{n} - 1, \quad k = 1, \ldots, \text{NV}; \tag{10.8}$$

$$x_{ijk} \in \{0.1\}, \qquad i, j = 0, \ldots, \hat{n}; k = 1, \ldots, \text{NV}; \tag{10.9}$$

$$y_{ik} \in \{0,1\}, \qquad i = 0, \dots, \hat{n}; k = 1, \dots, \text{NV}; \tag{10.10}$$

$$q_i \geq 0, \qquad i = 1, \dots, \hat{n}. \tag{10.11}$$

This formulation has a mixture of 0-1 integer linear VRP type constraints and nonlinear constraints (10.2). It is a single period formulation which generates deliveries driven essentially by the expected shortage costs. For the propane delivery setting the formulation unnecessarily assumes limited supply at the depot (constraint (10.3)) but does not contain the tank capacity constraints at the customer locations. It also incorporates an inventory holding cost per unit per unit time — which is not directly applicable to the propane case. More importantly, it charges inventory shortage costs per unit per unit time. In the propane case the shortage costs might be best represented by some step function representing customer specific cost of emergency delivery. However this is a very nice model and for a fixed vector y it decomposes into simpler problems. On the one hand we get the inventory allocation problem, and on the other NV-TSPs, one for each route-vehicle. It is an attractive approach but unfortunately not appropriate for the propane delivery long term optimization problem. The primary reason is that this model is a single period optimization which does not project short-term decisions into long term cost implications. Thus, it might attempt to myopically "paint" a sequence of one period solutions into a long-term bad solution. We will return to this point later when we compare this model with the later model taken from Dror and Ball (1987).

4. The initial propane delivery model (Dror and Ball, 1987)

We first describe a number of very simple principles guiding the efficiency of propane deliveries.

- Visit a customer as infrequently as possible. This translates simply to delivering as much as possible to a given customer on each visit. In other words, if it is feasible to deliver as much as the customer's tank capacity on each visit, then do so.
- If it does not cost extra to visit a customer then replenish him/her. That is, if you can save a future service visit which has a positive cost by delivering early at no (or small) cost, then go ahead and replenish.
- Replenishing earlier reduces the risk of stock-out and increases the present value of cash-flow (see Dror and Trudeau, 1996).

To develop the above principles more formally we describe a basic analysis of a single customer with a fixed sized tank (size T), a cost of refill b (b_i for customer i), daily consumption rate μ (deterministic for now) and initial inventory I_0. Assume the analysis for an n-day period (for "large" n). Note that the tank is refilled on each visit and the customer is serviced when the tank inventory reaches zero.

In this case, the (optimal) cost of service visits is:

$$\frac{b(n\mu - I_0)}{T}.$$

If we plan deliveries for the next m ($\ll n$) days, and select this planning horizon m small enough such that no customer will need more than one replenishment during the next m days, then for each customer we will have to examine two possible cases:

(1) If $I_0 - m\mu < 0$, this customer must be replenished during the next m days, otherwise a stockout occurs.

(2) If $I_0 - m\mu \geq 0$, the customer need not be replenished during the next m days.

If case 1 occurs, then the optimal policy would dictate that the replenishment take place on day $t^* = I_0/\mu$, allowing for non-integer t^* values. Clearly, if case 2 occurs it is best not to replenish the customer in this current m day period. This single customer analysis is very basic and does not communicate any problem dynamics. What if the capacity of the system is insufficient to replenish all the customers whose t^* day falls on a specific day during the current m-day period, but is sufficient to replenish all the customers which have their best delivery day t^* fall during some day of the current m-day period? Some of these customers will have to be replenished on a day different than their corresponding t^*. Thus, just for that reason we have to calculate for each customer the marginal cost over n days, denoted $c(t)$, of replenishment on day t deviating from day t^*. That is, $c(t^*) \equiv 0$, and $c(t) > 0$, $t \neq t^*$. There are other important reasons for evaluating this marginal cost $c(t)$, for instance balancing the work load over time. Another quantity is calculated for the customers who do not need to be replenished on day t during the current m-day period. This quantity is denoted by $g(t)$ and represents the decrease in future costs (over an n day horizon) if the customer is replenished during the current m-day period at no cost instead of being replenished at his "best" day in a future m-day period at cost b. Below we repeat the calculation from Dror and Ball (1987).

If replenishment is executed on day t, then the closing inventory I_c is defined by

$$I_c = T - (m - t)\mu.$$

Now a simple difference calculation for $c(t)$ and $g(t)$ is as follows:

$$\begin{aligned} b\frac{(n\mu - (T-(m-t)\mu))}{T} - b\frac{(n\mu - (T-(m-t^*)\mu))}{T} &= \frac{b}{T}(I_c - \mu t) \\ &= c(t) \\ b\frac{(n\mu - (I_c - m\mu))}{T} - b\frac{(n\mu - (T-(m-t)\mu))}{T} &= b\frac{(T - I_c + t\mu)}{T} \\ &= b - c(t) = g(t) \end{aligned}$$

The above simple deterministic single customer analysis is extended to a more realistic stochastic model and later $c(t)$ and $g(t)$ are used as a cost coefficient in a multi-customer setting. The major weakness of the above analysis lies in its assumption that we know b — the cost of visiting a customer, and that this value, even though customer specific, remains constant for all the replenishments in the n-period. Clearly, this is not entirely true in real-life propane distribution. We will return to this important point later on.

4.1 The stochastic single customer

Let r_t denote the amount of propane consumed by customer on day t. Normally, we do not know the value of r_t. We do not know its exact value for past days, which is less important, and, even more so, we do not know its value for future days. That is, r_t, $t = 1, 2, \ldots$ are random variables. For simplicity we assume that r_ts are independent identically distributed random variables for each t (consider that the seasonality effects are removed) with mean μ, variance σ^2, and cumulative distribution function $F(\cdot)$. The randomness (and variability) of consumption makes the replenishment scheduling a risky proposition. Guessing that there is enough propane in a customer's tank when there is not usually results in a high cost emergency replenishment. Guessing that there is little left in the tank when there is a lot left results in an almost equally costly visit. Thus, it is of value to calculate the replenishment day which balances the risk of the two cost penalties and at the same time accounts for future implications of an expected delivery volume that is less than the tank capacity. We describe below this calculation assuming that the customer's tank is full on day 1.

Let $R_t = \sum_{i=1}^{t} r_i$ denote the cumulative consumption over a t day period. Let $P_S(t)$ denote the probability that a stockout occurs on day t given that the tank has not been refilled prior to day t. Thus, assuming that $\mu < T$,

$$P_S(t) = \text{Prob}\{R_{t-1} \leq T < R_t\} = \text{Prob}\{R_{t-1} \leq T\} - \text{Prob}\{R_t \leq T\}$$

$$= F^{(t-1)}(T) - F^{(t)}(T)$$

where $F^{(k)}(T)$ is the k-fold convolution of F (set $F^{(0)} \equiv 0$).

Now we are in a position to write down the expression for the expected cost during the next $n + m$ days associated with a delivery on day t denoted by $E(t)$. Denote by $k^* = \max\{k : \text{such that } k\mu \leq T\}$,

$$\begin{aligned} E(t) &= \sum_{i=1}^{t-1} \big(S + c(i)\big) P_S(i) + \left(1 - \sum_{i=1}^{t-1} P_S(i)\right) c(i) \\ &= \sum_{i=1}^{t-1} \big(S + c(t) - c(i)\big) P_S(i) - c(t) \end{aligned}$$

for $1 \leq t \leq k^*$ (or $k^* + 1$).

What is particularly interesting is that in Dror and Ball (1987), it has been proven that $E(t)$, $1 \leq t \leq k^*$ is a strictly convex function by proving that $P_S(t) > P_S(t-1)$, $2 \leq t \leq k^*$, for r_ts normally distributed with coefficient of variation ≤ 1. Moreover, in Kreimer and Dror (1990), this result was strengthened by proving that the relation $P_S(t) > P_S(t-1)$, $2 \leq t \leq k^*$ holds for a number of other interesting distributions. In Dror (2002), the relation $P_S(t) > P_S(t-1)$, $2 \leq t \leq k^*$ (monotonicity) was stated formally as a more general conjecture.

In summary, the result is that $E(t)$, when viewed as a continuous function of t, is convex; thus it achieves its minimum at a single point (or at most 2 points as a discrete function). That is, let $E(t^*) = \min\{E(t) : 1 \leq t \leq k^*\}$ determines the "best" (minimal expected cost) day for replenishment — t^*. It is appropriate to note that a similar analysis (with similar results) has been conducted by Jaillet et al. (2002).

4.2 The propane routing model

The notation is similar but not identical to the presentation of the model by Federgruen and Zipkin (1984).

Constants.

NV = number of vehicles
M = the set of customers, with 0 indicating the depot location
Q = capacity of a vehicle (homogenous vehicles)
c_{ij} = cost of servicing customer i then travelling from i to j
m = number of days in planning period.

The quantities defined in the previous subsection now become customer specific, so that we have T_i, b_i, μ_i, S_i, $c_i(t)$, $g_i(t)$, I_{0i}, and t_i^*, defined for all $i \in M$. Customer i's expected demand on day t is denoted by $q_i(t)$ and equal to $T_i - I_{0i} - \mu_i t$. Since not all customers in M need to be replenished during the current planning period, we partition the customers into two subsets. Let $\widehat{M} = \{i \in M$ be such that $t_i^* \leq m\}$ as the customers who must be replenished during the current planning period, and $M^c = M \setminus \widehat{M}$ the rest of customers. In addition, to simplify the formulation we denote by TSP(N) a travelling salesman problem solution for customers in $N \subset M$.

Variables.

y_{iwt} = 1, if customer i is assigned to route (vehicle) w on day t, and is 0 otherwise.

Now the mathematical formulation expressing a single period cost minimization is stated as follows:

$$\min \sum_{w=1}^{\mathrm{NV}} \sum_{t=1}^{m} \left(\mathrm{TSP}(N_{wt}) + \sum_{i \in \widehat{M}} c_i(t) y_{iwt} - \sum_{i \in M^c} g_i(t) y_{iwt} \right) \qquad (10.12)$$

subject to the following constraints

$$\sum_{w=1}^{\mathrm{NV}} \sum_{t=1}^{t_i^*} y_{iwt} = 1, \qquad \forall\, i \in \widehat{M}, \qquad (10.13)$$

$$\sum_{w=1}^{\mathrm{NV}} \sum_{t=1}^{t_i^*} y_{iwt} \leq 1, \qquad \forall\, i \in M^c, \qquad (10.14)$$

$$\sum_{i \in M} q_i(t) y_{iwt} \leq Q, \qquad w = 1, \ldots, \mathrm{NV}; t = 1, \ldots, m \qquad (10.15)$$

$$N_{wt} = \{i : y_{iwt} = 1\}, \qquad w = 1, \ldots, \mathrm{NV}; t = 1, \ldots, m \qquad (10.16)$$

$$y_{iwt} \in \{0, 1\}, \qquad \forall\, i, w, t \qquad (10.17)$$

The y_{iwt} variables indicate for customer i the replenishment day and the replenishment vehicle. We artificially require that customers be replenished before or at their best day t^*. Customers who do not have their best day fall in the current m-period, do not have to be replenished (10.14). Other than the term for TSP(N) in the objective function followed by the appropriate set partition in (10.16), the formulation resembles that of the generalized assignment problem. The stochasticity is captured by the t^*'s and the dynamics (long-term implications) by the $c_i(t)$'s and $g_i(t)$'s. The "big" problem is that of calculating the individual b_i values required for calculation of $c_i(t)$'s ($g_i(t)$'s). Dror and Ball (1987) offered only an approximation of unproven quality, but their computational tests compared very favorably with the real-life results (Trudeau and Dror (1992)). For practitioners, a solution system based on this approach is best described in Dror and Trudeau (1988). This (10.12) – (10.17) mathematical formulation is similar in spirit to the formulation from Bell et al. (1983). There are however a number of differences. It is not a set-covering approach. That is, it is not a scheme to cover a given set of customers by selecting routes, each containing a subset of customers, from a large family of externally generated routes. It is a customer selection approach which selects subsets of customers together with the days in which to replenish these customer subsets (see also Dror et al., 1985, 1986). In addition, the amounts delivered are determined by the replenishment day since the policy is always to fill-up the tank, and the delivery implications are explicitly projected forward. In Bell et al. (1983) the future implications of a present delivery are not clearly spelled out.

5. The Markov decision process approach for inventory routing

Clearly propane delivery routing is merely one representative of a large class of practical significant problems. Yet due to the inherent combinatorial and stochastic nature of this class, it remains notoriously intractable. Formulating the control problem as a Markov decision process represents an attractive modelling approach which captures most of the system dynamics intrinsic to propane delivery. Following Minkoff (1993), there have been a number of attempts to do just that. Markov decision process modelling of inventory routing has taken-off in the work of Kelywegt, et al. (2002, 2003). However, the concomitant contribution by Adelman (2003a,b, 2004) is the most promising solution approach yet. We attempt below to provide a brief summary of the main ideas in these works.

5.1 The Markov decision process model (MDP)

The Markov decision process model is stated as follows (We modify the models of Minkoff, 1993; Kleywegt et al., 2002, slightly to unify notation and assumptions.):

(1) The state variable $I = (I_1, \ldots, I_{\bar{m}})$, where M is the customer set and $\bar{m} = |M|$, represents the current amount of inventory at each customer. The constant vector $T = (T_1, \ldots, T_{\bar{m}})$ represents the customers' tank capacities. Thus, the inventory can vary (continuously or discretely) in the product state space $\mathcal{I}$ bounded below by the zero vector and above by the vector of tank capacities. Let $I_t = (I_{1t}, \ldots, I_{\bar{m}t}) \in \mathcal{I}$ denote the inventory state at time t.

(2) Given a state vector $I \in \mathcal{I}$, denote by $A(I)$ the set of all feasible decisions. A decision $a \in A(I_t)$ in time t selects (i) the subset of customers for replenishment, and (ii) the vehicles' replenishment routes. Note that the amount to be replenished can be either a part of the decision, or, like in a partially observable MDPs, the outcome of customer inventory level observed on delivery if we always refill customer's tank. In the second case, the decision a will have to contain an estimate of what should be the replenishment volume. However, the actual delivery value might be quite different. Let $a_t \in A(I_t)$ be the decision chosen at time t. In our propane Markov decision model we assume that the exact demand is revealed only when the vehicle arrives at customer location and the policy is to fill-up the tank.

(3) The system's randomness is expressed in terms of the daily consumption rate $r = (r_1, \ldots, r_{\bar{m}})$. That is, the amount that can be delivered to customer i at time t (the demand q_{it}), equals $T_i - I_{it}$, which is a random variable dependent on r_i's since the last replenishment. The amount delivered to customer i (denoted by $d_{it}(a)$ in deference to q_{it}) by executing the policy a on day t can be either zero, a predetermined quantity d_{it}, or $T_i - I_{it}$ (if a replenishment always fills-up the tank). Let $U = \{I_{t+1} \in R_+^{\bar{m}} : ((I_{1t} - r_{1t} + d_{1t}(a)), \ldots, (I_{\bar{m}t} - r_{\bar{m}t} + d_{\bar{m}t}(a)))\}$. The known joint probability distribution F of customers demands $q_t = (q_{1t}, \ldots, q_{\bar{m}t})$ gives us a known transition function in the form of a conditional probability distribution. That is, for any state $I \in \mathcal{I}$, and decision $a \in A(I)$, we have

$$\text{Prob}\{U \mid I_t, a] = F(U \mid I_t, a).$$

(4) Let $\zeta(I, a)$ denote the expected single stage net reward (cost) if the process is in state I at time t, and decision $a \in A(I)$ is implemented. Note that not only the exact customer demands are random variables

but also the costs of the corresponding routing solution is a random variable since we do not know this cost until we execute the route and incur the additional recourse routing costs in response to route failures (see Trudeau and Dror, 1992).

(5) The objective is to maximize the expected total discounted value (or the present value of the cash flow), over an infinite (or finite "long" n-day) horizon. The decisions in time t, $a(t)$, are restricted to the feasible sets $A(I_t)$ for each t and depend only on the history $(I_0, a_0, \ldots, I_{t-1}, a_{t-1}, I_t)$ of the process up to time t. Let Π be the set of policies which depend on the history up to time t. Let $\alpha \in [0, 1)$ denote the discount factor. Let $\nu^*(I)$ denote the optimal expected value given the initial state is I, then

$$\nu^*(I) \equiv \sup_{\pi \in \Pi} E^{\pi} \left[\sum_{t=1}^{\infty} \alpha^t \zeta(I_t, a_t) | I_0 = I \right]$$

Following standard text book analysis (see Bertsekas and Shreve, 1978), a stationary deterministic policy π selects a decision $\pi(I) \in A(I)$ based only on the current state I. In principle, under some conditions, one can solve the above system by dynamic programming, computing the optimal value function ν^* and an optimal policy π^*. However, for the problem described here as the propane inventory problem, this is clearly impractical. The state space $\mathcal{I}$ is much too big (uncountable). The dimensionality is too high. The subproblems which need to be solved are NP-hard, etc. See the detailed arguments in Kleywegt et al. (2002, 2004). On the surface, this modelling approach seems to lead nowhere. Still, as a mathematical model it has the ability to represent the intrinsic problem details in a clear manner. Minkoff (1993) and Kleywegt et al. (2002, 2004) both attempted the ambitious undertaking of "salvaging" this Markov decision process approach to obtain reasonable solutions for inventory routing. (See also Berman and Larson (2001), for a different modelling approach.) In essence, Minkoff (1993) and Kleywegt et al. (2002, 2004), solution approach partitions the set of customers and estimates parameters for each subset by simulation. The optimal value function ν^* is approximated by $\hat{\nu}$ by choosing a collection of subsets (of size 1 or 2) of customers that partition the customer set. The approximate function $\hat{\nu}$ is computed for each subset and the sum over the subsets constitutes the approximate value. To simplify matters, Kleywegt et al. (2002, 2004) discretize their inventory demand state space. In all fairness, in Kleywegt et al. the focus is on designing vehicle routes which are limited to one or at most two customers, and the customers stockouts are due to lack of available vehicles. Since in our experience with propane delivery, vehicle availability was never the reason for stockouts, we do not consider these limitations here. There are however a number of questions regarding the modelling and solution methodology of Kleywegt et al. (2002, 2004). In addition, it is not clear how Kleywegt et al. solutions compare with real-life inventory routing since they did not conduct computational study which compares their results with real world data. However, many of the questions/reservations regarding their model are subsequently addressed in the work of Adelman (2003a, 2004) which we describe next.

5.2 Price-directed Markov models

Adelman (2004) states his decision model clearly: "The dispatcher chooses nonnegative integer-valued replenishment quantities $q =< q_1, q_2, \dots, q_{\bar{m}} >$ with q_i equal to the quantity replenished at i, $i = 1, \dots, \bar{m}$." For simplicity, we can adopt this notion of deciding the propane replenishment quantities a priori regardless of the amount I_{it} realized at time t in location i and the actual demand at i (remember that we do not know the exact inventory levels before service and therefore do not know how much is needed to fill the tank at i). The state space is as before the product space of estimated (and known) inventory levels $\mathcal{I}$. After estimating an inventory state $I \in \mathcal{I}$ the dispatcher selects the subset of customers who will be replenished in the current period. As before (for instance, Dror and Ball, 1987), it is assumed that no customer will be replenished more than once in a period. The customers are partitioned into non-empty (disjoint) subsets $M = \{M_1 \cup \dots \cup M_K\}$, where K is the number of subsets ($K \leq \bar{m}$) including the subset of customers who are not to be replenished in the current period, say M_K. Note that K and the particular partition are part of the action $a \in A(I)$. The idea is that the customers in each subset M_j, $j = 1, \dots, K - 1$ are replenished together (the same vehicle trip) in the current period (in Adelman, 2004, a period is a day). Based on the present state I, the corresponding action space $A(I)$ consists of determining the partition number K, the partition $M = \{M_1 \cup \dots \cup M_K\}$, and the vector q. The vehicle capacity constraints specify that $\sum_{j \in M_i} q_j \leq Q$, $i = 1, \dots, K - 1$. In addition, the components of the replenishment vector q as a function of the state I have to confirm to the customer tank constraints. That is, $q_i(I) \leq \max\{0, T_i - I_i\}$, $i = 1, \dots, \bar{m}$. In fact, one can replace the actual tank capacities T_i with artificial tank capacities T_i', or vehicle capacity Q with artificial vehicle capacity $Q' < Q$, as in a chance constrained models, to control the route failure probability for each subset of customers (see Trudeau and Dror, 1992).

After executing action $a \in A(I_t)$, the system observes (on delivery) a partial realization of demand. That is, the system observes only the demand quantities of the customers who were replenished at period t (day t). However, Adelman (2004) like Kleywegt et al. (2002, 2004) models the MDP as if the entire vector of demands $d = \langle d_1, \ldots, d_{\bar{m}} \rangle$ is observed after the decision a is taken. Here, we follow their modelling approach with respect to the probability distribution of the demand vector. That is, let $\eta(d)$ denote the probability that the demand equal d where $d_i \in D_i$, and D_i is a finite set of nonnegative integers.

Once the demand is realized, costs are computed. That is, given an action $a \in A(I_t)$, we obtain a partition of M as $M = M_1(a) \cup \cdots \cup M_{K(a)}(a)$ and the corresponding cost equal to $\sum_{i=1}^{K(a)-1} C_i\big(M_i(a)\big)$, where the cost for replenishing a given subset $M_i(a)$ is $C_i\big(M_i(a)\big)$ — the cost of the replenishment route (a TSP route) through M_i. Clearly, if our convention is that the subset $M_{K(a)}(a)$ does not get replenished in the current period, then the cost $C_{K(a)}\big(M_{K(a)}(a)\big) \equiv 0$. In addition to the delivery (routing) cost, Adelman (2004) also uses a traditional linear form to account for inventory holding and shortage costs in each location in the form of $g_i(I_i, q_i, d_i) = h_i(I_i + q_i - d_i)^+ + b_i\big(d_i - (I_i + q_i)\big)^+$.

Adelman (2004) derives an infinite horizon, expected discounted cost MDP which requires the dispatcher to find an optimal expected cost minimizing policy. After deriving the optimality equations (following Puterman, 1994) for finding an optimal policy that is Markovian, stationary, and deterministic, a linear program is proposed to solve the problem. Again, because of the huge size of the subsequent model, approximation solution schemes must be proposed. That is, the optimality equations are:

$$\nu^*(I) = \min_{a \in A(I)} \left\{ \sum_{i \in M} g_i\big(I_i, q_i(a)\big) + \sum_{j=1}^{K(a)-1} C_j\big(M_j(a)\big) + \alpha \sum_{I' \in \mathcal{I}} p(I'|I,a)\nu^*(I') \right\}, \quad \forall\, I \in \mathcal{I}$$

where $g_i\big(I_i, q_i(a)\big)$ is the expected holding and stockout cost for item i given the current state I_i and $q_i(a)$ is replenished. The linear program is:

$$\mathrm{LP}_0 = \max_{\nu} \sum_{I \in \mathcal{I}} s(I)\nu(I)$$

$$\nu(I) \le \sum_{i \in M} g_i\big(I_i, q_i(a)\big) + \sum_{j=1}^{K(a)} C_j(M_j(a)) + \alpha \sum_{I' \in \mathcal{I}} p(I' \mid I, a)\nu^*(I'), \quad \forall\, a \in A(I); I \in \mathcal{I}$$

where $s(I) > 0$ can be arbitrary positive constants for all $I \in \mathcal{I}$.

Substituting $\nu(I)$ by the sum of customer dependent value functions $V_i(I)$, that is, $\nu(I) = \sum_{i \in M} V_i(I), \forall\, I \in \mathcal{I}$ we can rewrite the above linear program.

An important modelling novelty introduced in Adelman's (2004) math programming based solution scheme, is his approximation of $C_i(M_i)$ from below with $\sum_{j \in M_i} W_j(q_j)$, where $W_j(q_j)$ represents the allocated cost of replenishing customer j with quantity q_j. Without recasting the full analysis of Adelman (2004), we note that the inventory replenishment solution uses the $W_j(q_j)$ in a similar role to the customer specific b_j value in Dror and Ball (1987). The optimal $W_j^*(q_j)$ values have to satisfy cost allocation efficiency conditions. That is, $\sum_{j \in M_i} W_j^*(q_j) = C(M_i)$ = the cost of the TSP tour through the subset of customers M_i including the depot. Thus, one approximating model proposed by Adelman (2004) is

$$\mathrm{LP}_{\mathrm{app}} = \max_{V,W} \sum_{i \in M} \sum_{I_i \in \mathcal{I}} s(I_i) V_i(I_i)$$

$$V_i(I_i) \leq g_i(I_i, q_i) + W_i(q_i) + \alpha \sum_{I_i' \in \mathcal{I}} p_i(I_i' \mid I_i, q_i) V_i(I_i'), \qquad \forall\, q_i \in \Upsilon(I_i), I \in \mathcal{I}$$

$$\sum_{i \in M'} W_i(q_i) \leq C(M'), \quad \forall M' \in M, q \in \Upsilon(M')$$

where $\Upsilon(M') = \{q_i, i \in M' : \sum_{i \in M'} q_i \leq Q \cdot \mathrm{NV}\}, \forall M' \subset M$, and $C(M')$ is the cost on an optimal VRP solution.

Adelman (2004) has shown that $\mathrm{LP}_{\mathrm{app}}$ gives the same results as forcing separable V in LP_0, but LP_{app} is much easier to solve. The optimal vector W^* of $W_i^*(q_i)$'s is coupled with the optimal vector V^*. When the optimal prices $V_i^*(I_i)$ are used to obtain the control solution then Adelman calls it a *price-directed* control policy. We note that our definition of $\Upsilon(M')$ is different than that in Adelman (2004) since the cost function $C(\cdot)$ must also depend on the full vector q, because it is now the solution to a VRP instead of a TSP. However, we believe that the math goes through in this case if we ignore the travel time. There are a number of key technical details in Adelman (2004) which we omit here for the sake of space. Incidently, in Adelman (2003b), Proposition 2 shows that when $C(M')$ is the cost of the optimal VRP solution, it can be decomposed into individual TSP solutions for the purpose of solving the relaxed LP. Based on the computational results, this (Adelman, 2004) solution methodology is proven superior to that of Minkoff (1993) and Kleywegt et al. (2002).

5.3 Cost allocation for subsets and inventory

In Adelman (2003a) the "price-directed" solution methodology for inventory routing receives an additional boost in terms of clarification of ideas, solution philosophy, and results. However, we should note that this paper looks for optimal policies in a deterministic setting like the one in the example described in Figure 1. The key concept in this paper is that of incremental cost when considering, in current time, the replenishment for customer i. That is, the key value which "real-world" dispatcher ought to examine is $C(M_j \cup i) - C(M_j), M_j \subset M, i \notin M_j$, together with the future cost implication of delivering quantity $d_i > 0$ to i.

Since all the costs have to be absorbed by the customers, Adelman's (2003a) analysis requires a cost allocation process which is applied simultaneously to routing and inventory replenishment decisions. (For cost allocation in vehicle routing see Gothe-Lundgren et al., 1996.) The propane delivery problem is formulated as that of minimizing long-run time average replenishment costs. This objective corresponds nicely to the objective of maximizing the long-run average number of units (gallons) delivered per hour of delivery operation which is used in real-life propane distribution. Adelman (2003a) formulates the problem as a control problem using dynamic system equations. Without restating the evolution of the problem modelling and the technical details involved, we note that the main thrust is to reformulate the deterministic control problem as a nonlinear program in which "in the long-run averages, replenishment must equal consumption." Solving the nonlinear program leads to the development of what is called the *price-directed* operating policy which maximizes the net-value of the replenishment. Incidently, Adelman proved that the objective used by Dror and Ball (1987), is also a net-value replenishment maximizing objective justifying its apparent success. Next, we sketch out Adelman's (2003a) modelling approach. We attempt to keep the notational convention of the earlier sections.

Adelman links the initiation of a replenishment action to any subset of customers with an occurrence of one stockout (or more than one, if occurring simultaneously) in the system which triggers a "must" replenishment response. (This is not how propane replenishment systems behave in practice, but it is quite fitting in this setting.) Now the time is measured not in day units, but as the elapsed time between the successive initiations of new replenishment activities. Thus, $\overline{T}_t$ represents the time elapsed between replenishment epochs t and $t+1$, $t = 1, 2, 3, \ldots$, and I_{it} be the inventory level at customer i just before the tth replenishment operations activation. We require that at least one day lapses between consecutive replenishment activities. Given a set M of customers, let $\widehat{M}_{\hat{A}} \subseteq M$ denote a subset of customers, and let the zero-one variable $Z_{\widehat{M}_{\hat{A}},t} = 1$, if customers in $\widehat{M}_{\hat{A}}$ are replenished during epoch t. The corresponding control problem is formulated as follows:

(CONTROL)

$$\inf \lim_{N\to\infty} \sup \frac{\sum_{t=1}^{N}\sum_{\widehat{M}_{\hat{A}}\subseteq M} C(\widehat{M}_{\hat{A}}) Z_{\widehat{M}_{\hat{A}},t}}{\sum_{t=1}^{N}\bar{T}_t} \tag{10.18}$$

$$I_{i,t+1} = I_{i,t} + d_{i,t} - \mu_i \bar{T}_t, \qquad \forall \text{ positive integer } t; \forall i \in M \tag{10.19}$$

$$d_{i,t} \le (T_i - I_{i,t}) \cdot \sum_{\widehat{M}_{\hat{A}}\subseteq M : i \in \widehat{M}_{\hat{A}}} Z_{\widehat{M}_{\hat{A}},t}, \qquad \forall \text{ positive integer } t; \forall i \in M \tag{10.20}$$

$$\sum_{i \in \widehat{M}_{\hat{A}}} d_{i,t} \le Q \cdot \mathrm{NV}, \qquad \forall \text{ positive integer } t \tag{10.21}$$

$$\sum_{\widehat{M}_{\hat{A}}\subseteq M} Z_{\widehat{M}_{\hat{A}},t} = 1, \qquad \forall \text{ positive integer } t \tag{10.22}$$

$$Z_{\widehat{M}_{\hat{A}},t} \in \{0,1\}, \qquad \forall \widehat{M}_{\hat{A}} \subseteq M, \forall \text{ positive integer } t \tag{10.23}$$

$$s, I, \bar{T} \ge 0 \tag{10.24}$$

The objective (10.18) minimizes the long-run average replenishment costs. Note that $C(\widehat{M}_{\hat{A}})$ denotes the cost of the corresponding VRP solution through the subset $\widehat{M}_{\hat{A}}$. Constraints (10.19)) state the conservation of inventory for each customer i. Constraints (10.20) insure that the individual tank capacities are respected. Constraints (10.21) make sure that for the replenishments scheduled in an epoch the vehicle fleet capacity is not exceeded. Constraints (10.22) state that exactly one subset is selected for replenishment in each replenishment epoch. The other constraints are just state $0-1$ selection for subsets and nonnegativity of the corresponding vectors. We note that it is straight forward in this formulation to limit the choice of the subsets $\widehat{M}_{\hat{A}}$ which can be considered for replenishment and thus manage the size of the corresponding control problem.

In order to solve the above problem, a nonlinear programming model is proposed which is a relaxation of the original. Denote by $Z_{\widehat{M}_{\hat{A}}}$ a nonnegative decision variable representing the long-run time average rate that the subset $\widehat{M}_{\hat{A}}$ is replenished together. For each such subset $\widehat{M}_{\hat{A}}$ which contains i, let $d_{i,\widehat{M}_{\hat{A}}}$ denote the decision variable representing the average replenishment quantity delivered to i when replenishing the subset $\widehat{M}_{\hat{A}}$. The corresponding program is stated as:

$$\textbf{(NLP)}\qquad \min \sum_{\widehat{M}_{\hat{A}} \subseteq M} C(\widehat{M}_{\hat{A}}) Z_{\widehat{M}_{\hat{A}}} \tag{10.25}$$

$$\sum_{\widehat{M}_{\hat{A}} \subseteq M; i \in \widehat{M}_{\hat{A}}} d_{i,\widehat{M}_{\hat{A}}} Z_{\widehat{M}_{\hat{A}}} = \mu_i, \qquad \forall i \in M \tag{10.26}$$

$$\sum_{i \in \widehat{M}_{\hat{A}}} d_{i,\widehat{M}_{\hat{A}}} \leq Q \cdot \mathrm{NV}, \qquad \forall \widehat{M}_{\hat{A}} \subseteq M \tag{10.27}$$

$$d_{i,\widehat{M}_{\hat{A}}} \leq T_i, \qquad \forall \widehat{M}_{\hat{A}} \subseteq M, i \in \widehat{M}_{\hat{A}} \tag{10.28}$$

$$Z, d \geq 0 \tag{10.29}$$

Adelman (2003a) shows that solutions to (NLP) may not be necessarily implementable because it does not capture all the dynamics in the system.

Next, a dual problem to (NLP) is formulated below with decision variables V_i and data d_i derived from a set $\mathcal{D}_O$.

$$\textbf{(D)}\qquad \max \sum_{i \in M} \mu_i V_i \tag{10.30}$$

$$\sum_{i \in \widehat{M}_{\hat{A}}} d_i V_i \leq C(\widehat{M}_{\hat{A}}), \quad \forall\ (\widehat{M}_{\hat{A}}, d) \in \mathcal{D}_O \tag{10.31}$$

The interpretation of the V_i's is that "at optimality they are the marginal costs, or prices, associated with satisfying constraints (10.26) of (NLP)" and "$\mu_i V_i$ at optimality can be interpreted as the total allocated cost rate for replenishing customer i in the optimal solution to (NLP)."

As far as solution, Adelman (2003a) solves the (NLP) by solving a version of the dual problem (D) by column generation procedure. We do not describe here the technical details and the difficulties involved. With this solution scheme he can prove that the solution for the motivating example (Figure 10.1) is indeed optimal!

The computational study of Adelman's (2003a) price-directed solution methodology demonstrates its viability. It produces results superior to all the previously proposed solution schemes.

6. Summary

This chapter is about solving the problem of propane deliveries. It is commonly viewed as a representative problem of a much larger family of hard problems of considerable practical significance. This problem has been on the "front burner" of the logistics academic and practitioners community for over twenty years. In fact, it was voted as the "most important/interesting" current OR problem in an unofficial gathering of Operations Research professionals which took place in early 1983 at Cornell University. Has it been solved now?

Clearly, Bell et al. (1983) describe a workable solution to the problem. They were able to construct a mathematical optimization module which routinely solved mixed integer programs with 800,000 variables and 200,000 constraints. In 2004, with our present computing power, this model should be able to solve problems 100 times larger. Could they prove optimality of their solution for the example in Figure 10.1? I do not think so. This problem was solved by Adelman (2003a).

In another solution scheme, Dror and Ball (1987) proposed and implemented a solution methodology which routinely solved problems with 5,000 customers. While it was (and may still be) a very promising solution methodology, it did not claim or deliver optimal solutions.

Presently, the work of Adelman (2003a,b, 2004) stands out as the reigning incumbent. Adelman (2003a), describes computational testing of his approach on a number of instances form Praxair, Inc. available online. These computational results are very promising. We would hope that more testing on "real-world" problems and operational implementation would follow.

Final Note. Propane deliveries are made every day in the US, Canada, and Europe (to my knowledge). The real-world operators are making money replenishing customers with propane and are in the market for improved solution methodologies for their operations.

Acknowledgment The author is grateful to Dan Adelman from The University of Chicago, Graduate School of Business, for many comments and constructive critique of this chapter. I claim all the remaining mistakes.

References

Adelman, D. (2003a). Price-directed replenishment of subsets: Methodology and its application to inventory routing. *MSOM*, 5(4):348 - 371.

Adelman, D. (2003b). Internal Transfer Pricing for a Decentralized Operation with a Shared Supplier. Working paper, October 2003, The University of Chicago, Graduate School of Business.

Adelman, D. (2004). A price-directed approach to stochastic inventory/routing. Forthcoming in*Operations Research.*

Bard, J.F., Huang,L., Jaillet,P., and Dror, M. (1998). A decomposition approach to the inventory routing problem with satellite facilities. *Transportation Science*,

32:189 – 203.

Bell, E.T., Dalberto, L.M., Fisher, M.L., Greenfield, A., Jaikumar, R., Kedia, P., and Prutzman, P. (1983). Improving the distribution of industrial gases with on-line computerized routing and scheduling optimizer. *Interfaces*, 3(6):4 – 23.

Berman, O. and Larson, R.C. (2001). Deliveries in an inventory/routing problem using stochastic dynamic programming. *Transportation Science*, 35:192 – 213.

Bertsekas, D.P. and Shreve, S.E. (1978). *Stochastic Optimal Control: The Discrete Time Case.* Academic Press, New York.

Dror, M. (1983). *The Inventory Routing Problem.* Ph.D. Thesis, University of Maryland at College Park.

Dror, M. (2002). Routing with stochastic demands: A survey. In: M. Dror, P. L'Ecuyer, and F. Szydarovszky (eds.), *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pages 629 – 653, Kluwer Academic Publishers.

Dror, M. and Ball, M.O. (1987). Inventory/routing: Reduction from annual to a short period problem. *Naval Research Logistics*, 34:891 – 905.

Dror, M., Ball, M.O., and Golden, B. (1985/6). Computational comparison of algorithms for inventory routing. *Annals of Operations Research*, 4:3 – 23.

Dror, M. and Trudeau, P. (1988). Inventory routing: Operational design. *Journal of Business Logistics*, 9:165 – 183.

Dror, M. and Trudeau, P. (1996). Cash flow optimization in delivery scheduling. *European Journal of Operational Research*, 88:504 – 515.

Federgruen, A. and Zipkin, P. (1984). A combined vehicle routing and inventory allocation problem. *Operations Research*, 32:1019 – 1037.

Fisher, M., Greenfield, A., Jaikumar, R., and Kedia, P. (1982). *Real-Time Scheduling of Bulk Delivery Fleet: Practical Application and Lagrangian Relaxation.* Technical report, The Wharton School, University of Pennsylvania, Department of Decision Science.

Gothe-Lundgren, M., Jornsten, K., and Varbrand, P. (1996). On the nucleolus of the basic vehicle routing game. *Mathematical Programming*, 72:83 – 100.

Herron, D. (1979). Managing physical distribution for profit. *Harvard Business Review*, 79:121 – 132.

Jaillet, P., Huang, L., Bard, J.F, and Dror, M. (2002). Delivery cost approximations for inventory routing problems in a rolling horizon framework. *Transportation Science*, 36:292 – 300.

Kleywegt, A.J., Nori, V.S., and Savelsbergh, M.W. (2002). The stochastic inventory routing problem with direct deliveries. *Transportation Science*, 36:94 – 115.

Kleywegt, A.J., Nori, V.S., and Savelsbergh, M.W. (2004). Dynamic programming approximations for stochastic inventory routing problem, *Transportation Science*, 38:42 – 70.

Kreimer, J. and Dror, M. (1990). The monotonicity of threshold detection probability in stochastic accumulation processes. *Computers & Operations Research*, 17:63 – 71.

Minkoff, A.S. (1993). A Markov decision model and decomposition heuristics for dynamic vehicle dispatching. *Operations Research*, 41:77 – 90.

Puterman, M.L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley.

Trudeau, P. and Dror, M. (1992). Stochastic inventory routing: Stockout and route failure. *Transportation Science*, 26:172 – 184.

Chapter 11

SYNCHRONIZED PRODUCTION-DISTRIBUTION PLANNING IN THE PULP AND PAPER INDUSTRY

Alain Martel
Nafee Rizk
Sophie D'Amours
Hanen Bouchriha

Abstract This chapter examines the short term production, transportation and inventory planning problems encountered in the fine-paper industry. After positioning the problems in the context of a general supply chain planning system for the pulp and paper industry, a comprehensive synchronized production-distribution model is gradually developed. First, a model for the dynamic lot-sizing of intermediate products on a single paper machine with a predetermined production cycle is proposed. The model also plans the production and inventory of finished products. Then, we consider the lot-sizing of intermediate products on multiple parallel paper machines with a predetermined production sequence. Finally, simultaneous production and distribution planning for a single mill multiple distribution centers network is studied by considering different transportation modes between the mill and its Distribution Centers (DCs).

1. Introduction

The pulp and paper industry is one of the most important industries of Canada in terms of contribution to its balance of trade. In 2001, it represented 3% of Canada's Gross Domestic Product (FPAC, 2002). The expertise of the Canadian pulp and paper industry is well renowned. Over the years, the industry has been confronted with different market pressures. For example, currently, global production capacity is abundant due to major consolidations in the sector. Companies are working closely to integrate the different business units of their supply chain due to this consolidation. They are reengineering their supply chain, which means they are trying to define the optimal network structure and planning approach in order to maximize profit.

1.1 The pulp and paper supply chain

Total shipments within the industry supply chain in 2002 included pulp (10.5 million tons), newsprint (8.5 million tons), printing and writing paper (6.3 million tons) as well as other paper and paperboard (5.2 million tons). These products are produced and distributed in complex supply chains composed of harvesting, transformation, production, conversion and distribution units, as shown in Figure 11.1. The main components of the pulp and paper supply chains are their supply network, their manufacturing network, their distribution network and the product-markets targeted. Different companies in the world are structured in different ways. Some are vertically integrated: they possess and control all the facilities involved in this value creation chain, from woodlands to markets. Others are not integrated and they rely on outsourcing to fulfill part of their commitments to their customers. For example, some companies buy pulp on the market, produce the paper and convert it through a network of external converters, before distributing the final products. All these possibilities are illustrated in Figure 11.1. The links between the external network and the internal network define these outsourcing alternatives.

An important problem is therefore to determine the supply chain structure and capacity, to decide how and where intermediate and finished products should be manufactured and how they should be distributed. These decisions relate to the company's business model as well as to its strategic supply chain design. In the pulp and paper industry, these decisions are tightly linked to the availability of fiber and the supply of raw materials. For example, Canadian paper is made from 55% chips and sawmill residues, 20% recovered paper and 25% round wood. The quality of the paper produced depends directly on the quality of the fiber used. Therefore, designing the supply chain imposes a thorough analysis of the supply network. Also, the industry is very capital intensive. Even the modification of a single paper machine is a long-term investment project. A planning horizon of at least five years must be considered to evaluate such projects. The final output of this strategic decision process defines the supply chain network structure, that is its internal and external business units (woodlands, mills warehouses, etc.), their location, their capacity, their technology as well as the transporta-

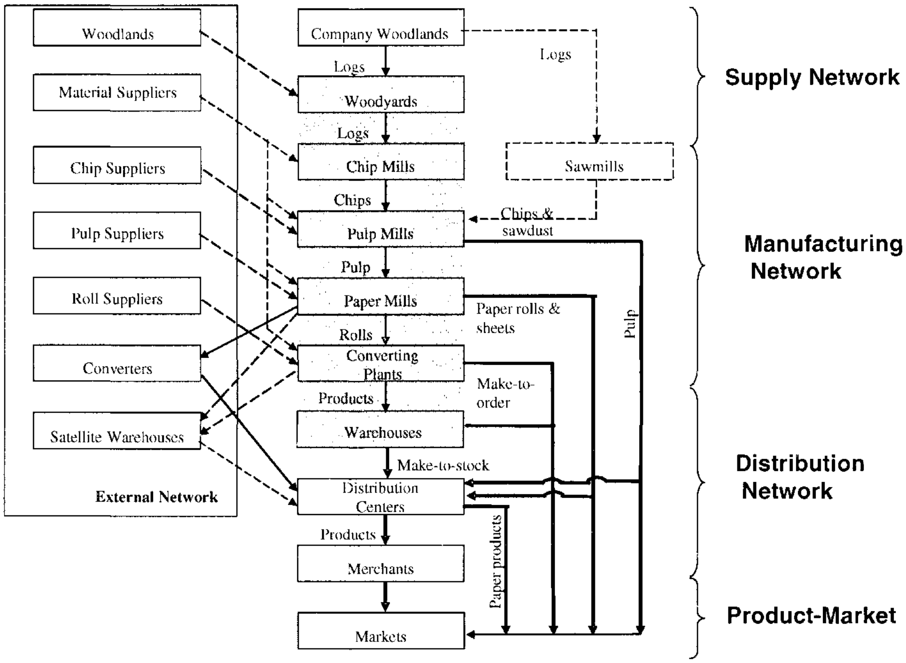


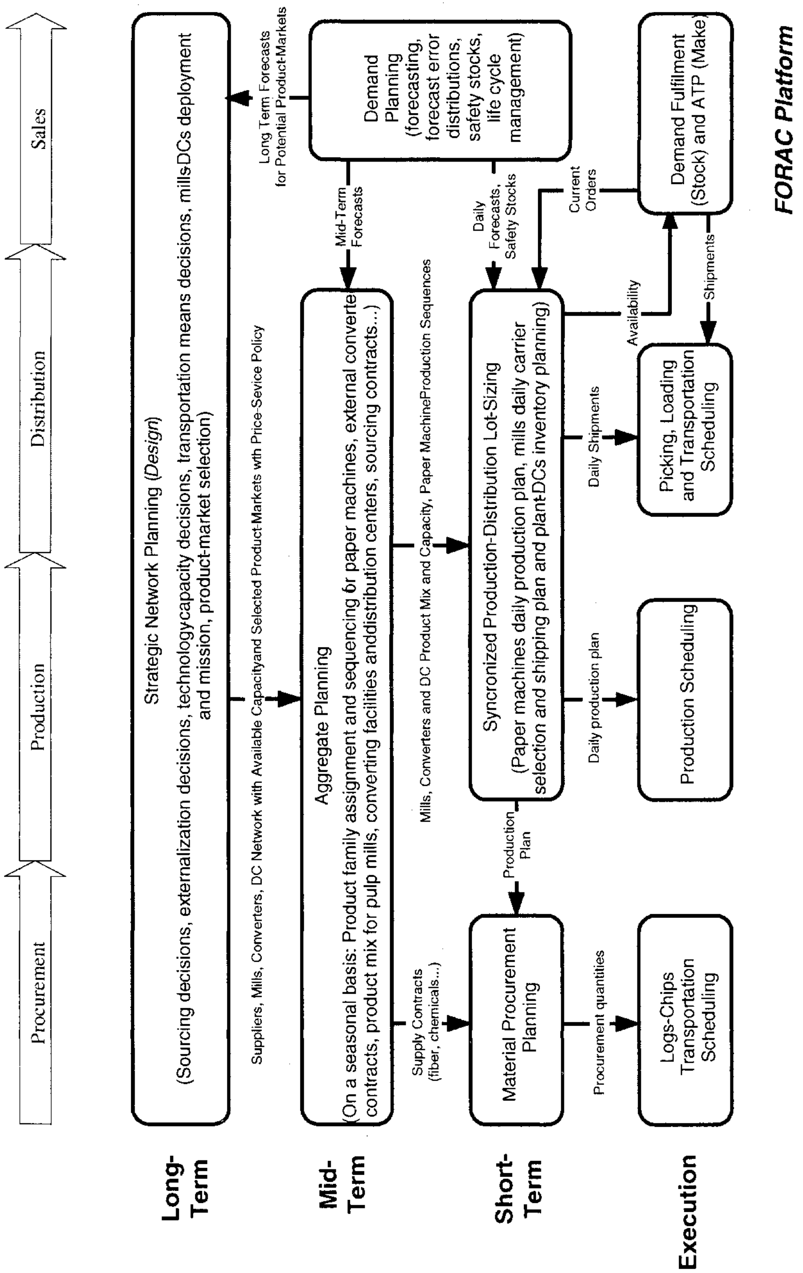
Figure 11.1. The pulp and paper supply chain

tion modes to favor. Large scale mixed-integer programming models can often be formulated to support this complex design process. In order to take the uncertainty of the future business environment into account, these models must be used in conjunction with a scenario planning approach. The application of such a modeling approach to capital budgeting problems at Fletcher Challenge Canada and Australasia are documented respectively in Everett, Philpott and Cook (2000) and in Everett, Aoude and Philpott (2001).

Once the structure of the supply chain is decided, managers need to plan supply, production and distribution over a rolling horizon. Usually this process is conducted in two phases: tactical planning and operational planning. Tacticai planning deals with resource allocation problems and it defines some of the rules-of-the-game to be used at the operational planning level. The tactical plans elaborated usually cover a one year horizon divided into enough planning periods to properly reflect seasonal effects. The rules-of-the-game relate to supply, production, distribution and transportation policies such as: customer service levels, safety stocks, the assignment of customers to warehouses or to mills, the selection of external converters, the size of parent rolls to manufacture, the assignment of paper grades to paper machines and the determination of their production sequence, sourcing decisions for the mills, etc. These tactical decisions frame the operational planning decisions by identifying operational targets and constraints. They are made to convey an integrated view of the supply chain without having to plan all activities for all business units within a central planning engine. Again, mathematical programming models can often be used to support tactical planning decisions. Philpott and Everett (2001) present the development of such a model for Fletcher Challenge Paper Australasia.

At the operational planning level, managers are really tackling material, resource and activity synchronization problems. They have to prepare short-term supply, production and distribution plans. Usually at this planning level, information is no longer aggregated and the planning horizon considered covers a few months divided into daily planning periods. The plans obtained are usually sufficiently detailed to be converted into real-time execution instructions without great difficulty. The procurement, lot-sizing, scheduling and shipping plans made at this level are based on trade-offs between set-up costs, production and trim loss costs, inventory holding costs and transportation economy of scales, and they take into consideration production and delivery lead times, capacity, etc. The objective pursued at the operational planning level is usually to minimize operating costs while meeting targeted service levels and resource availability constraints. Mathematical programming models can often be used to support operational planning decisions. Everett and Philpott (2002) describe a mixed integer programming model for scheduling mechanical pulp production with uncertain electricity prices. Bredström et al. (2003) present an operational planning model for a network of pulp mills. Keskinocak et al. (2002) propose a production scheduling system for make-to-order paper companies. An integrated diagram of the system of strategic, tactical and operational planning decisions required to manage the pulp and paper supply chain is provided in Figure 11.2.

In an integrated pulp and paper plant, the production process can be decomposed in four main stages. The first stage (the chip mill) transforms logs into chips. The second stage (the pulp mill) transforms chips and chemicals into pulp. The third stage (the paper mill) transforms pulp into paper rolls. The paper mill is usually composed of a set of parallel paper machines. Finally, the last stage (convention mill) converts paper rolls into the smaller rolls or sheets which are demanded by external customers. Figure 11.3 illustrates the material flow within an integrated pulp and paper mill. As can be noted, some production stages can be partially or completely bypassed through external provi-


Procurement
Production
Distribution
Sales
Long-Term
Mid-Term
Short-Term
Execution
Strategic Network Planning (Design)
(Sourcing decisions, externalization decisions, technologycapacity decisions, transportation means decisions, millsDCs deployment and mission, product-market selection)
Long Term Forecasts for Potential Product-Markets
Suppliers, Mills, Converters, DC Network with Available Capacityand Selected Product-Markets with Price-Sevice Policy
Demand Planning (forecasting, forecast error distributions, safety stocks, life cycle management)
Mid-Term Forecasts
Aggregate Planning
(On a seasonal basis: Product family assignment and sequencing on paper machines, external converter contracts, product mix for pulp mills, converting facilities anddistribution centers, sourcing contracts...)
Mills, Converters and DC Product Mix and Capacity, Paper MachineProduction Sequences
Daily Forecasts, Safety Stocks
Current Orders
Supply Contracts (fiber, chemicals...)
Production Plan
Material Procurement Planning
Syncronized Production-Distribution Lot-Sizing
(Paper machines daily production plan, mills daily carrier selection and shipping plan and plant-DCs inventory planning)
Availability
Demand Fulfilment (Stock) and ATP (Make)
Shipments
Procurement quantities
Daily production plan
Daily Shipments
Logs-Chips Transportation Scheduling
Production Scheduling
Picking, Loading and Transportation Scheduling
FORAC Platform

Sophie D'Amours, Alain Martel and Alain Rousseau, FORAC, May 2003

Figure 11.2. Supply chain decision processes in the pulp and paper industry

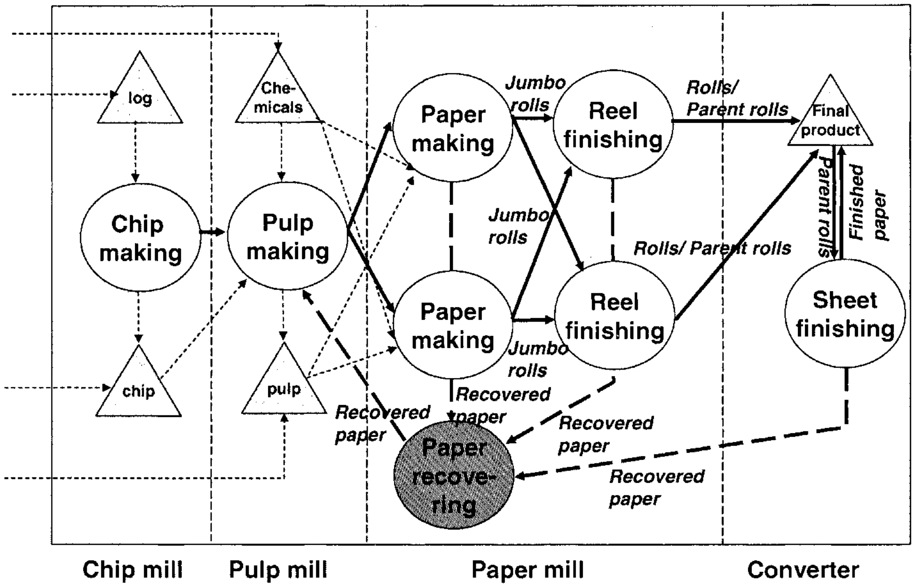


Figure 11.3. Processes and material flows in an integrated pulp and paper mill

sioning of intermediate products (chips and/or pulp). Also, although, some paper is lost during the paper making, reel finishing and sheet finishing operations, it is recovered and fed back into the pulp production process.

The planning challenge is to synchronize the material flow as it moves through the different production stages, to meet customer demand and to minimize operations costs. The paper machine is often the bottleneck of this production system and this is why production plans are usually defined in terms of this bottleneck.

The problem we focus on in this chapter is the synchronized production-distribution planning problem for a single mill and the set of distribution centers it replenishes for a make to stock and to order paper company. It is addressed gradually, starting by current industrial practices where production and distribution are planned independently and moving toward the integration of production and distribution decisions. Under the first paradigm two business contexts have attracted our attention. The first one refers to production planning on a single machine constrained by a production cycle, within which all different products are produced in a pre-defined sequence. It is referred to as the *Single-Machine Lot-Sizing Model.* The second context considered relates to production planning on several parallel machines each constrained by a pre-defined production sequence. It is referred to as the

Multiple-Machine Lot-Sizing Model. Finally, distribution considerations are introduced in the last part of this paper and the problem is set in its complete form as the *Synchronized Production-Distribution Planning Model.* Harvesting decisions and pulp making planning decisions are not taken into consideration in this chapter as they are in Bredström et al. (2001). Their work, however, fits within the planning paradigm presented in this chapter.

1.2 Production and distribution planning problems

In what follows, we concentrate our attention on the short term production and distribution planning problems encountered in the fine-paper industry. The specific context considered is illustrated in Figure 11.4 (see Tables 11.1 and 11.2 for notations). In this industry, some products are made-to-stock, others are made-to-order and others are shipped to external converting plants. Although the demand for products is partly planned and partly random, we assume, as is customary in ERP and APS systems, that it is deterministic and time-varying (dynamic). This demand is based on orders received and on forecasts, and we assume that the safety stocks required as protection against the randomness of demand are determined exogenously, prior to the solution of our problem. In order to provide a competitive service level, the make-to-stock products must be stored in distribution centers (DCs) which are close to the market. Part of the company demand is therefore fulfilled from these DCs. Make-to-order demand, converter demand and local make-to-stock demand is however fulfilled directly from the mills.

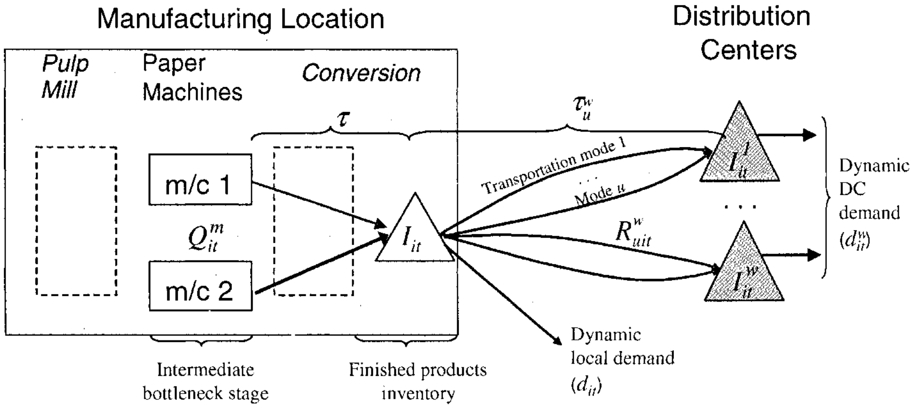


Figure 11.4. Paper industry production and distribution context

As indicated before, production at the mills involves multiple stages, with one of them, the paper machines, creating a bottleneck. Paper machines can run 24 hours a day during the whole year, but they can also be stopped (or slowed down) from time to time to adapt to low market demand or for maintenance purposes. In the bottleneck stage, a small number of *intermediate products* (IP) are manufactured by parallel paper machines, each machine producing a predetermined set of intermediate products with a *fixed* production sequence. A changeover time is required to change products on a paper machine, which means that capacity is lost when there is a production switch. In the succeeding stages, the intermediate products are transformed into a large number of finished products (FP). However, in the paper industry, any given finished product is made from a single intermediate product (divergent Bill-of-Material). The conversion operations can be done within a predetermined planned lead-time. We assume that no inventory of intermediate products is kept, but the finished products can be stocked at the plant before they are shipped.

Several transportation modes (mainly truck, train and intermodal), can be used to ship products from the plants to the warehouses. The transit time for a given origin-destination depends on the transportation mode used. For each mode, there are economies of scale in transportation costs, depending on the total loads shipped during a time period, independently of the type of finished products in the shipments.

Planning is done on a rolling horizon basis, with daily time buckets. Within this context, three different problems are examined in the next sections of the chapter:

(1) Single machine lot-sizing of intermediate and finished products with a predetermined IP production cycle.

(2) Single-mill multiple-machine lot-sizing of intermediate and finished products with a predetermined IP production sequence.

(3) Synchronized production-distribution planning for a single-mill multiple-DC subnetwork, with a predetermined IP production sequence.

The general notation used in the chapter is introduced in Tables 11.1 and 11.2. Additional notation specific to the three problems studied is defined in their respective sections.

1.3 Literature review

The three problems studied in the chapter relate to the multi-item capacitated dynamic lot-sizing literature. A recent survey of the lot-sizing literature covering these problems is found in Rizk and Martel (2001). Under the assumptions that there is a single production stage, that set-up costs and times are sequence independent and that capacity is constrained by a single resource, three formulations of the problem have been studied extensively: the *Capacitated Lot-Sizing Problem* (CLSP), the *Continuous Setup Lot-Sizing Problem* (CSLP) and the *Discrete Lot-Sizing and Scheduling Problem* (DLSP). The CLSP involves

Table 11.1. Indices, parameters and sets

T	Number of planning periods in the planning horizon.
t	A planning period ($t = 1, \ldots, T$).
IP	Set of intermediate products ($\{1, \ldots, N\}$).
FP	Set of finished products ($\{n+1, \ldots, N\}$).
i, i'	Product type indexes ($i, i' \in \mathrm{IP} \cup \mathrm{FP}$).
m	A paper machine for the production of IP ($m = 1, \ldots, M$).
M_i	The set of machines m that manufacture product i ($i \in \mathrm{IP}$).
IP_m	The set of intermediate products manufactured by machine $m(\mathrm{IP}_m \subset \mathrm{IP})$.
f_m	Number of intermediate products manufactured on machine m (i.e., $\lvert\mathrm{IP}_m\rvert$).
W	Set of distribution centers, $w \in W$.
U^w	Set of transportation modes available to ship products to DC w ($u \in U^w$).
τ	Planned production lead-time.
τ_u^w	Supply lead-time of DC w when transportation mode u is used.
d_{it}^w	Effective external demand for product i at DC w during period t.
d_{it}	Effective demand at the mill for product i during period t.
C_t^m	Production capacity of machine m in period t (in time units).
k_{it}^m	Changeover time required at the beginning of period t to produce $i \in \mathrm{IP}_m$ on machine m.
K_{it}^m	Product i changeover cost on machine m in period t ($i \in \mathrm{IP}_m$).
r_i	Transportation resource absorption rate for product i (in tons).
h_{it}^w	Inventory holding cost of product i at DC w in period t.
h_{it}	Inventory holding cost of product i at the mill in period t.
$g_{ii'}$	Number of product i units required to produce one unit of product i'.
a_{it}^m	Machine m capacity consumption rate of product $i \in \mathrm{IP}_m$ in period t.
SC_i	Set of finished products manufactured with intermediate product i ($\mathrm{SC}_i = \{i' \mid g_{ii'} > 0\}$).

Table 11.2. Decision variables

R_{it}	Quantity of finished product $i \in \mathrm{FP}$ added to the mill inventory for the beginning of period t.
Q_{it}^m	Quantity of intermediate product $i \in \mathrm{IP}$ produced with machine m during period t.
I_{it}	Inventory level of finished product $i \in \mathrm{FP}$ on hand in the mill at the end of period t.
I_{it}^w	Inventory level of finished product $i \in \mathrm{FP}$ on hand at DC w at the end of period t.
R_{uit}^w	Quantity of item i shipped by transportation mode u from the mill to DC w at the beginning of period t.

the elaboration of a production schedule for multiple items on a single machine over a planning horizon, in order to minimize total set-up, production and inventory costs. The main differences between the CLSP and the CSLP are that in the latter, at most one product is produced in a period and a changeover cost is incurred only in the periods where the production of a new item starts. In the CLSP, several products can be produced in each period and, for a given product, a set-up is necessary in each period that production takes place. For this reason, CLSP is considered as a large time bucket model and CSLP as a *small* time bucket model. DLSP is similar to CSLP in that it also assumes at most one item to be produced per period. The difference is that in DLSP, the quantity produced in each period is either zero or the full production capacity.

The first problem studied in this chapter can be considered as an extension of the CLSP to the case where the items manufactured include both intermediate products and finished products made from the IP products. When a predetermined fixed production cycle is used, production planning on a single paper machine reduces to such a problem. The length of the IP production cycle to use can be determined by first solving an *Economic Lot-Sizing and Scheduling Problem* (Elmaghraby, 1978; Boctor, 1985). Florian et al. (1980) and Bitran and Yanasse (1982) showed that CLSP is NP-hard even when there is a single product and Trigeiro et al. (1989) proved that when set-up times are considered, even finding a feasible solution is NP-hard. Exact *mixed integer programming* solution procedures to solve different versions of the problem were proposed by Barany et al. (1984), Gelders et al. (1986), Eppen and Martin (1987), Leung et al. (1989) and Diaby et al. (1992). Heuristic methods based on mathematical programming were proposed by Thizy and Wassenhove (1985), Trigeiro et al. (1989), Lasdon and Terjung (1971) and Solomon et al. (1993). Specialized heuristics were also proposed by Eisenhut (1975), Lambrecht and Vanderveken (1979), Dixon and Silver (1981), Dogramaci et al. (1981), Gunther (1987), and Maes and Van Wassenhove (1988).

When set-up costs are sequence dependent, the sequencing and lot-sizing problems must be considered simultaneously and the problem is more complex. This problem is known as *lot sizing and scheduling with sequence dependent set-up* and it has been studied by only a few authors (Haase, 1996; Haase and Kimms, 1996). Particular cases of the problem were also examined by Dilts and Ramsing (1989) and by Dobson (1992).

The second problem studied in this chapter can be considered as an extension of the CSLP to the case of several parallel machines with a predetermined production sequence, and with a two level (IP and FP) product structure. The multi-item CSLP has been studied by Karmarkar and Scharge (1985) who presented a Branch and Bound procedure based on Lagrangean relaxation to solve it. An extension to the basic CSLP that considers parallel machines was studied by De Matta and Guignard (1989) who proposed a heuristic solution method based on Lagrangean relaxation. The DLSP, which is also related to our second problem, has been studied mainly by Solomon (1991).

The third problem studied in this chapter is an extension of the second one involving the simultaneous planning of the production and distribution of several products. Coordinating flows in a one-origin multi-destination network has attracted the attention of some researchers (see Sarmiento and Nagi (1999), for a partial review). Most of the work done involves a distributor and its retailers and it considers a single product. Anily (1994), Gallego and Simchi-Levi (1990), Anily and Federgruen (1990, 1993), and Herer and Roundy (1997) tackle this problem in the case of a single product and deterministic static demand. In these papers, transportation costs are made up of a cost per mile plus a fixed charge for hiring a truck. The objective is to determine replenishment policies that specify the delivery quantities and the vehicle routes so as to minimize long-run average inventory and transportation costs. Viswanthan and Mathur (1997) generalized Anily and Federgruen (1990) with the multi-item version of the problem. Diaby and Martel (1993) and Chan et al. (2002) consider the single-item deterministic dynamic demand case with a general piece-wise linear transportation cost. Martel et al. (2002) consider the multi-item dynamic demand case with a general piece-wise linear transportation cost but they do not include production decisions in their model. To the best of our knowledge, the only models including production-distribution decisions for multi-item dynamic demands are Chandra and Fisher (1994), Haq et al. (1991) and Ishii et al. (1988).

2. Single-machine lot-sizing problem

2.1 Problem definition and assumptions

In order to reduce the complexity of the complete production-distribution problem defined in Figure 11.4, the current practice in most paper mills is to plan production for each paper machine separately. Furthermore, as indicated earlier, in order to simplify the planning problem and the implementation of the plans produced, the set of IP products to be manufactured on a given paper machine, the sequence in which the products must be manufactured and the length of the production cycles (in planning periods) to be used are predetermined (at the tactical planning level). The fixed sequence context also implies that the intermediate products are all manufactured in each cycle. However, as illustrated in Figure 11.3, the paper rolls (jumbo) coming out of the paper machines are not inventoried: they are transformed immediately into finished products. The finished products however are stored in the mill warehouse and it is from this stock that products are shipped, every planning period, to distribution centers or customers. In order to prepare adequate production plans, the relationships between the IP lot-sizes and the FP inventories and demands must be considered explicitly. Our aim in this section is to present a model to determine the lot-size of the IP to manufacture on a single paper machine which minimizes total relevant costs for all the production cycles in the planning horizon considered.

In order to relate the model proposed to the general problem, the timing conventions used must be clarified. Figure 11.5 illustrates the relationships between *planning periods, production cycles, production lead-times* and the *planning horizon.* As can be seen, a production cycle p, is defined by a set T_p of planning periods and there are P production cycles in the planning horizon. For the finished products, the planning horizon is offset by the production lead-time. This planned lead-time is assumed to be the same for all finished products and it is expressed in planning periods. It includes the total elapsed time from the beginning of the period in which an IP production order is released until the finished products are available to be shipped from the mill warehouse. In

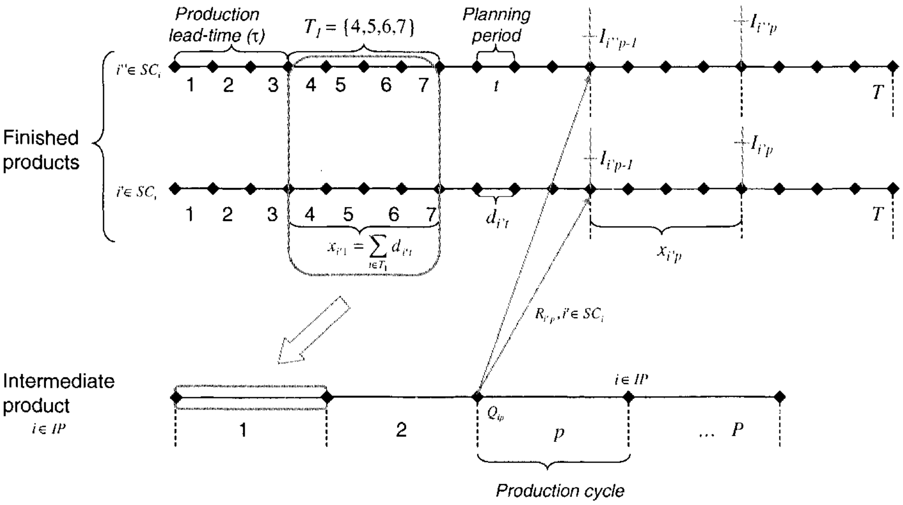


Figure 11.5. Planning horizon for IP and FP products

other words, it is assumed that the finished products made from the intermediate products produced in a production cycle will be available in inventory τ planning periods after the beginning of the cycle, independently of the position of the IP product in the predetermined machine production sequence. Clearly, this is a gross approximation and it is reasonable only when the production cycles are relatively short. This assumption, however, provides a rational for aggregating planning period effective demands into production cycles effective demands.

In what follows, we assume that planning is based on the finished products *effective demands*. Following Hax and Candea (1984), the effective demand d_{it} of a finished product $i \in \mathrm{FP}$ in planning period $t > \tau$ is defined as the demand for the period which cannot be covered by the projected inventory on hand $I_{i\tau}$ at the end of period τ, taking the desired safety stock level SS_i for the product into account. More precisely,

$$d_{it} = \begin{cases} \max\{0, \sum_{t'=\tau+1}^{t} \underline{d}_{it'} - I_{i\tau} + SS_i\}, & \text{if } d_{i,t-1} = 0, \\ \underline{d}_{it}, & \text{otherwise} \end{cases}, \qquad t = \tau + 1, \ldots, T \ (d_{i\tau} = 0),$$

where $\underline{d}_{it}$ is the demand for finished product i at the mill in planning period t. We also assume that, for each cycle p in the planning horizon, the cumulative capacity available is greater than or equal to the cumulative effective demand. When this condition is not satisfied, there is no feasible solution. We also assume that there is a lower bound on the production lot-size for each product made in a cycle. Finally, we assume that the unit production costs for an intermediate product are the same in every production cycle.

2.2 Single-machine lot-sizing model

Since this is a single-machine problem, the index m is dropped in what follows from the notation defined in Tables 11.1 and 11.2. The additional notations in Table 11.3, are also required to formulate our fixed cycle lot-sizing model.

In order to formulate the model, we first need to define the aggregate effective demand for the production cycles. As illustrated in Figure 11.5, the cycles' effective demands are given by:

$$x_{ip} = \sum_{t \in T_p} d_{it}, \quad i \in \mathrm{FP}, \ p = 1, \ldots, P.$$

Note next that one of the implications of using predetermined fixed production cycles is that every IP is manufactured during each cycle.

Table 11.3. Additional notation

P	Number of production cycles in the planning horizon.
p	A production cycle.
T_p	Set of planning periods in production cycle p.
$\tilde{C}_p$	Production capacity available in cycle p, net of set-up times (in time units).
$\underline{Q}_i$	Minimum lot-size for product $i \in$ IP (minimum hours/a_i).
x_{ip}	Product $i \in$ FP effective demand for production cycle p.

This implies that the total set-up costs over the planning horizon are constant and that they do not have to be taken into account explicitly.

In order to economize set-up times for the entire planning horizon, while maintaining the fixed sequence, as illustrated in Figure 11.6, we can impose that the last item scheduled at the end of a given cycle is scheduled at the beginning of the next cycle. The example in Figure 11.6 assumes that product 1 was the last product manufactured in cycle 0. Given this, the net capacity available in each cycle, $\tilde{C}_p$, $p = 1, \dots, P$, can be calculated *a priori*. For example, for cycle 2 in Figure 11.6, the net capacity available is $\tilde{C}_p = \sum_{t \in T_2} C_t - k_1 - k_2$, where k_i is the changeover time for product i. More generally, the capacity available can be calculated with the expression:

$$\tilde{C}_p = \sum_{t \in T_p} C_t - \sum_{i \neq \text{first}(p)} k_i,$$

where first(p) is the index of the product scheduled for production at the beginning of cycle p.

Also, since we have a deterministic demand and since the variable production costs do not change from cycle to cycle, the total production

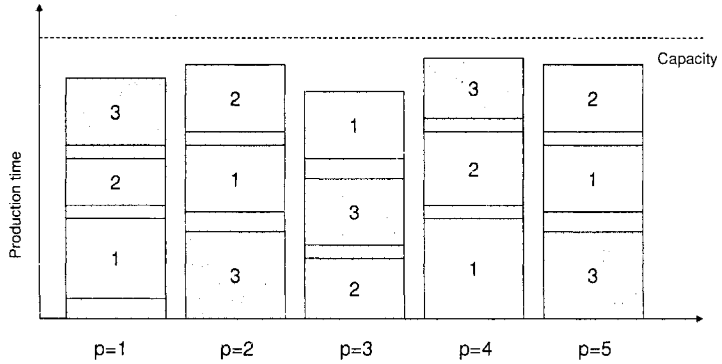


Figure 11.6. Example of a fixed cycle production plan for three products

cost is a constant and it does not have to be taken into account explicitly. Consequently, the only relevant costs under our assumptions are the intermediate products inventory holding costs. The lot-sizing problem to solve in order to minimize these costs is the following:

$$\text{Min} \sum_{p=1}^{P} \sum_{i \in \text{IP}} h_{ip} I_{ip} \tag{11.1}$$

subject to:

$$Q_{ip} - \sum_{i' \in \text{SC}_i} g_{ii'} R_{i'p} = 0 \qquad i \in P \tag{11.2}$$

$$R_{ip} + I_{i(p-1)} - I_{ip} = x_{ip} \qquad i \in \text{FP};\ p = 1, \ldots, P\ (I_{i0} = 0) \tag{11.3}$$

$$\sum_{i \in \text{IP}} a_{ip} Q_{ip} \leq \tilde{C}_p \qquad p = 1, \ldots, P \tag{11.4}$$

$$Q_{ip} \geq \underline{Q}_{ip} \qquad i \in \text{IP}; p = 1, \ldots, P \tag{11.5}$$

$$I_{ip} \geq 0 \qquad i \in \text{FP} \tag{11.6}$$

$$Q_{ip} \geq 0 \qquad i \in \text{IP};\ p = 1, \ldots, P \tag{11.7}$$

$$R_{ip} \geq 0 \qquad i \in \text{FP};\ p = 1, \ldots, P \tag{11.8}$$

The constraints include product bills of material (11.2), inventory accounting equations for finished products (11.3), and production output capacity of the machine (11.4), taking into account set-up times incurred for each cycle. Constraints imposing a minimum production quantity for each cycle (11.5) were also included. These constraints guarantee that each product can be manufactured in each cycle according to the predetermined sequence. Backorders are not allowed (11.6). Finally, nonnegativity constraints for production variables are also included (11.7) and (11.8). The experimental evaluation of the impact of this model and its various parameters are discussed is Bouchriha, D'Amours, and Ouhimmou (2003).

Although it is common practice in the fine paper industry to prepare fixed cycle length production plans for the paper machines and to use all the capacity available (i.e., to replace the inequality by an equality in constraint (11.4)), it is clear that the planning approach, developed in this section, is not really satisfactory. When the market demand for paper is low, as it currently is, the approach may lead to the production of products which are not required or to high inventory levels which could be avoided. It is therefore clear that this approach is suboptimal. Moreover, depending on the cycle length used and the demand variability, the approach could even lead to unfeasible solutions. For all these reasons, in the following sections, the fixed cycle length assumption is relaxed but the assumption that the production sequence is predetermined is maintained.

3. Multiple-machines lot-sizing problem

3.1 Problem definition and assumptions

In this section, we consider the simultaneous planning of the lot-sizes of intermediate products on all the paper machines in a mill, as well as the production and inventory planning of its finished products. We assume that the paper machines are capacity constrained but that the conversion stages are not capacity constrained. This is realistic, since it is always possible to subcontract part of the finishing operations if additional capacity is required. Although it is important in the industry to preserve the predetermined production sequence on the paper machines (launching production according to increasing paper thickness minimizes paper waste and set-up times), the use of fixed length production cycles is not imposed by any technological constraints. In this section we therefore relax the assumption of a fixed length production cycle. However, we assume that at most one production changeover is allowed per paper machine per planning period. This is reasonable provided that the planning periods used are relatively short (a day or a shift). We also assume that it is not necessary to use the total capacity available in a given period. Although in practice this is rarely the case, it is possible to reduce the production paste in order to produce less during a planning period without stopping the machine. The approach proposed in this section and the next is based on Rizk, Martel, and D'Amours (2003).

Let $g_{ii'}$ be the number of units of IP i required to produce one unit of FP i', taking any waste incurred in the transformation process into account. Since each FP is made from a single IP product, the set of FP can be partitioned according to the IP it is made of. In addition, it is assumed that a standard production sequence of IP must be maintained for each machine $m = 1, \ldots, M$, and that at most one product type can be produced in a given time period. Let e_m denote the index of the IP in the eth position in machine m production sequence, so that $e_m = 1_m, \ldots, f_m$, where f_m represents the product in the final position in machine m production sequence. Thus, when $e < f$ product $(e+1)_m$ can be produced on machine m, only after product e_m has finished its production batch (see Figure 11.7). The production resource consumption for intermediate products is assumed to be concave, that is, a fixed resource capacity consumption is incurred whenever production switches from one IP to another (changeover time), and linear resource consump-

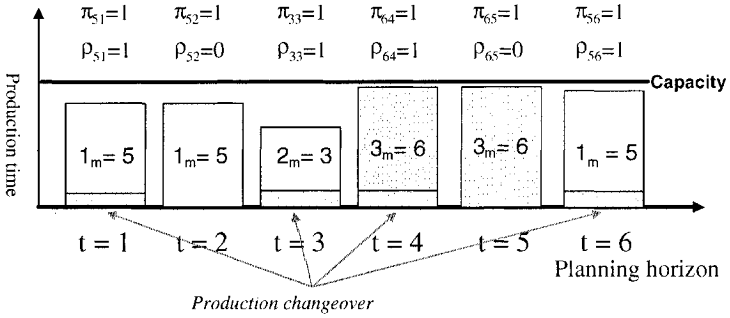


Figure 11.7. Example of a production plan for machine m

tion is incurred during the production of a batch of IP. Inventory holding costs are assumed to be linear.

3.2 Multiple-machines lot-sizing model

Using the notation in Tables 11.1, 11.2 and 11.4, the production planning problem of the manufacturing plant can be represented by the following optimization model:

$$\text{Min} \sum_{t=1}^{T}\left[\sum_{m=1}^{M}\sum_{i\in \text{IP}_m} K_{it}^{m}\rho_{it}^{m}\right] + \sum_{t=\tau+1}^{T+\tau}\left[\sum_{i=n+1}^{N} h_{it}I_{it}\right] \tag{11.9}$$

subject to

$$\sum_{\eta\in M_i} Q_{it}^{m} - \sum_{i'\in \text{SC}_i} g_{ii'}R_{i't} = 0 \qquad i = 1,\ldots,n; t = 1,\ldots,T \tag{11.10}$$

$$R_{it} + I_{i(t+\tau-1)} - I_{i(t+\tau)} = d_{i(t+\tau)} \qquad i = n+1,\ldots,N; t = 1,\ldots,T; I_{i\tau} = 0 \tag{11.11}$$

$$k_{it}^{m}\rho_{it}^{m} + a_{it}^{m}Q_{it}^{m} - C_{t}^{m}\pi_{it}^{m} \leq 0 \qquad m = 1,\ldots,M; i \in \text{IP}_m; t = 1,\ldots,T \tag{11.12}$$

Table 11.4. Additional notation

e_m	The eth item in the production sequence of machine m, $e_m = 1_m,\ldots,f_m$ $(f \leq n)$
ρ_{it}^{m}	Binary variable equal to 1 if a new production batch of product i is started on machine m at the beginning of period t and to 0 otherwise
π_{it}^{m}	Binary variable equal to 1 if product i is made on machine m in period t and to 0 otherwise

$$\pi^m_{e_m t} - \sum_{u=1}^{t} \rho^m_{e_m u} + \sum_{u=1}^{t} \rho^m_{(e+1)_m u} = 0 \qquad m = 1,\dots,M; e_m = 1_m,\dots,(f-1)_m; \; t = 1,\dots,T \qquad (11.13)$$

$$\pi^m_{f_m t} - \sum_{u=1}^{t} \rho^m_{f_m u} + \sum_{u=1}^{t} \rho^m_{1_m u} = 1 \qquad m = 1,\dots,M; t = 1,\dots,T \qquad (11.14)$$

$$\sum_{i \in \mathrm{IP}_m} \pi^m_{it} \leq 1 \qquad m = 1,\dots,M; t = 1,\dots,T \qquad (11.15)$$

$$\rho^m_{it} \leq \pi^m_{it}, \pi^m_{it} \in \{0,1\}, \rho^m \in \{0,1\}, Q^m_{it} \geq 0 \qquad m = 1,\dots,M; u \in \mathrm{IP}_m; t = 1,\dots,T \qquad (11.16)$$

$$I_{it} \geq 0 \qquad i = n+1,\dots,N; t = \tau+1\dots,\tau+T \qquad (11.17)$$

$$R_{it} \geq 0 \qquad i = n+1,\dots,N; t = 1,\dots T \qquad (11.18)$$

In model 2, (11.10) and (11.11) are the flow conservation constraints of IP and FP products at the manufacturing location. Constraints (11.12) ensure that production capacity is respected. Constraints (11.13) and (11.14) make sure that the production sequence is respected for each machine. For a given machine m, when $e < f$, constraint (11.13) enforces the number of product $(e{+}1)_m$ changeovers to be less than or equal to the number of product em changeovers for any given period of time. Hence, it forces product $(e+1)_m$ production to start only after the production batch of product e_m is completed. Constraints (11.14) do the same job for product f_m which has the particularity of being last in the machine m production sequence. Thus, after its production batch, machine m has to switch production to product 1_m and start another sequence. Constraints (11.15) makes sure that at most one product is manufactured per period of time for each machine. Finally, constraints (11.16) restrict the changeovers on a machine to the periods in which there is some production.

This is a mixed-integer programming model of moderate size and it can be solved efficiently with commercial solvers such as Cplex. Rizk, Martel and D'Amours (2003) showed, however, that its solution time can be decreased significantly by the addition of appropriate valid inequalities (cuts).

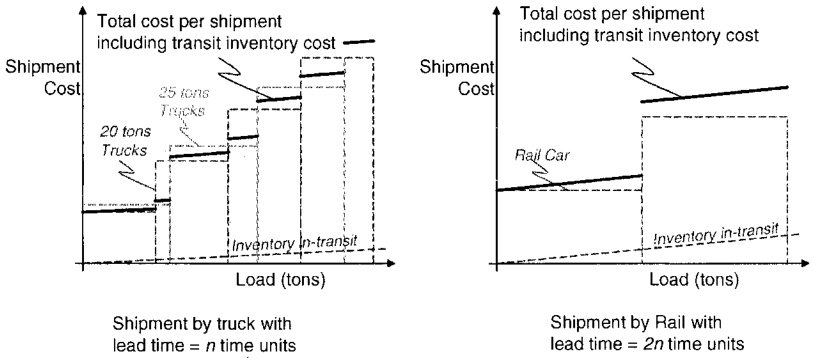
4. Synchronized production-distribution planning problem

4.1 Problem definition and assumptions

In this section, we consider the flow coordination problem of multiple products in a single plant multi-warehouse network. In this network, one or multiple transportation modes are used to replenish different distribution centers with finished goods. The different transportation modes may have different transportation lead times from the plant to its clients and their cost structure can be represented by a general piece-wise linear function $z(S)$ to reflect economies of scale. These transportation economies of scale may have a major impact on inventory planning and replenishment strategies for both the plant and its clients. Transit inventory costs may have an impact on which transportation mode to use between the plant and a destination. Transit inventory costs can be embedded in each transportation mode cost structure as shown in Figure 11.8. Figure 11.8 also shows that, when different transportation modes have the same lead time to a given destination, their cost structures can be amalgamated in a single piece-wise linear function. Major cost savings can be achieved by integrating inventory control and transportation planning.

4.2 Synchronized production-distribution planning model

The type of general piece-wise linear function used to model transportation costs can be represented as a series of linear functions, as shown in Figure 11.9. Let S_j, $j = 0, \ldots, \gamma$, $S_0 = 0$ denote the break points of



Shipment by truck with lead time = *n* time units

Shipment by Rail with lead time = *2n* time units

Figure 11.8. Cost structures for two different transportation modes

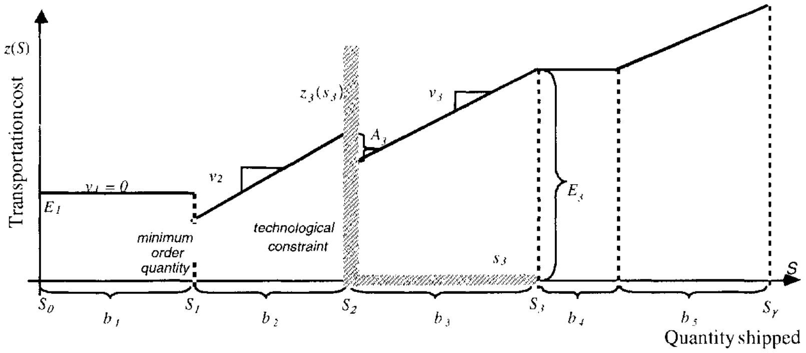


Figure 11.9. General transportation cost function

the piece-wise linear function and let $b_j = S_j - S_{j-1}$, $j = 1 \ldots, \gamma$ denote the length of the jth interval on the S-axis defined by the break points $(S_0, \ldots, S_\gamma)$. Finally, for interval j, let ν_j be the slope of its straight line (variable cost), A_j be the discontinuity gap at the beginning of the interval and E_j be the value of the function at the end of the interval, i.e. $E_j = z(S_j)$. Then, it is seen that for $S_{j-1} < S < S_j$, we have $z(S) = (E_{j-1} + A_j) + \nu_j s_j, s_j = (S - S_{j-1})$.

For an amount S to be shipped in a given period of time, let j be the interval for which $S_{j-1} < S < S_j$, $j \geq 1$, $S_0 = 0$. S can then be expressed as $S = \lambda_j S_j$ where $\lambda_j = S/S_j$ for $j \geq 1$. Based on the above, S can be written in general as $S = \sum_{j=0}^{\gamma} \lambda_j S_j$ where $(S_{j-1}/S_j) < \lambda_j \leq 1$ if $S_{j-1} < S \leq S_j$ and $\lambda_j = 0$ otherwise, for $j = 1, \ldots, \gamma$.The last two conditions can be represented by a binary variable α_j where

$$\alpha_j = \begin{cases} 1, & \text{if } S_{j-1} < S \leq S_j; \\ 0, & \text{otherwise,} \end{cases} j = 1, \ldots, \gamma \text{ and } \alpha_0 = \begin{cases} 1, & \text{if } S = 0; \\ 0, & \text{otherwise.} \end{cases}$$

Using the above observation, S can be expressed in an LP model by the following set of constraints:

$$S = \sum_{j=0}^{\gamma} \lambda_j S_j \tag{11.19}$$

$$\frac{S_{(j-1)}}{S_j} \alpha_j \leq \lambda_j \leq \alpha_j, \quad j = 1, \ldots, \gamma \tag{11.20}$$

$$\sum_{j=0}^{\gamma} \alpha_j = 1 \tag{11.21}$$

$$\alpha_j \in \{0, 1\}, \qquad j = 0, \ldots, \gamma \tag{11.22}$$

From the definition of E_j and the above set of constraints, it is seen that $z(S)$ can be expressed as a linear function of variables α_j and λ_j, $j = 1, \ldots, \gamma$:

$$z(S) = \sum_{j=1}^{\gamma} [(E_j - v_j S_j)\alpha_j + (v_j S_j)\lambda_j] \tag{11.23}$$

In addition, we can observe from constraints (11.21) and (11.22) that α_j, $j = 0, \ldots, \gamma$ form a *Special Ordered Set of type 1* (SOS1) as defined by Beale and Tomlin (1970). Declaring α_j, $j = 0, \ldots, \gamma$ as SOS1, the process of Branch and Bound can be further improved (see Beale and Tomlin, 1970). In addition, by defining α_j, $j = 0, \ldots, \gamma$ as SOS1 along with constraints (11.21), constraints (11.22) are not needed.

For a given destination $w \in W$, let $\beta^w = \text{Min}_{u \in U^w}(\tau_u^w)$. β^w is the shortest transportation lead time to destination w. Let's assume that in period 1, a quantity of product $i \in$ FP is manufactured at the plant and at the end of period 1 we decide to ship an amount of product i to destination w. Because of the production and transportation lead times, the quantity of product i shipped cannot get to destination w earlier than time period $\tau + \beta^w + 1$. Thus, destination w replenishment planning can only start at period $\tau + \beta^w + 1$. Figure 11.10 illustrates different transportation mode shipments (R_{uit}^w) to satisfy the demand for product i at destination w. In this example, the planning horizon includes five ($T = 5$) planning periods. There are three transportation modes available to ship finished products from the plant to destination w ($U^w = \{1, 2, 3\}$). The transportation modes lead time from the plant

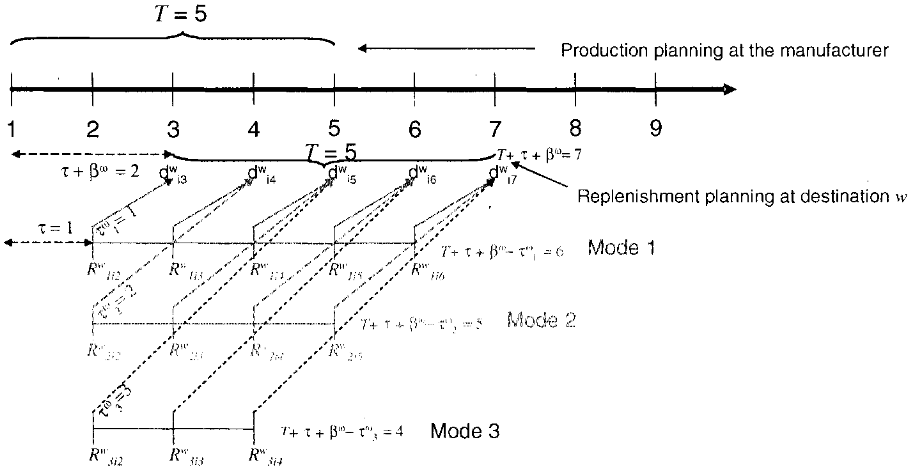


Figure 11.10. Example of multiple transportation mode shipments

to destination w are $\tau_1^w = 1$, $\tau_2^w = 2$, and $\tau_3^w = 3$. Production planning in the plant starts at period 1 and ends at period $T = 5$. On the other hand, because of production and transportation lead times, as stated above, replenishment planning for destination w starts at period $\tau + \beta^w + 1$ and ends at period $T + \tau + \beta^w$. Note that for a given transportation mode $u \in U^w$, only shipments that are made before time period $T + \tau + \beta^w - \tau_u^w$ can get to destination w within its replenishment planning horizon ($[\tau + \beta^w + 1, T + \tau + \beta^w]$). In practice, to get around this difficulty, planning must be done on a rolling horizon basis and the number of periods in the planning horizon must be sufficiently long to have a significant horizon for all the transportation modes, i.e, $T \gg \tau_u^w$, $\forall w \in W$, $u \in U^w$.

Using the notation in Tables 11.1, 11.2, 11.4 and 11.5, the flow coordination problem in a single manufacturer multi-destination network with multiple transportation modes can be formulated as follows:

Table 11.5. Additional notation

j	jth interval of the piece wise linear cost function of a transportation mode u to destination w in period t, $j = 0, \ldots, \gamma_{ut}^w$
S_{utj}^w	The maximum volume (in tons) that can be shipped by transportation mode u to destination w to incur the fixed plus linear cost associated to interval j in period t.
A_{utj}^w	Fixed cost associated to the jth interval of transportation mode u piece-wise linear cost function to destination w in period t.
E_{utj}^w	Cost of shipping the volume S_{utj}^w to destination w by transportation mode u in period t.
v_{utj}^w	Variable cost associated to the jth interval of transportation mode u to destination w piece-wise linear cost function in period t.
τ_u^w	Transportation lead-time to destination w by transportation mode u in period t.
α_{utj}^w	Binary variable associated with the jth interval of mode u to destination w transportation cost function in period t.
λ_{utj}^w	Multiplier associated to interval j of the quantity shipped by transportation mode u from the plant to location w for period t.
d_{it}^w	Effective external demand at destination w for item i during period t.
r_i	Transportation resource absorption rate for item i (in cwt, cube...).
I_{it}^w	Inventory level of finished item i in destination w at the end of period t.
R_{uit}^w	Quantity of finished item i shipped by transportation mode u from the plant to destination w in period t.

Min

$$\sum_{t=1}^{T}\left[\sum_{m\in M}\left[\sum_{i\in \mathrm{IP}_m} K_{it}^{m}\rho_{it}^{m}\right]+\sum_{i\in \mathrm{FP}} h_{i(t+\tau)}I_{i(t+\tau)}\right]+\sum_{w\in W}\left[\sum_{t=\tau+\beta^w+1}^{T+\tau+\beta^w}\left[\sum_{i\in \mathrm{FP}} h_{it}^{w}I_{it}^{w}\right]\right]$$

$$+\sum_{w\in W}\left[\sum_{u\in U^w}\left[\sum_{t=\tau+1}^{T+\tau-\eta_u^w}\left[\sum_{j=1}^{\gamma_{ut}^{w}}(E_{utj}^{w}-v_{utj}^{w}S_{utj}^{w})\alpha_{utj}^{w}+(v_{utj}^{w}S_{utj}^{w})\lambda_{utj}^{w}\right]\right]\right] \tag{11.24}$$

subject to

$$\sum_{m\in M_i} Q_{it}^{m}-\sum_{i'\in \mathrm{SC}_i} g_{ii'}R_{i't}=0, \quad i\in \mathrm{IP}; 1\le t\le T \tag{11.25}$$

$$R_{it}+I_{i(t+\tau-1)}-I_{i(t+\tau)}-\sum_{w\in W}\left[\sum_{u\in U^w} R_{ui(t+\tau)}^{w}\right]=d_{i(t+\tau)},$$
$$i\in \mathrm{FP}; 1\le t\le T; R_{uit}^{w}=0, \forall t\ge T+\tau-\eta_u^w \tag{11.26}$$

$$K_{it}^{m}\rho_{it}^{m}+a_{it}Q_{it}^{m}-C_t^{m}\pi_{it}^{m}\le 0, \quad m\in M; i\in \mathrm{IP}_m; 1\le t\le T \tag{11.27}$$

$$\pi_{e_m t}^{m}-\sum_{u=1}^{t}\rho_{e_m u}^{m}+\sum_{u=1}^{t}\rho_{(e+1)_m u}^{m}\le 0, \quad m\in M;$$
$$1_m\le i_m\le (f-1)_m; 1\le t\le T \tag{11.28}$$

$$\pi_{f_m t}^{m}-\sum_{u=1}^{t}\rho_{f_m u}^{m}+\sum_{u=1}^{t}\rho_{1_m u}^{m}=1, \quad m\in M; 1\le t\le T \tag{11.29}$$

$$\sum_{i\in \mathrm{IP}_m}\pi_{it}^{m}\le 1, \quad m\in M; 1\le t\le T \tag{11.30}$$

$$\sum_{u\in U^w} R_{ui(t-\tau_u^w)}^{w}+I_{i(t-1)}^{w}-I_{it}^{w}=d_{it}^{w}, \quad w\in W; i\in \mathrm{FP};$$
$$\tau+\beta^w+1\le t\le T+\tau+\beta^w; R_{uit}^{w}=0, \forall t<\tau+1 \tag{11.31}$$

$$\sum_{i\in \mathrm{FP}} r_iR_{uit}^{w}-\sum_{j=1}^{\gamma_{ut}^{w}}\lambda_{utj}^{w}S_{utj}^{w}=0,$$
$$w\in W; u\in U^w; \tau+1\le t\le T+\tau-\eta_u^w \tag{11.32}$$

$$(S_{ut(j-1)}^{w}/S_{utj}^{w})\alpha_{utj}^{w}\le\lambda_{utj}^{w}\le\alpha_{utj}^{w},$$
$$w\in W; u\in U^w; \tau+1\le t\le T+\tau-\eta_u^w; 1\le j\le\gamma_{ut}^{w} \tag{11.33}$$

$$\sum_{j=0}^{\gamma_{ut}^{w}}\alpha_{utj}^{w}=1, \quad w\in W; u\in U^w; \tau+1\le t\le T+\tau-\eta_u^w \tag{11.34}$$

$$\rho_{it}^{m}\le\pi_{it}^{m}, \pi_{it}^{m}\in\{0,1\}, \rho^{m}\in\{0,1\}, Q_{it}^{m}\ge 0$$
$$m\in M; i\in \mathrm{IP}_m; 1\le t\le T \tag{11.35}$$

$$I_{it}^{w}\ge 0, \quad w\in W; i\in \mathrm{FP}; \tau+\beta^w+1\le t\le T+\tau+\beta^w \tag{11.36}$$

$$R_{it} \geq 0, I_{i(t+\tau)} \geq 0, \quad i \in \mathrm{FP}; 1 \leq t \leq T \tag{11.37}$$

$$R_{uit}^{w} \geq 0, \quad w \in W; u \in U^{w}; i \in \mathrm{FP}; \tau + 1 \leq t \leq T + \tau - \eta_{u}^{w} \tag{11.38}$$

$$0 \leq \alpha_{utj}^{w} \leq 1, 0 \leq \lambda_{utj}^{w} \leq 1,$$
$$w \in W; u \in U^{w}; \tau + 1 \leq t \leq T + \tau - \eta_{u}^{w}; 1 \leq j \leq \gamma_{ut}^{w} \tag{11.39}$$

$$(\alpha_{ut0}^{w}, \ldots, \alpha_{utj}^{w}) \in \mathrm{SOS1},$$
$$w \in W; u \in U^{w}; \tau + 1 \leq t \leq T + \tau - \eta_{u}^{w}. \tag{11.40}$$

This is a large scale mixed-integer programing model and only small cases can be solved efficiently with commercial solvers such as Cplex. For the case when there is a single distribution center, Rizk, Martel, and D'Amours (2003) proposed valid inequalities which can be added to the model to speed up the calculations. Work on the development of an efficient heuristic method to solve the problem is also currently under way.

5. Conclusion

This chapter presents a review of the supply chain decision processes needed in the pulp and paper industry, from strategic supply chain design to operational planning, but with a particular emphasis on production and distribution planning for a paper mill logistic network. Gradually more relevant and comprehensive planning models are sequentially introduced starting from current industry practice and ending with a sophisticated synchronized production-distribution planning model.

The implementation of these models raises some interesting questions. From a practical point of view, solving the distribution and the production planning problem in sequence may seem interesting, since it reduces the problem size and complexity. Although the size of the problem may increase with the number of intermediary products and planning periods, large linear problems of this sort are easily solved with today's commercial solvers. Under this planning approach, the multi-machine lot-sizing problem provides a better solution than the single-machine lot sizing model where a production cycle constraint is imposed. However, for some demand contexts, experimental work has shown that the potential gains may be small in regard to the planning simplicity induced by the latter approach. Moreover, the imposition of a production cycle time is often useful to synchronize sales and operations, especially when order-promising is conducted on the web.

The last model proposed integrates both production and distribution planning processes. It takes advantage of transportation economies of scale and permits a better selection of transportation modes. However, in order to solve the model within practical time limits, specialized solution methods taking the structure of the problem into account must be developed. An approach which has shown interesting potential, is the addition of valid inequalities (cuts) to the original model. Initial experimentation has shown that the use of appropriate cuts can reduce computation times by an order of magnitude for this class of problem. The application of various decomposition approaches to the solution of the problem is also under study.

It is important to remember that the synchronized production-distribution model assumes that converting facilities are in-house (transportation between roll production and converting facilities is not considered) and over-capacitated in comparison with the bottleneck which was assumed in this chapter to be the paper making machines. Therefore, rolls can be converted within a known delay. Obviously, before applying the model this assumption should be assessed with regard to the company's situation.

Finally, the models presented in this chapter were designed to plan production and distribution over a two-week to a month rolling planning horizon. Since such a short horizon may limit visibility over seasonal parameters, tactical planning models should be used to supply key information to the production-distribution planning model. More specifically they should define end-of-horizon inventory targets for each product produced. Not doing so may results in very bad planning decisions over time, especially in the context of cyclic or highly variable demand. Including such end-of-horizon inventory targets in the model proposed presents no difficulty.

Acknowledgments We would like to thank the business partners of the FOR@C Research Consortium for their financial support and their contributions to our understanding of their planning problems. We would also like to thank the Natural Science and Engineering Research Council (NSERC) for its support through grant # CAP 248987. Finally, we express special thanks to internship student Annie Larochelle who helped in defining the problem with our business partners.

References

Anily, S. (1994). The general multi-retailer EOQ problem with vehicle routing costs. *European Journal of Operational Research*, 79:451–473.

Anily, S. and Federgruen, A. (1990). One warehouse multiple retailers systems with vehicle routing costs. *Management Science*, 36:92–114.

Anily, S. and Federgruen, A. (1993). Two-echelon distribution systems with vehicle routing costs and central inventories. *Operations Research*, 41:37–47.

Barany, I., Van Roy, T.J., and Wolsey, L.A. (1984). Strong formulations for multi-item capacitated lotsizing. *Management Science*, 30:1255–1261.

Beale, E.M.L. and Tomlin, J.A. (1970). Special facilities in general mathematical programming system for non-convex problems using ordered sets of variables. In: *Proceedings of the Fifth International Conference on Operational Research*, pages 447–454, Tavistock Publications, London.

Bitran, G. and Yanasse, H.H. (1982). Computational complexity of the capacitated lot size problem. *Management Science*, 28:1174–1185.

Boctor, F.F. (1985). Single machine lot scheduling: A comparison of some solution procedures. *RAIRO Operations Research*, 19:389–402.

Bouchriha, H., D'Amours, S., and Ouhimmou M. (2003). *Lot Sizing Problem on a Paper Machine Under a Cyclic Production Approach.* FORAC, Working paper, Université Laval.

Bredström, D., Lundgren, J., Mason, A., and Ronnqvist, M. (2003). Supply chain optimization in the pulp mill industry. *EJOR*, In press.

Chan, L.M.A., Muriel, A., Shen, Z.J., Simchi-Levi, D., and Teo, C. (2002). Effective zero-inventory-ordering policies for the single-warehouse multiretailer problem with piecewise linear cost structures. *Management Science*, 48:1446–1460.

Chandra, P. and Fisher, M.L. (1994). Coordination of production and distribution planning. *Journal of the Operational Research Society*, 72:503–517.

De Matta, R. and Guignard, M. (1989). Production Scheduling with Sequence-Independent Changeover Cost. Technical Report, Wharton School, University of Pennsylvania.

Diaby, M., Bahl, H.C., Karwan, M.H., and Zionts, S. (1992). A Lagrangean relaxation approach for very-large-scale capacitated lot- sizing. *Management Science*, 38:1329–1340.

Diaby, M. and Martel, A. (1993). Dynamic lot sizing for multi-echelon distribution systems with purchasing and transportation price discounts. *Operations Research*, 41:48–59.

Dilts, D.M. and Ramsing, K.D. (1989). Joint lotsizing and scheduling of multiple items with sequence dependent setup costs. *Decision Science*, 20:120–133.

Dixon, P.S. and Silver, E.A. (1981). A heuristic solution procedure for the multi-item single level, limited capacity, lotsizing problem. *Journal of Operations Management*, 2(1):23–39.

Dobson. G. (1992). The cyclic lot scheduling problem with sequence-dependent setups. *Operation Research*, 40:736–749.

Dogramaci, A., Panayiotopoulos, J.C., and Adam, N.R. (1981). The dynamic lot sizing problem for multiple items under limited capacity. *AIIE Transactions*, 13(4):294–303.

Eisenhut, P.S. (1975). A dynamic lotsizing algorithm with capacity constraints. *AIIE Transactions*, 7:170–176.

Elmaghraby, S.E. (1978). The economic lot scheduling problem (ELSP): Review and extensions. *Management Science*, 24(6):587–598.

Eppen, G.D. and Martin, R.K. (1987). Solving multi-item capacitated lotsizing problems using variable redefinition. *Operations Research*, 35:832–848.

Everett, G., Aoude, S., and Philpott, A. (2001). Capital planning in the paper industry using COMPASS. In: *Proceedings of 33th Conference of ORSNZ.*

Everett, G. and Philpott, A. (2002). Pulp mill electricity demand management. In: *Proceedings of 33th Conference of ORSNZ.*

Everett, G., Philpott, A., and Cook, G. (2000). Capital Planning under uncertainty at fletcher challenge Canada. In: *Proceedings of 32th Conference of ORSNZ.*

Florian, M., Lenstra, J.K., and Rinnooy Kan, A.H.G. (1980). Deterministic production planning: Algorithms and complexity. *Management science*, 26:12–20.

Forest Products Association of Canada — FPAC. (2002). *2002 Annual Review.*

Gallego, G. and Simchi-Levi, D. (1990). On the effectiveness of direct shipping strategy for the one warehouse multi-retailer r-systems. *Management Science*, 36:240–243.

Gelders, L.F, Maes, J., and Van Wassenhove, L.N. (1986). A branch and bound algorithm for the multi-item single level capacitated dynamic lotsizing problem. In: S. Axsater, Ch. Schneeweiss and E. Siver (eds.), *Multistage Production Planning and Inventory Control, Lecture Notes in Economics and Mathematical Systems 266*, pages 92–108, Springer, Berlin.

Gunther, H.O. (1987). Planning lot sizes and capacity requirements in a single stage production system. *European Journal of Operational Research*, 31(2):223–231.

Haase, K. (1996). Capacitated lot-sizing with sequence dependent setup costs. *OR Spektrum*, 18:51–59.

Haase, K. and Kimms, A. (1996). *Lot Sizing and Scheduling with Sequence Dependent Setup Costs and Times and Efficient Rescheduling Opportunities.* Working paper No. 393, University of Kiel.

Haq, A., Vrat, P., and Kanda, A. (1991). An integrated production-inventory-distribution model for manufacture of urea: A case. *International Journal of Production Economics*, 39:39–49.

Hax A.C. and Candea, D. (1984). *Production and Inventory Management.* Prentice-Hall.

Herer, Y. and Roundy, R. (1997). Heuristics for a one-warehouse multiretailer distribution problem with performance bounds. *Operations Research*, 45:102–115.

Ishii, K., Takahashi, K., and Muramatsu, R. (1998). Integrated production, inventory and distribution systems. *International Journal of Production Research*, 26-3:473–482.

Karmarkar, U.S. and Schrage, L. (1985). The deterministic dynamic product cycling problem. *Operation Research*, 33:326–345.

Keskinocak, P., Wu, F., Goodwin, R., Murthy, S., Akkiraju, R., Kumaran, S., and Derebail, A. (2002). Scheduling solutions for the paper industry. *Operations Research*, 50-2:249–259.

Lambrecht, M.R. and Vanderveken, H. (1979). Heuristic procedure for the single operation, multi-item loading problem. *AIIE Transactions*, 11(4):319–326.

Lasdon, L.S. and Terjung, R.C. (1971). An efficient algorithm for multi-item scheduling. *Operations Research*, 19(4):946–969.

Lehtonen, J.M. and Holmstrom, J. (1998). Is just in time applicable in paper industry logistics? *Supply Chain Management*, 3(1):21–32.

Leung, J.M.Y., Magnanti, T.L., and Vachani., R. (1989). Facets and algorithms for the capacitated lotsizing. *Mathematical Programming*, 45:331–359.

Maes, J. and Van Wassenhove, L.V. (1988). Multi-item single-level capacitated dynamic lot sizing heuristics: A general review. *Journal of the Operational Research Society*, 39(11):991–1004.

Martel, A., Rizk, N., and Ramudhin, A. (2002). *A Lagrangean Relaxation Algorithm for Multi-Item Lot-Sizing Problems with Joint Piecewise Linear Resource Costs.* CENTOR Working paper, Université Laval.

Philpott, A. and Everett, G. (2001). Supply chain optimisation in the paper industry. *Annals of Operations Research*, 108(1):225–237.

Rizk, N. and Martel, A. (2001). *Supply Chain Flow Planning Methods: A Review of the Lot-Sizing Literature.* CENTOR Working paper, Université Laval.

Rizk, N., Martel, A., and D'Amours, S. (2003). *The Manufacturer-Distributor Flow Coordination Problem.* CENTOR Working paper, Université Laval.

Sarmiento, A.M. and Nagi, R. (1999). A review of integrated analysis of production-distribution systems. *IIE Transactions*, 3:1061–1074.

Solomon M. (ed.) (1991). Multi-stage production planning and inventory control. *Lectures Notes in Economics and Mathematical Systems*, 355:92–108, Springer-Verlag, Berlin.

Solomon, M., Kuik, R., and Van Wassenhove, L.N. (1993). Statistical search methods for lotsizing problems. *Annals of Operations Research*, 41:453–468.

Thizy, J.M. and Wassenhove, L.N. (1985). Lagrangean relaxation for the multi-item capacitated lotsizing problem: A heuristic approach. *IIE Transactions*, 17:308–313.

Trigeiro, W.W., Thomas, L.J., and McClain, J.O. (1989). Capacitated lot sizing with setup times. *Management Science*, 35:353–366.

Viswanathan, S. and Mathur, K. (1997). Integrating routing and inventory decisions in one-warehouse multiretailer multiproduct distribution systems. *Management Science*, 43:294–312.

Chapter 12

PRODUCTION PLANNING OPTIMIZATION MODELING IN DEMAND AND SUPPLY CHAINS OF HIGH-VALUE CONSUMER PRODUCTS

Benoit Montreuil

Abstract This chapter is about production planning optimization modeling in the production centers in a demand and supply chain manufacturing, distributing and selling high value consumer products. First, it contrasts demand and supply chain alternatives in terms of collaboration, agility, customer-centricity and personalization offering, with a focus on the implications for production planning optimization. Second, it introduces a comprehensive production planning optimization model applicable to a large variety of centers in demand and supply chains. Third, it contrasts the production planning optimization model instance required as a demand and supply chain is transformed from a rigid and pushy implementation to integrate more collaboration, customer-centricity, agility and personalization. It also puts in perspective the importance of production planning optimization knowledge and technology. Finally it draws conclusive remarks for the research and professional communities.

1. Introduction

This chapter aims to clearly demonstrate that defining and modeling the production planning optimization problems of manufacturing centers in a demand and supply chain is an important activity which depends highly on the collaboration, agility, customer-centricity and personalization offering implemented through the demand and supply chain, as well as on the production planning optimization knowledge and technology available.

In order to contain complexity while insuring widespread representativeness, the chapter deals strictly with the demand and supply chain of manufacturers of high-value products such as vehicles, computers and equipment, sold to consumers in a large geographical region through a network of dealers. Furthermore, it focuses on the production planning optimization of the centers where the products are assembled. Finally, the emphasis is on problem modeling rather than on solution methodologies.

First the chapter presents a comprehensive description of alternate demand and supply chains, and the implications on production planning at the centers assembling the high value products. Second, it introduces a comprehensive production planning model encompassing a large variety of demand and supply chains. Third, it contrasts the production planning optimization model required when a demand and supply chain is transformed from a rigid and pushy mass production and distribution oriented implementation to integrate more collaboration, customer-centricity, agility and personalization. It also puts into perspective the importance of knowledge and technology. Fourth, it draws conclusive remarks for both the research and professional communities.

1.1 Contrasting alternative demand and supply chains

Demand and supply chains define the networks and processes through which demand and supply are expressed, realized and managed by an enterprise and its partners, all the way from suppliers to final customers. The demand and supply chain of an enterprise can take multiple forms, ranging widely in terms of customer-centricity, agility, collaboration and personalization capabilities.

In order to illustrate the spectrum of potential alternatives, Figures 12.1 and 12.2 synthesize two alternative demand and supply chains for an enterprise developing, manufacturing and distributing high value seasonal products to customers, such as recreational vehicles. In both cases the business sells products to several hundred thousand end-user clients spread throughout a large geographical region such as North America or Europe.

1.2 Mass production and distribution oriented demand and supply chain

In Figure 12.1, the demand and supply chain is built according to a mass production and distribution paradigm. It produces a mix of a few hundred standard products in a centralized factory with limited agility, requiring significant setups when its assembly center switches from one product to another. Each product is assembled from thousands of parts, components and modules. The enterprise has selected to produce some

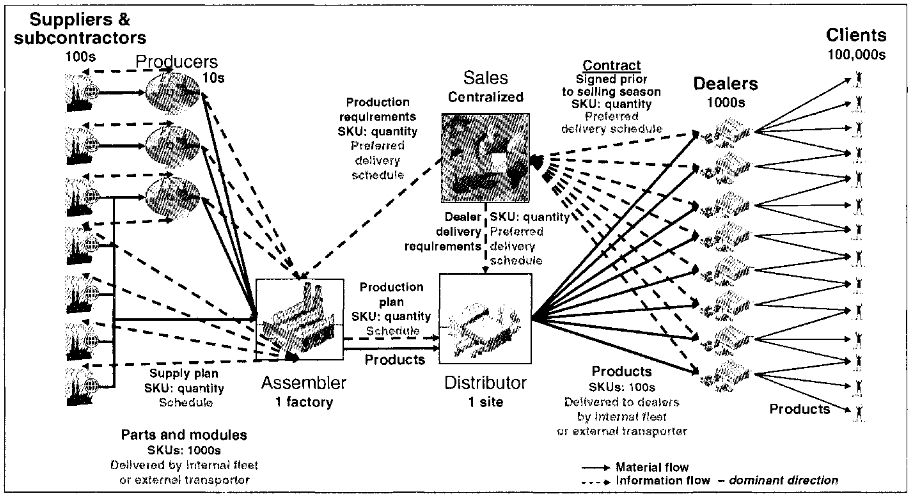

Suppliers & subcontractors
100s
Producers
10s
Production requirements
SKU: quantity
Preferred delivery schedule
Sales
Centralized
Contract
Signed prior to selling season
SKU: quantity
Preferred delivery schedule
Dealers
1000s
Clients
100,000s
Dealer delivery requirements
SKU: quantity
Preferred delivery schedule
Production plan
SKU: quantity
Schedule
Products
Supply plan
SKU: quantity
Schedule
Assembler
1 factory
Distributor
1 site
Products
SKUs: 100s
Delivered to dealers by internal fleet or external transporter
Products
Parts and modules
SKUs: 1000s
Delivered by internal fleet or external transporter
Material flow
Information flow – dominant direction

Figure 12.1. Demand and supply chain of a manufacturer with limited agility offering standard products and imposing pre-season sales contracts to dealers

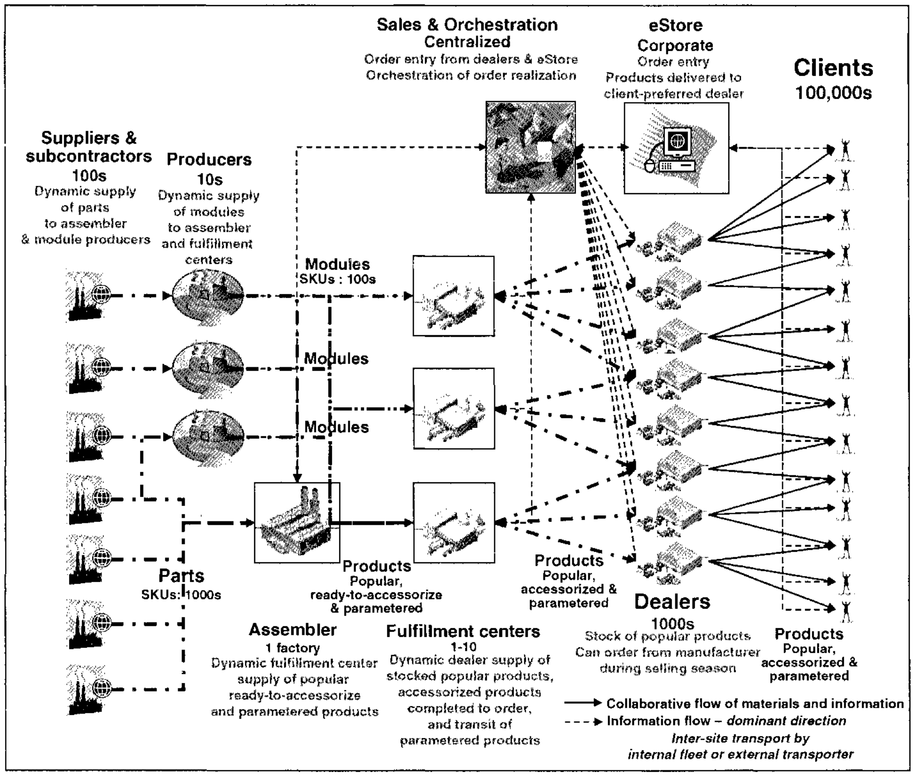

Sales & Orchestration
Centralized
Order entry from dealers & eStore
Orchestration of order realization
eStore
Corporate
Order entry
Products delivered to client-preferred dealer
Clients
100,000s
Suppliers & subcontractors
100s
Dynamic supply of parts to assembler & module producers
Producers
10s
Dynamic supply of modules to assembler and fulfillment centers
Modules
SKUs : 100s
Modules
Modules
Parts
SKUs: 1000s
Products
Popular, ready-to-accessorize & parametered
Products
Popular, accessorized & parametered
Dealers
1000s
Stock of popular products
Can order from manufacturer during selling season
Products
Popular, accessorized & parametered
Assembler
1 factory
Dynamic fulfillment center supply of popular ready-to-accessorize and parametered products
Fulfillment centers
1-10
Dynamic dealer supply of stocked popular products, accessorized products completed to order, and transit of parametered products
Collaborative flow of materials and information
Information flow – dominant direction
Inter-site transport by internal fleet or external transporter

Figure 12.2. Demand and supply chain of an agile manufacturer offering both popular and personalized products with dynamic dealer supply

of these in one of its tens of internal production centers. The others are supplied from external suppliers and subcontractors.

The enterprise sells its products to a network of over a thousand dealers who have the responsibility of selling them to final customers. The dealers are independent businesses, not owned by the enterprise. Several months prior to the selling season, the enterprise forces each dealer to sign a contract stipulating how many units of each product it is buying.

Once all sales to dealers are known, the enterprise knows all production requirements for every product. This allows the factory to establish a master product assembly plan, deciding the sequence of products to be assembled through the entire production season. This plan can be very precise. Illustratively, it may state that from June 15th at 14:00 to June 17th at 10:00, the assembly center is planning to assemble 42 units of product 123 using a single eight-hour shift per day, with a takt time of 15 minutes per product unit. At 10:00 begins a period of 45 minutes, corresponding to three 15-minute cycles, required to change over to product 46 which is the next to be assembled.

The master assembly plan is transposed into an optimized supply plan for every part from every supplier and subcontractor. The supply plans take into account the cost structure, ordering constraint, and lead time speed and reliability of the supplier or subcontractor.

Once products are assembled, the enterprise assigns them to dealers. It optimizes the transportation of the products to their assigned dealers, taking into consideration its internal vehicle fleet and/or its external transporters. The dealers receive their ordered products prior to the heart of the selling season. They must attempt to satisfy clients as best as possible from their available product stock since the enterprise does not allow any reordering after their initial order.

The enterprise imposes such constraints to dealers and clients due to the generalized lack of agility through its supply chain. The assembly center requires significant setup times and costs when switching from a product to another. Its network of internal component/part/module production centers and external suppliers and subcontractors generally does not have the capability and capacity necessary to operate without the stability and visibility offered by the pre-season contract system.

In one variant of this rigid demand and supply chain, the pre-season contract with each dealer gives the enterprise full freedom in deciding when it is to ship the ordered products to the dealer, as long as each receives showcase products early on and the remainder prior to the selling season peak. In another variant, the enterprise is more collaborative with the dealers and lets them stipulate preferred target dates for receiving each unit. In a limited accountability version, the enterprise simply states that it will try to satisfy these targets as closely as possible. In an alternative version of this variant, the enterprise may offer dealers rebates proportional to the deviation between the delivered date and the target date for each ordered product.

1.3 Personalized, customer-centric, collaborative and agile demand and supply chain

Figure 12.2 depicts a much more customer-centric and agile demand and supply chain. Here the enterprise is geared to deliver a personalization offer to customers (Montreuil and Poulin, 2004; Poulin et al., 2004). It offers popular products, expected to be available off-the-dealer-shelves or to be delivered within a few days to the client through its selected dealer. It also offers two types of personalized products: accessorized products and parametered products. Accessorized products are assembled from ready-to-accessorize products to which are added personalized sets of modules. Parametered products are selected by customers through the setting of parameters or options. The personalized products, either accessorized or parametered, are promised to be delivered in a specific number of days to the client through its selected dealer. The order-to-delivery time is promised to be shorter for accessorized products than for parametered products. All products can be ordered by customers either at a dealership or through the web-based eStore operated by the enterprise. In the latter case, the client selects a dealer where he wants the product delivered and where he wants after-sales service.

The main factory has the mandate to assemble standard products, ready-to-accessorize products and parametered products. It is more agile and has lower changeover time from one product to the next. The completion of personalized products from modules and ready-to-personalize products is performed in one of the few fulfillment centers strategically distributed throughout the territory. These fulfillment centers are highly agile, capable of finishing the personalized products on a first-come-first-serve basis with no changeover time from one product to another. The fulfillment centers also serve as transhipment points for parametered products. Both the factory and the fulfillment centers operate during the selling season.

The demand and supply chain has a collaborative nature. From the demand side, on one hand, dealers are allowed to reorder as often as they want. On the other hand, they are asked to collaborate by providing regular forecast updates on their forthcoming demand. The forecasts allow the enterprise to speculatively assemble standard products and ready-to-personalize products. The speculative stocks allow the enterprise to offer faster delivery, especially during the selling season peak where demand may exceed production capacity. The opportunity to build speculative stocks may also permit the enterprise to smooth its production, especially its manpower and supply requirements. From the supply side, the chain exploits collaborative exchange of information, plans and constraints between the partners.

1.4 Production planning implications

In the pushy demand and supply chain of Figure 12.1, the enterprise imposes dealer pre-season contracts and thereafter operates mostly in deterministic high visibility mode. The enterprise is thus free of the demand uncertainty during the production season. It faces demand uncertainty once the selling season starts. Its only reactive mechanism relies on a combination of publicity and rebates as the assembly center is closed and assembled products are already shipped to specific dealers. At the end of the season, it faces the sales results, including a percentage of unsold product units at most dealerships. These unsold units will have to be offered at discount price during the next selling season, cannibalizing the new vehicle market.

In the demand and supply chain of Figure 2, the enterprise is facing a much lower visibility and a much higher uncertainty, as dealers decide to order whenever they want, and clients may similarly order whenever they want directly on the web. High availability and fast and reliable delivery are promised. This implies that the enterprise may regularly find itself with a few-day order booking, having to take decisions based on forecasts. To compensate, the much more agile chain is such that there is much less pressure to produce and order in large lots, which allows faster reaction to occurring events. However this agility is never perfect and numerous constraints may still have to be taken into account.

Production planning, the focus of this chapter, is a seasonal event at the assembly factory of Figure 1. The master assembly plan covers the entire production season in a mostly deterministic fashion. The only probabilistic events are engineering changes, machine breakdowns, quality problems, manpower strikes and supplier lateness. When these occur, they are taken care of by local adjustments to the master schedule. However, the enterprise constantly works at reducing their occurrence by better controlling its chain.

In Figure 12.2, assembly production planning occurs both at the main factory and at fulfillment centers. In both cases the planning horizon is much shorter than when operating according to Figure 12.1. In fact the master plan is an ongoing creation, constantly rejuvenating itself in light of recent events, systematically advancing through a short rolling planning horizon. Beyond the planning horizon lies the forecasting and resource planning horizon, mostly necessary to smooth production and deal with long lead time suppliers. Constant work by the enterprise to become more agile, internally as well as through the network of external suppliers and subcontractors, aims at reducing the need to rely on forecasts. Yet, especially in seasonal markets such as is the focus of the demand and supply chains studied in this chapter, there remains a dependency on some forecasting, mostly related to decisions to build or not anticipatory stock to smooth future production and help to fulfill demand in peak periods.

Both cases share many common features relative to production planning. They are also distinct relative to many others due to differences in customer centricity, agility and personalization offering. Yet, it is important to recognize that they reflect extreme situations. Real situations often lie in between these two extremes, often gradually evolving from the more rigid type to the more agile, customer-centric and personalized type. In light of such insights, section two introduces a comprehensive problem formulation which sustains both cases. The formulation allows dealing with each case by appropriately selecting sets of constraints, variables and by setting parameters.

In both cases the production planning problem definition is highly dependent on the level of consciousness of the decision makers about the impact of the planning on the operations of other stakeholders in the network and about the impact of these stakeholders on the global feasibility and optimality of the production plan. The presentation of the problem formulation in Section 2 is structured to highlight this phenomenon.

2. Formulating the production planning problem

This section provides a comprehensive formulation of the production planning problem in an assembly center driven by a stable takt time establishing the pace of finished products output from the center. The center may consist of a single assembly line or an assembly tree composed of sub-lines recursively feeding a master line. The main decision consists of determining which product to assemble in each takt time slot and when to switch from one product to another, which often involves changeover operations requiring time off in each station, consequently creating a production gap in the center output. This is illustrated in Figure 12.3. In the short extract displayed, it is shown that 33 units of product B are to be assembled, followed by at least 16 units of product M. A unit is to be produced every 15 minutes according to the takt

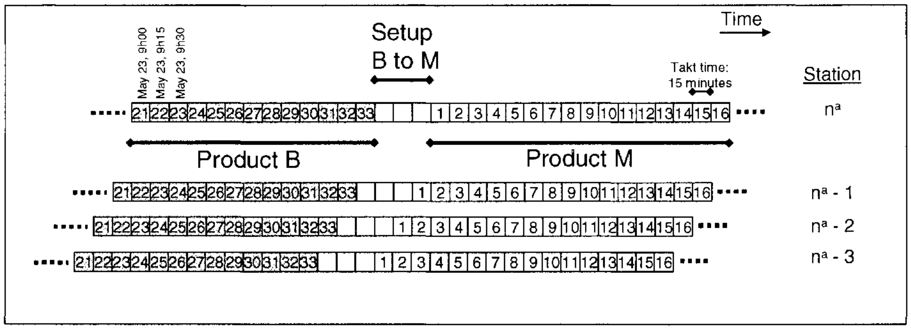


Figure 12.3. Illustrating the production plan of the assembly center

time. For example, unit 21 of product B is to be finished at 9:00 on May 23rd while the 22nd unit is to be finished at 9:15. Between products B and M, the center requires three slots of 15-minute takt time to put in effect the product changeover. Figure 12.3 also shows the forward shift in time of the plans associated with the last station, then the second to last, and so on up to the first station. For example, the 23rd unit of product B is to be finished at 9:30 in the last station, at 9:15 in the second-to-last station, and so on.

In real settings corresponding to Figure 12.1, the production plan (alternatively named master assembly plan thereafter) may readily comprise six months of production, roughly about 120 active days. Often, such centers operate one or two shifts. Assuming a single 8-hour shift per day and a takt time of three minutes, this cumulates to about 19,200 time slots available for assembling a product or making a changeover. Assuming 200 products, this means an assignment matrix of 200 by 19,200, which involves the assignment of 3,840,000 entries in the matrix. These entries are the key decision variables in the problem formulation presented here. This should make it clear that the production planning problem in such a context is a large scale problem.

It should be understood that when the production lot sizes are known to be large, reducing the overall potential number of production runs, then an alternative modeling framework based on start and end times for each production run may be more economical in terms of the number of variables and constraints. However, with the intended goal of the formulation presented here to sustain the full spectrum of possibilities relative to production agility, then the time slot assignment modeling framework is preferred.

The problem formulation is presented below in a modular fashion. First is listed the entire set of sets, indices, parameters and variables.

Second, the objective function is presented from an overall perspective, shown to be adapted depending on problem scoping options. These options depend both on the level of network stakeholder consciousness and collaboration, and on the type of demand and supply chain. Then are iteratively introduced the constraint sets associated with modeling features depending on problem scoping options.

2.1 Sets, indices, parameters and variables

The formulation being comprehensive in nature, it encompasses a large number of decision variables and parameters, requiring a significant number of sets and indices to permit its coherent representation. These are listed hereafter. An effort has been made to have the identifiers as meaningful as possible, however the sheer number of them has forced to use such tricks as superscripts to control complexity.

Sets.

$\boldsymbol{A}$: Set of production/assembly stations composing the production center

$\boldsymbol{B_{sr}}$: Set of time buckets to be used for planning critical resource r of capacitated supplier s, each defined through a starting time t^s_{srb} and a finish time t^e_{srb}

$\boldsymbol{C_p}$: Set of products p' requiring a nonzero changeover time when switching from product p to product p' ($e_{pp'} > 0$)

$\boldsymbol{C_{pa}}$: Set of products p' requiring a positive number of workers at assembly station a to perform the changeover work during the nonzero changeover time when switching from product p to product p'

$\boldsymbol{M}$: Set of all modules m

$\boldsymbol{M_{sr}}$: Set of modules m whose supply requires a positive amount of critical resource r of supplier s

$\boldsymbol{N^{sf}_p}$: Set of cost segments n for speculative stock of product p at the end of the planning horizon

$\boldsymbol{P}$: Set of all products

$\boldsymbol{P_{ma}}$: Set of products requiring module m at assembly station a

$\boldsymbol{R_s}$: Set of critical resources r of capacitated supplier s

$\boldsymbol{S^c}$: Set of capacitated suppliers s

$\boldsymbol{T}$: Set of time periods, linearly sequenced from 0 to t^l

$\boldsymbol{T^a}$: Set of assignable time periods, linearly sequenced from 1 to t^l

$\boldsymbol{T^m}$: Set of time periods at which a change in manpower is allowed

$\boldsymbol{T^s}$: Set of allowed supply time periods, linearly sequenced from t^{ss} to t^l

$\boldsymbol{T^w}$: Set of working time periods, linearly sequenced from t^{sw} to t^l

$\boldsymbol{U}$: Set of cost segments for unused time slots in the assembly center

$\boldsymbol{W}$: Set of worker types
$\boldsymbol{Z}$: Set of geographical zones in the dealership network

Indices.

$\boldsymbol{a}$: A station in the production center
$\boldsymbol{b}$: A planning time bucket
$\boldsymbol{m}$: A module (component, part, etc.)
$\boldsymbol{n}$: A linear cost segment of speculative product inventory, cost increasing with n
$\boldsymbol{p}$, $\boldsymbol{p'}$: A product (when equal to zero, it means "no product")
$\boldsymbol{r}$; A constraining resource
$\boldsymbol{t}$: A time period of duration equal to the takt time of the production center
$\boldsymbol{t^l}$: The last time period
$\boldsymbol{w}$: A worker type
$\boldsymbol{z}$: A geographical zone in the dealership network

Parameters.

$\boldsymbol{c^c_{pp't}}$: Actualized marginal changeover cost when the production center switches from making product p to making product p' at time t
$\boldsymbol{c^e_t}$: Actualized expected unit cost for not finishing a product in time t
$\boldsymbol{c^{im}_{mt}}$: Actualized unit inventory cost for module m at time t
$\boldsymbol{c^{ip}_{pt}}$: Actualized unit inventory cost for product p at time t
$\boldsymbol{c^{om}_{mt}}$: Actualized cost for ordering module m from its supplier at time t, including administration and transport
$\boldsymbol{c^o_t}$: Actualized marginal cost of opening the production center at time slot t as perceived from the end of the production center
$\boldsymbol{c^{qm}_{mt}}$: Actualized unit purchasing cost for module m from its supplier at time t
$\boldsymbol{c^{s-}_{pt}}$: Actualized unit cost per deviation from minimal safety stock target for product p at time t
$\boldsymbol{c^{sf}_{pn}}$: Actualized unit cost of speculative product stock cost at the end of the planning horizon for product p, in cost segment n
$\boldsymbol{c^{sl}_{srb}}$: Actualized marginal cost for using critical resource r of capacitated supplier s during bucket b
$\boldsymbol{c^{sl+}_{srb}}$: Actualized marginal cost for exceeding the average load on critical resource r of capacitated supplier s during bucket b
$\boldsymbol{c^{sl-}_{srb}}$: Actualized marginal cost for underachieving the average load on critical resource r of capacitated supplier s during bucket b
$\boldsymbol{c^T_u}$: Actualized expected marginal cost for not using a number of available time slots in cost segment u

$\boldsymbol{c^v_{zt}}$: Actualized cost for a round-trip transport to zone z departing at time t

$\boldsymbol{c^w_{wt}}$: Actualized marginal cost per period for each worker of type w at time t

$\boldsymbol{c^{w+}_{wt}}$: Actualized marginal cost for adding a worker of type w at time t

$\boldsymbol{c^{w-}_{wt}}$: Actualized marginal cost for removing a worker of type w at time t

$\boldsymbol{d_{pzt}}$: Preferred cumulative deliveries of product p in zone z at time t, summed over the individual preferences of each dealer in zone z

$\boldsymbol{e_{pp'}}$: Number of time periods during which the changeover from product p to product p' requires to stall production at each station in the center

$\boldsymbol{e^T_u}$: Number of unused time slots belonging to cost segment u

$\boldsymbol{f^v}$: Number of vehicles in the fleet

$\boldsymbol{i^s_{pt}}$: Target safety stock for product p at time t

$\boldsymbol{i^{sf}_{pn}}$: Maximum inventory allowed in cost segment n for product p

$\boldsymbol{l_a}$: Time lag between station a and the end of the production center, time between the end of production for a product at station a and its exit of the production center

$\boldsymbol{l^d}$: Required time lag between the production completion of product p at the factory and its availability at the distribution center for delivery to dealers

$\boldsymbol{l^m_{ma}}$: Required time lag between the delivery of module m from its supplier and its use in assembly station a

$\boldsymbol{l^r_{srb}}$: Maximum load to be imposed on critical resource r of capacitated supplier s during planning time bucket b

$\boldsymbol{l^{ra}_{srb}}$: Average load of critical resource r of capacitated supplier s over the planning horizon, adjusted to the length of time bucket b

$\boldsymbol{l^s_m}$: Lead time from order to delivery of module m by its supplier to the factory

$\boldsymbol{n^a}$: Number of assembly stations

$\boldsymbol{n^p}$: Number of products

$\boldsymbol{n^f_p}$: Number of cost segments for anticipatory stock of product p at the end of the planning horizon

$\boldsymbol{n^{\min}_{wt}}$: Minimal allowed number of workers of type w at time t

$\boldsymbol{n^{\max}_{wt}}$: Maximal possible number of workers of type w at time t

$\boldsymbol{o_p}$: Total order for product p

$\boldsymbol{q_{mpa}}$: Quantity of modules m required per unit of product p at assembly station a

$\boldsymbol{q^h_m}$: Total quantity of modules m required to assemble all products demanded by the dealers over the planning horizon

$\boldsymbol{q^r_{srm}}$: Quantity of resource r required to produce one unit of module m at supplier s

q_m^s: Minimum ordering lot size imposed by the supplier of module m

r_{pzt}: Actualized revenue from delivering a unit of product p to dealers in zone z at time t

r_{wap}^a: Number of workers of type w required at assembly station a when assembling product p

$r_{wapp'}^c$: Number of workers of type w required at assembly station a when changing over from product p to product p'

s_p: Space occupied by a unit of product p in a transport vehicle

s^v: Space availability in a transport vehicle

t^{ss}: First time period at which a module may be ordered from a supplier ($t^{ss} \leq 0$)

t^{sw}: First time period at which any change can be made to the work assignment of any station of the production center (generally ≤ 0)

t_{srb}^e: End time of bucket b

t_{srb}^s: Start time of bucket b

v_z^t: Travel time for a round-trip to zone z by a delivery vehicle

Variables.

A_{pt}: Binary variable stating whether or not a unit of product p is to be finished at time t

$C_{pp't}$: Binary variable stating whether or not a changeover from product p to p' starts at time t, as perceived at the end of the production center

C^c: Nonnegative real variable summing the actualized total changeover cost

C^e: Nonnegative real variable summing the actualized total cost for empty production slots, not finishing products at each time in the planning horizon

C^i: Nonnegative real variable summing the actualized total product inventory cost

C^o: Nonnegative real variable summing the actualized total line opening cost

C^s: Nonnegative real variable summing the actualized total supply cost

C^{s-}: Nonnegative real variable summing the actualized total cost for deviation from minimal safety stock targets for products

C^{sf}: Nonnegative real variable summing the actualized total speculative product stock cost at the end of the planning horizon

C^v: Nonnegative real variable summing the actualized total vehicle transport cost

C^w: Nonnegative real variable summing the actualized total personnel cost

$\boldsymbol{D_{pzt}}$: Nonnegative real variable computing the cumulative dealer network delivery of product p at time t

$\boldsymbol{D^{+}_{pzt}}$: Nonnegative real variable computing the over-delivery of product p to dealer network at time t

$\boldsymbol{D^{-}_{pzt}}$: Nonnegative real variable computing the under-delivery of product p to dealer network at time t

$\boldsymbol{D^{t}_{pzt}}$: Nonnegative real variable computing the punctual delivery of product p to dealer network at time t

$\boldsymbol{E_t}$: Binary variable equal to one only when no product is to be finished at time t

$\boldsymbol{E^{T}_{u}}$: Nonnegative real variable computing the number of unused assembly time slots (not producing products) belonging to the cost segment u, during the entire planning horizon

$\boldsymbol{I_{pt}}$: Nonnegative real variable computing the distribution center inventory of product p at time t

$\boldsymbol{I^{m}_{mt}}$: Nonnegative real variable computing the inventory of modules m at time t

$\boldsymbol{I^{s+}_{pt}}$: Nonnegative real variable computing the positive deviation from the safety stock target for a product p at time t

$\boldsymbol{I^{s-}_{pt}}$: Nonnegative real variable computing the negative deviation from the safety stock target for a product p at time t

$\boldsymbol{I^{sf}_{pn}}$: Nonnegative real variable computing the speculative stock of product p at the end of the planning horizon, belonging to cost segment n

$\boldsymbol{L_{srb}}$: Nonnegative real variable computing the load on critical resource r of capacitated supplier s during a planning time bucket b

$\boldsymbol{L^{+}_{srb}}$: Nonnegative real variable computing the above average loading on critical resource r of capacitated supplier s during a planning time bucket b

$\boldsymbol{L^{-}_{srb}}$: Nonnegative real variable computing the under average loading on critical resource r of capacitated supplier s during a planning time bucket b

$\boldsymbol{N_{wt}}$: Nonnegative integer variable computing the number of workers of type w active in the production line at time t

$\boldsymbol{N^{+}_{wt}}$: Nonnegative integer variable computing the number of workers of type w added in the production line at time t

$\boldsymbol{N^{-}_{wt}}$: Nonnegative integer variable computing the number of workers of type w removed from the production line at time t

$\boldsymbol{O^{f}_{t}}$: Binary variable stating whether the production center is open at time period t, as perceived from the end of the center

$\boldsymbol{O^{s}_{mt}}$: Binary variable stating whether or not an order of modules m is transmitted to its supplier at time t

$\boldsymbol{P_{pt}}$: Nonnegative real variable cumulating production of product p up to time t

$\boldsymbol{Q^s_{mt}}$: Nonnegative real variable computing the quantity of modules m ordered from its supplier at time t

$\boldsymbol{R_{wt}}$: Nonnegative real variable computing the number of workers of type w required at time t

$\boldsymbol{R^d}$: Nonnegative real variable computing the actualized total revenue generated from deliveries to dealer network

$\boldsymbol{R^s_{mt}}$: Nonnegative real variable computing the cumulative total number of modules m required from its supplier up to time t

$\boldsymbol{U_u}$: Binary variable stating whether or not there is a greater-than-zero number of unused assembly time slots during the planning horizon corresponding to cost segment u

$\boldsymbol{V_{zt}}$: Nonnegative integer variable computing the number of transport vehicles departing to zone z at time t

2.2 Objective function

Production planning greatly influences the flow of revenues and costs through the planning horizon. From the revenue side, in the studied cases, sales are registered once a product is delivered to a dealer. Dealers order a variety of products, with distinct margins associated to each product. Therefore the production sequence affects the availability of products for delivery, which affects the deliveries to dealers, which affects the revenue stream. Also, especially in the agile and client-centric demand and supply chain of Figure 12.2, time spent on changeovers in the assembly center reduces the potential for producing products required by customers, thus having an impact on overall sales. This influence of production planning on revenue leads to using a maximizing objective function as stated in equation (12.1).

Objective function.

$$\text{Maximize } R^d - (C^o + C^c + C^w + C^i + C^v + C^s + C^{s-} + C^{sf} + C^e) \quad (12.1)$$

The above statement of the objective function is purposefully limited to identifying the main aggregate revenue and cost variables. Detailed specification of each cost variable is to be addressed in a modular fashion in the next sections. However it is important to state that all variables in the objective function are actualized, taking into consideration the present value of future costs and revenues. All cost and revenue parameters are also allowed to be time dependent.

The costs in the objective function include, in their presentation order in (12.1):

(1) Center opening cost;
(2) Product changeover cost;
(3) Personnel cost;
(4) Product inventory cost;
(5) Transport-to-dealer cost;
(6) Supply cost;
(7) Safety stock target deviation cost;
(8) Speculative product stock cost;
(9) Empty production slot cost.

Cost variables (1) and (2) encompass the most basic costs which an enterprise is conscious of when planning production. Their computation is modelled in Section 2.3 dealing with operation and changeover in the assembly center. Third, the personnel cost variables are cumulating costs associated to manpower requirements and variations. They are modelled in Section 2.4 dealing with personnel. Product inventory and transport-to-dealer cost variables (4) and (5) are addressed in Section 2.5 dealing with the dealer network. The sixth cost variable corresponds to the supply cost and is modelled in Section 2.6 dealing with the supply network. The last three cost variables (7) to (9) are modeling costs associated with dealing with the dynamic uncertainty in a rolling horizon mode, which is dealt with in Section 2.7.

2.3 Dealing with operation and changeover in the assembly center

The core operational decision variables in planning production in the assembly center are the A_{pt} and $C_{pp't}$ variables. The former variables state whether or not product p is to be assigned to production time slot t, finishing the product in the last assembly station at time t. The latter variables state whether or not a changeover from product p to product p' is to be initiated in time slot t. From these are derived the following sets of constraints defining the core operations and costs of the assembly center.

Operations and changeover constraints.

$$P_{pt} = P_{p,t-1} + A_{pt}, \quad \forall p \in P; \forall t \in T^a \tag{12.2}$$

$$P_{pt^l} \geq o_p, \quad \forall p \in P \tag{12.3}$$

$$\sum_p A_{pt} + E_t = 1, \quad \forall t \in T^a \tag{12.4}$$

$$(1 - C_{pp't}) \geq \sum_p A_{pt'}, \quad \forall t' \in [t, t-1+e_{pp'}]; \forall t \in T; \quad \forall (p \in P, p' \in P) \mid p' \in C_p \tag{12.5}$$

$$\sum_{p' \neq p} C_{pp't} \leq A_{p,t-1}, \quad \forall p \in P, \forall t \in T^a \tag{12.6}$$

$$A_{p't} \leq A_{p'(t-1)} + \sum_{p} C_{pp'(t-e_{pp'})}, \quad \forall p' \in P; \forall t \in T^a \tag{12.7}$$

$$\sum_{p'} C_{pp't} = 1, \quad p = 0; t = 1 \tag{12.8}$$

$$O_t^f = \sum_{p \in P} A_{pt} + \sum_{p \in P} \sum_{\substack{p' \in P \\ e_{pp'} > 0}} \sum_{t' \in [t, t-1+e_{pp'}]} C_{pp't'}, \quad \forall t \in T^a \tag{12.9}$$

$$C^o = \sum_{t \in T^a} c_t^o O_t^f \tag{12.10}$$

$$C^c = \sum_{p \in P} \sum_{\substack{p' \in P \\ p' \neq p}} \sum_{t \in T^a} c_{pp't}^c C_{pp't} \tag{12.11}$$

Constraint set (12.2) computes the cumulative production of each product p at each time t. Constraint set (12.3) imposes that the cumulative production of each product p at the end of the planning horizon be at least as high as the total number of orders for product p, ensuring that all orders are planned to be fulfilled. Constraint set (12.4) limits each production time slot to be assigned at most one product unit. It also identifies the time slots during which no product units are produced, the center being either idle or changing over from one product to another. Constraint set (12.5) imposes that no product be assembled during the changeover time from a product p to another p' starting at time t. For example, in Figure 12.3, this imposes the three-time-slots changeover time from product B to product M in each assembly station. Note that this changeover time may be dependent on the pair of from-to products. For example, changing over from B to M may take 3 time slots, but changing from B to S may take 20 time slots. Constraint set (12.6) restricts a changeover from p to p' to be allowed at time t only if product p is allocated for production in time $t-1$. Conversely, constraint set (12.7) restricts production of product p' in time t to be allowed only if product p' was already produced in the previous time slot or if a changeover has been performed from some product p to product p' in the previous time slots, according to the specified changeover duration. Constraint set (12.8) initializes the production by deciding which product is to be produced first, requiring beforehand an initial setup. It uses product zero as surrogate for stating the initial changeover time requirements. Constraint set (12.9) determines whether or not each time slot is open or not, active either in producing a product unit or changing over from one product to another.

Constraints (12.10) and (12.11) respectively cumulate the costs for opening production time slots and for the inter-product changeovers.

The operational and changeover constraint sets can be generalized to deal with multiple parallel assembly centers, each with specific capabilities in terms of which products it can assemble. This generalization is beyond the scope of the chapter and left as an exercise to the interested reader.

2.4 Dealing with personnel

Assembly centers such as those modelled in this chapter may readily employ multiple hundreds of people. These represent important costs. The production planning decisions influence the need for personnel, and therefore the personnel costs. Opening time slots implies having an adequate number of persons to operate the center during those slots. Furthermore, it is often the case that each product requires a specific number of workers of each type in each assembly station. For example, a small and simple product may require less people in the assembly center than a large complex product. Two products may have very similar personnel requirements at each station, except a few where they differ dramatically due to their specifications.

In many enterprises, personnel cost is not dealt with explicitly when developing the master assembly plan. In such cases the plan is forwarded to the human resources center which has the responsibility of providing and assigning the right set of people to the center. The production plan is thus optimized without considering the personnel costs, and the personnel costs are afterward minimized given the production plan constraints. Adjustments may be made to the production plan to deal with infeasibilities.

Below are presented the constraint set allowing to consciously integrate personnel into the production planning optimization.

Personnel constraints.

$$n_{wt}^{\min} \leq N_{wt} \leq n_{wt}^{\max}, \quad \forall w \in W; \forall t \in T^w \tag{12.12}$$

$$N_{w,t-1} + N_{wt}^{+} - N_{wt}^{-} = N_{wt}, \quad \forall w \in W; \forall t \in T^m \tag{12.13}$$

$$N_{w,t-1} = N_{wt}, \quad \forall w \in W; \forall t \in \{T^w - T^m\} \tag{12.14}$$

$$R_{wt} \leq N_{wt}, \quad \forall w \in W; \forall t \in T^w \tag{12.15}$$

$$R_{wt} \geq \left[\sum_{a \in A} \sum_{p \in P} r^a_{wap} A_{p,t+l_a} + \sum_{a \in A} \sum_{p \in P} \sum_{p' \in C_{pa}} r^c_{wapp'} \sum_{s=t+l_a}^{t+l_a-1+e_{pp'}} C_{pp's} \right], \qquad \forall w \in W; \forall t \in T^w \tag{12.16}$$

$$C^w = \sum_{t \in T^a} \sum_{w \in W} c^w_{wt} N_{wt} + \sum_{t \in T^m} \sum_{w \in W} (c^{w+}_{wt} N^+_{wt} + c^{w-}_{wt} N^-_{wt}). \tag{12.17}$$

The key decision variables related to personnel are the N_{wt} and R_{wt} variables which state the number of workers of type w to be respectively working and required in the assembly center in time t. For each time t, constraint set (12.12) bounds this number to be lower than the available pool of workers of the specific type and to be higher than the union-negotiated and/or strategically-planned lower limit on the number of workers of this type. Constraint set (12.13) computes the increase and decrease in workforce of each type occurring at each time period. Constraint set (12.14) forces workforce increases or decreases to be occurring only at allowable times. It does so by forcing the workforce to be the same as in the precedent time slot whenever workforce changes are not allowed in the time slot. This set is included for presentation clarity. However it leads to variable set reduction prior to solving the problem.

Constraint set (12.15) insures that the number of workers of each type in each time slot is always greater or equal to the required number to realize the production at each assembly station. Constraint set (12.16) computes the number of workers of each type required at each assembly station at time t given the production and changeover decisions.

Constraint (12.17) cumulates the total personnel cost, combining for each time slot the cost of working employees, the cost of adding employees and the cost of removing employees. These latter costs may be expensive when numerous variations of staffing level occur. Constraint (12.17) could be made even more rigorous by adding the cost of moving personnel around in the center to deal with varying personnel requirements at each station from one time slot to the next, whenever this cost becomes significant and influenced by the production plan. This is left as an exercise to the interested reader.

2.5 Dealing with the dealer network

Demand and supply chains such as those studied in this chapter involve thousands of dealers geographically spread throughout large regions. At production planning it is generally too cumbersome to explicitly deal with each dealer. So enterprises make compromises.

Most mass production oriented enterprises completely discard them from their modeling. In fact the dealers only appear in aggregate form in constraint set (12.3) which makes sure that somehow during the production season all orders from all dealers are produced. Dealing with the assignment of production to dealers is left as an aftermath decision to be dealt with by distribution managers.

Even within the framework of the demand and supply chain of Figure 12.1, there is potential and reward for production planning to explicitly integrate the dealer network in its optimization modeling. In a demand and supply chain that is customer-centric, agile and/or offering personalization, explicitly dealing with the dealer network is a requisite. Below is presented a set of constraints which allows modeling the issues relevant to production planning that are related to the dealer network.

By the way, this is where the notion of inventory is introduced. If the dealer network is not explicitly modelled, then the production planner either assumes that finished products will be shipped efficiently to dealers in a prompt manner after their availability for delivery or that the inventory-transport decisions will be subordinated to the production plan without significant loss of optimality. Dealing with product inventory involves defining the set of variables I_{pt} stating the inventory of product p at time t.

Dealer network constraints.

$$I_{pt} = P_{pt} - \sum_{z \in Z} D_{pzt}, \qquad \forall p \in P; \forall t \in T^a \tag{12.18}$$

$$P_{p(t-l^d)} \geq \sum_{z \in Z} D_{pzt}, \qquad \forall p \in P; \forall t \in T^a \tag{12.19}$$

$$D_{pzt} - d_{pzt} = D^+_{pzt} - D^-_{pzt}, \qquad \forall p \in P; \forall z \in Z; \forall t \in T^a \tag{12.20}$$

$$D^t_{pzt} = D_{pzt} - D_{pz,t-1}, \qquad \forall p \in P; \forall z \in Z; \forall t \in T^a \tag{12.21}$$

$$s^v V_{zt} \geq \left[\sum_{p \in P} \sum_{t \in T^a} s_p D^t_{pzt} \right], \qquad \forall z \in Z; \forall t \in T^a \tag{12.22}$$

$$\sum_{z \in Z} \sum_{t' \in [t+1-v^z_t, t]} V_{zt'} \leq f^v, \qquad \forall t \in T^a \tag{12.23}$$

$$R^d = \sum_{p \in P} \sum_{t \in T^a} \sum_{z \in Z} r_{pzt} D^t_{pz,t-v^t_z/2} \tag{12.24}$$

$$C^i = \sum_{t \in T^a} \sum_{p \in P} c^{ip}_{pt} I_{pt} \tag{12.25}$$

$$C^v = \sum_{t \in T^a} \sum_{z \in Z} c^v_{zt} V_{zt} \tag{12.26}$$

In order to model the dealer network, it is clustered in a set of dealer zones grouping nearby dealers. Furthermore, the delivery timing preferences of each dealer are pooled to generate the set of parameter d_{pzt} stating the preferred cumulative deliveries of product p in zone z at time t. Two key sets of variables allow modeling the dealer related decisions, these are D_{pzt} and V_{zt}. The D_{pzt} variables compute the cumulative delivery of product p to the dealers in zone z up to time t. The V_{zt} variables stipulate the number of transport vehicles departing to zone z at time t, so as to deliver products to dealers in that zone z.

Constraint set (12.18) computes the inventory of product p in time t as the difference between the cumulative production of product p up time t and the sum over all regions of the cumulative deliveries of product p to these regions up to time t. Constraint set (12.19) insures that sufficient production of product p is realized prior to its delivery to dealers, with enough lead time to permit transit from the factory to the distribution center and preparation for delivery. Constraint set (12.20) computes the positive and negative differences between the preferred and achieved cumulative deliveries of product p to dealers in zone z at time t.

The punctual deliveries of product p to zone z shipped at time t are determined through constraint set (12.21). Cumulating these punctual product-to-zone deliveries at time t for each zone, constraint set (12.22) determines the corresponding number of transport vehicles departing to the zone, given the space requirements of each product in the vehicle and the spatial capacity of the vehicles. Constraint set (12.23) limits the number of vehicles simultaneously travelling to never exceed the fleet size. These constraint sets can easily be generalized to concurrently deal with volume and weight capacities, distinct types of transport vehicles, and combinations of internal fleet and external transporters.

The total actualized revenues are computed through constraint set (12.24). It assumes revenues to be registered at delivery time to dealers, approximated to mid round trip to the dealer zone. Other revenue actualization can be similarly modelled to reflect specific situations. Constraints (12.25) and (12.26) respectively compute the total actualized inventory and transport costs.

2.6 Dealing with the supplier network

By their intrinsic nature, supply networks studied in this chapter involve a large number of supplied parts, components and/or modules, supplied by many suppliers and subcontractors. Indeed it is common to deal with several thousands of items by hundreds of organizations throughout the world.

Everybody having been involved in assembly factories readily recognizes that a lot of the operational problems leading to difficulties in delivering productively, fast and reliably are related to missing or incorrect supplies. In fact, when supplies are always on time at the right assembly stations, piloting the assembly center becomes much easier. Furthermore, even though everyone dreams of agile nonconstraining suppliers delivering perfect products just in time with short notice, there are nearly always suppliers with significant and unreliable delivery times, imposing minimal supply lots, and subject to limited supply capacity. These may have significant impact on the feasibility and profitability of production plans. Yet in most cases, supply planning is performed subject to a predetermined production plan and forecasts, according to a material requirement planning (MRP) logic. Below are presented sets of constraints allowing for integration of supply planning of influential inputs and suppliers to production planning optimization.

Dealing with the supply network.

$$R^s_{mt} = R^s_{m,t-1} + \sum_{a\in A}\sum_{p\in P_{ma}} q_{mpa} A_{p,(t+l_a+l^m_{ma})}, \qquad \forall m \in M; \forall t \in T^a \tag{12.27}$$

$$\sum_{\substack{t'\in T^s \\ t'\le t-l^s_m}} Q^s_{mt'} = R^s_{mt} + I^m_{mt}, \quad \forall m \in M; \forall t \in T^s \tag{12.28}$$

$$Q^s_{mt} \le q^h_m(1 - O^s_{mt}), \quad \forall m \in M; \forall t \in T^s \tag{12.29}$$

$$Q^s_{mt} \ge q^s_m O^s_{mt}, \quad \forall m \in M; \forall t \in T^s \tag{12.30}$$

$$L_{srb} = \sum_{m\in M_{sr}} q^r_{srm}\big(R^S_{mt^s_{srb}} - R^S_{mt^e_{srb}}\big), \qquad \forall s \in S^c; \forall r \in R_s; \forall b \in B_{sr} \tag{12.31}$$

$$L_{srb} \le l^r_{srb}, \quad \forall s \in S^c; \forall r \in R_s; \forall b \in B_{sr} \tag{12.32}$$

$$L_{srb} - l^{ra}_{srb} = L^+_{srb} - L^-_{srb}, \quad \forall s \in S^c; \forall r \in R_s; \forall b \in B_{sr} \tag{12.33}$$

$$C^s = \sum_{t\in T^s}\sum_{m\in M}(c^{om}_{mt}O^s_{mt} + c^{qm}_{mt}Q^s_{mt} + c^{im}_{mt}I^m_{mt} + c^{sl}_{srb}L_{srb}) + \sum_{s\in S^c}\sum_{r\in R_s}\sum_{b\in B_{sr}}(c^{sl+}_{srb}L^+_{srb} + c^{sl-}_{srb}L^-_{srb}). \tag{12.34}$$

The key variables allowing to deal explicitly with supply are the R^s_{mt}, Q^s_{mt} and I^m_{mt} variables, respectively deciding the supply requirements, supply quantity ordered and the current inventory of module (part, component, etc.) m at time t. The link between supply and production is set by parameters q_{mpa} stating the quantity of modules m required per unit of product p treated at assembly station a of the assembly center.

Constraint set (12.27) transposes the assembly plan decisions into their supply implications. Explicitly, they update the cumulative supply requirements for each module m at time t by adding to the previous cumulative requirements at time $t-1$ the material requirements generated by the products assigned to each assembly station at time t. They allow for setting a required time lag between the delivery of a module m from its supplier and its use in assembly station a for internal logistic considerations and protection against supplier delivery time unreliability.

Constraint set (12.28) balances on one side the cumulative quantity of ordered modules m planned to have been delivered at time t, taking delivery lead time in consideration, and on the other side the combination of current module inventory and cumulative consumption of these modules due to production requirements up to time t.

Constraint set (12.29) makes sure that an order of modules m is transmitted at time t to the supplier for actually allowing the ordered quantity of module m to be greater than zero. This then allows constraint set (12.30) to impose supplier specified minimum order quantities whenever an order of modules m is transmitted to its supplier at time t.

Constraint sets (12.31) to (12.33) allow supporting collaborations with critical suppliers allowing the enterprise to know and exploit in its planning their key resource constraints so as to insure supply feasibility and minimize their joint supply costs. For example, if a supplier of a specialized module is known to have a production capacity of 10 modules per day, then the enterprise can integrate this knowledge in the assembly plan and therefore avoid both keeping unnecessary product inventory and avoiding supply disruptions associated with infeasibilities due to limited capacity at the supplier site. Similarly, if it is important for a supplier to smooth its loading on a critical resource, then this can be taken into consideration through the planning optimization. The constraint sets permit at the extreme to synchronize the constraints to the assembly center takt time, but allow for aggregating the resource constraints limitations in terms of a time bucket (shift, day,. week, etc.) specified for each critical resource r of critical supplier s. These time buckets can be set to variable durations stated in terms of shifts, days, weeks, and so on through the use of parameters t^s_{srb} and t^e_{srb} setting the start and end times of bucket b for resource r of supplier s.

Constraint set (12.31) transposes the supply requirements into loads on resource r of supplier s during time bucket b. Constraint set (12.32) limits the load on resource r not to exceed its capacity during time bucket b. Constraint set (12.33) permits to model supplier resource smoothing by computing punctual positive and negative deviations from the ideally smoothed average loading of resource r of supplier s during bucket b.

Constraint (12.34) adds up all supply related costs. These include the order costs, the purchase costs, the inventory costs, the critical resource usage costs and the critical resource smoothing costs.

The costs are linearly modelled through constraint (12.34). Variants can be developed to allow, for example, the module purchase cost to be a stepwise function of the quantity ordered exhibiting economies of scale. Similarly, constraint sets (12.28) to (12.34) assume a single supplier per item supplied as is currently the most common case for significant supplied items. The model can be readily upgraded to deal with multiple suppliers per item. This is again left as an exercise for the interested reader.

2.7 Dealing with future dealer demand uncertainty

The mass production oriented demand and supply chain of Figure 12.1 does not allow dealers to reorder after signing their original pre-season order. In such a context, all products to be assembled during the production season are known a priori. On the contrary, in the agile customer-centric demand and supply chain of Figure 12.2, the dealers may order any time they want prior to and during the selling season. This means that at planning time, the enterprise has in its hands a set of orders from dealers and a set of forecasts for future demands. These forecasts can be collected and synthesized from forecasts provided by individual dealers and/or can be generated by the enterprise based on sales history and a variety of predictive indicators.

When continuous ordering is allowed, it does not make sense for the planning to specify production assignments of products to time slots in the assembly centers way ahead in the future, far beyond the range of actual orders, since the future demand is unknown and forecasts are bound to be imperfect. Therefore the production planning horizon is generally much shorter than up to the end of the entire selling season, often in the order of days or weeks. Furthermore, contrary to the more stable and deterministic case of Figure 12.1 where the production planning horizon covers the entire production season and the production plan is only to be re-optimized due to major events in the supply chain such as supply problems, in the agile and customer-centric context of Figure 12.2 the production plan is to be re-optimized in a very frequent rhythm, often daily, in order to adjust to events in the demand chain such as new orders and adjusted forecasts as well as events in the supply chain.

There are multiple ways to model demand uncertainty. In this chapter, it is treated by differentiating four uses for products assembled in a period:

(1) Fulfillment of an actual order from a dealer in a zone;
(2) Expected fulfillment of probable orders within the planning horizon, based on average expected demand per time period for each product;
(3) Amplification of a safety stock specified for allowing fast response to forthcoming yet unknown dealer orders;
(4) Amplification of a speculative stock to deal with future demand beyond the planning horizon.

As an illustrative example, assume that the production planning horizon is set to two weeks and that the current time is set in the early stages of the selling season, before the selling peaks. The first usage covers all officially registered orders from dealers. This may correspond to 25% of the assembly center capacity during the two-week horizon. The second usage corresponds to the expected average consumption of products, forecast to be coming from not yet registered dealer orders during the next two weeks. For example, this may correspond to an average of two units of product 46 per day. The combination of all these forecasts may occupy 40% of the center capacity. This leaves 35% remaining capacity to deal with the third and fourth usages. Relative to the third, given the forecast uncertainty and the required level of service, the enterprise may set a target safety stock to be maintained at all times during the planning horizon. For example, it may set a target safety stock of 10 units of product 46 in the first week and 12 units during the second week. Given the current level of stock of product 46, equal to eight units, then this implies raising it by two units in the first week and two more units in the second week. Adjusting the safety stocks may, for example, require 10% of the center capacity. This leaves 25% remaining capacity for usage four, which is to build anticipatory stock to be used after the planning horizon, to prepare for the forthcoming peaks. If the forthcoming peaks are not so high and that future capacity is expected to be able to handle them, then the enterprise may decide not to produce anything more during the current two-week horizon, as planned at the current time. This decision may be altered in the coming days if new increased forecasts become available. To the contrary, the enterprise may profit from the currently available capacity to build inventory to sustain high future peaks beyond future capacity to handle by itself. The task is then to decide what quantity of each product to assemble for anticipatory stock. For example, the enterprise may decide to increase its production of product 46 by 20 units. The above options leave a lot of decisions to be made which are subjects of the production planning optimization.

The first usage is already modelled through the previous sets of equations, those related to the operational constraints and those related to the dealer network. In order to avoid having to express new sets of constraints, the third usage uses for the in-horizon forecast orders the same variable and constraint sets as the registered orders. The only difference lies in setting the parameters. For example, the revenue per unit can be weighted to express the confidence level in the forecast. A posteriori analysis of the solution is to show that some planned transports to dealer zones are to deal with registered orders while others are mostly potential transports delivering expected forecast orders. The shorter the planning horizon, the less important is to be the set of forecast orders to be added to the set of registered orders.

The constraint sets (12.35) to (12.43) below describe how the second and fourth usages are to be dealt with.

Future dealer demand uncertainty constraints.

$$I_{pt} = i^s_{pt} + I^{s+}_{pt} - I^{s-}_{pt}, \qquad \forall p \in P; \forall t \in T^a \tag{12.35}$$

$$C^{s-} = \sum_{t \in T^a} \sum_{p \in P} c^{s-}_{pt} I^{s-}_{pt} \tag{12.36}$$

$$I^{s+}_{pt^l} = \sum_{n \in N^{sf}_p} I^{sf}_{pn}, \qquad \forall p \in P \tag{12.37}$$

$$I^{sf}_{pn} \leq i^{sf}_{pn}, \qquad \forall p \in P; \forall n \in N^{sf}_p \tag{12.38}$$

$$C^{sf} = \sum_{p \in P} \sum_{n \in N^{sf}_p} c^{sf}_{pn} I^{sf}_{pn} \tag{12.39}$$

$$\sum_{t \in T^a} E_t = \sum_{u \in U} E^T_u \tag{12.40}$$

$$E^T_u \leq e^T_u U_u, \qquad \forall u \in U \tag{12.41}$$

$$U_u \leq E^T_{u-1} / e^T_{u-1}, \qquad \forall (u \neq 1) \in U \tag{12.42}$$

$$C^e = \sum_{t \in T^a} c^e_t E_t + \sum_{u \in U} c^T_u E^T_u . \tag{12.43}$$

Constraint sets (12.35) and (12.36) deal with the goal of maintaining target safety stocks for each product p at each time t. Set (12.35) simply contrasts the current inventory level of product p with its target and computes the positive and negative deviations from target. Constraint set (12.36) computes the cost associated with lower than targeted safety stocks. Larger stocks are not disruptive from a safety stock perspective and therefore are not penalized from this perspective. Setting the expected marginal cost for lower than targeted safety stock requires an a priori statistical and economical analysis.

The build up of inventory for dealing with anticipated demand beyond the planning horizon results in an inventory of the product at the last time slot in the planning horizon that is above the level required for safety stock purposes. Therefore the level reached by variable $I_{pt^l}^{s+}$ corresponds to the anticipatory inventory of product p. Constraint sets (12.37) to (12.42) serve the purpose of correctly balancing the compromise between using current production capacity for creating such anticipatory stocks and not using this current production capacity, thus avoiding production and inventory costs.

Consider the first unit of product 46 stocked for anticipation. How much does it cost to stock it? A deterministic answer is impossible since its usage depends on future demands which have not yet materialized. However it is easy to differentiate between two levels of stocked quantities. The first level includes units which are practically certain to be ordered by dealers before the end of the selling season. For example, if the future demand for product 46 for the entire season beyond the planning horizon is forecasted to behave according to a Normal distribution with an average of 300 and a standard deviation of 30, then the 200 first products stocked in anticipatory mode are almost certain to be ordered by dealers. At this first level, the only uncertainty lies in the timing of the eventual order. Based on the time phased forecasts, it is possible to compute the expected duration-of-stay of each additional unit in anticipatory stock.

The second level includes the stocked units that are beyond the practical certainty of eventual demand. At this level, there is a significant probability that a stocked unit may never be ordered by any dealer during the selling season. In the above example for product 46, the 300th unit stocked in anticipation has a 50% chance of never being ordered by a dealer during the selling season. If not sold, it would either have to be dismantled, sold to a bargain market, or kept in stock until the next selling season, to be sold at discounted price in competition with next year's products. Therefore the cost for stocking it must incorporate both the duration-of-stay factor and the probability-of-demand factor. Again, through statistical and economical analysis, the marginal cost of an additional unit in anticipatory stock may be computed at this second level. For practical purposes, in the product 46 example, the second level ends at about 400 units. Beyond that it makes practically no sense to produce any further product unit.

Given the above logic, the enterprise is to prepare a priori an anticipatory stock cost function for each product and to generate a piecewise linear approximation of this convex cost function. Constraint sets (12.37) to (12.39) implement this approximation. For each product, the speculative end-of-planning-horizon inventory is split in n_p^f linear cost segments. Each has a fixed unit cost c_{pn}^{sf} and a maximum allowed inventory i_{pn}^{sf}. The first segment has the lowest unit cost while the nth segment has the highest unit cost. Each segment n of every product p has its inventory variable I_{pn}^{sf} stating the current level of anticipatory inventory at time t. Constraint set (12.37) insures that the sum of the current anticipatory inventories associated with each segment for a product p equals the total anticipatory stock for that product, as expressed by I_{ptl}^{s+}. Constraint set (12.38) simply bounds the segment-specific anticipatory stocks not to exceed the specified maximum for that segment. Constraint set (12.39) cumulates the anticipatory stock costs for all segments of all products.

In the above constraint sets, stocking a product for anticipatory use beyond the production planning horizon has been considered to be a cost. There must be a balancing cost promoting the build up of such stock, otherwise the optimal solution to the problem will never construct such stock. This balancing cost is associated with lost capacity when nothing is produced during a time slot in the assembly center. Consider for example the first time slot in the planning horizon. It is clear that if nothing is planned to be produced during this time, then nothing will ever be produced during this time, and potential capacity will be lost forever.

How does one compute an expected value for this capacity? The answer is similar in nature as what has been explained for the expected anticipatory stock cost. Assume that when summing the forecasts for all products in all dealer zones, the global demand remaining beyond the production planning horizon is forecast to behave according to a Normal distribution with an average of 20,000 units and a standard deviation of 1,000. Now assume that beyond the current production planning horizon, there remains only 15,000 potential time slots in the assembly center, with no possibility to increase that number. Then it is clear that with almost certainty there is a remaining demand of at least 17,000 units. Thus there is lack of 2,000 time slots. If time slots are not found for producing them, then the result will be a loss of at least 2,000 sales and their associated margins. Assuming that the current planning horizon covers ten eight-hour days with a 3-minute takt time, there are 1,600 time slots during the current planning horizon. This means that each one of them not used for producing a product is practically certain to result in a lost sale and associated lost margin.

It is easy to set up another example depicting the other extreme situation where expected subsequent demand is to be so small that producing anticipatory stock cannot make any business sense, as is often the case late in the selling season.

In between these extreme situations it is possible through Monte Carlo simulations and statistical and economical analysis to generate an expected concave cost function for each additional assembly slot left empty during the planning horizon, not being used to assemble a product. The enterprise must then develop a piecewise linear approximation of this cost function. Constraint (12.40) computes the total number of empty time slots in the planning horizon and then redistributes this sum over all cost segments corresponding to the piecewise linear approximation of the concave cost function. Constraint set (12.41) simply states whether or not a cost segment u is used, with a greater than zero membership. Constraint set (12.42) insures the validity of the linear relaxation of the concave cost function by insuring that a lower numbered cost segment is fully used prior to allowing the opening of the next cost segment.

Constraint (12.43) adds up all the unused capacity costs over all cost segments. It also adds up a sum of unused capacity costs specific to each time slot. The reasoning for these costs is to factor in the fact that when using a rolling horizon with frequent re-optimizations, then among all time slots of a given planning horizon, the first slots are more costly to leave empty than the latest slots. The first time slots, if not planned to be used for production, have a probability of one of never being used, their potential capacity gone forever. So in a planning horizon, if a time slot is to be planned to be unused, it is preferable for that time slot to be among the last slots rather than the first slots. By computing an expected unused marginal cost differentiation among time slots, the enterprise is in a position to set cost parameters c_t^e for each time slot. The contribution of these costs is then summed in constraint (12.43), added whenever a time slot t is not planned to be used for production.

3. Production planning optimization in alternative demand and supply chains

This section examines the production planning optimization modeling of assembly centers in gradually more collaborative, customer-centric, agile and personalized demand and supply chains. It describes the implications of these transformations as well as of production planning optimization knowledge and technology.

3.1 Pushy and rigid demand and supply chain

The rigid demand and supply chain of Figure 12.1, in its pushy variant which does not take care of delivery timing preferences of dealers, still generically faces the entire production planning problem described in section two, except the portions dealing with the dealer network and with dealer demand uncertainty.

The assembly center(s) of such chains are generally designed from a mass production paradigm. They are efficient at assembling a product once setup for it. However the changeover from one product to the next can require highly significant time and generate significant costs. It is common to let a product run for days prior to switching to the next. In fact a product is often run only a few times, if not a single time, during the production season. The operations and changeover constraint sets (12.2) to (12.11) are therefore at the core of the production planning problem.

In order to size the complexity of the problem, assume that only these constraint sets are considered by the planning optimization, and that the enterprise imposes a single run per product. Then the problem reduces to the Traveling Salesman Problem, well known to be NP-complete, where a city becomes a product, the travelling distance between pairs of cities becomes the inter-product changeover cost (including the cost of lost time slots in the assembly center during the changeover), the traveling salesman becomes the assembly center, and the objective of touring all cities in minimal travelled distance becomes the objective of touring all products in minimal overall changeover cost. In the contexts studied in this chapter, it is common to deal with hundreds of products. Furthermore, the single run per product is an extreme solution which can be examined but is not to be a priori imposed except in very precise conditions guaranteeing its optimality. Such conditions involve a complete dominance of changeover costs and an absolute ignorance of dealer delivery preferences.

The sheer complexity and size of the overall problem helps understand why in most cases, the planning decisions result from a decomposition of the problem, from relying on decision rules and heuristics, and from heavy reliance on human intelligence.

Problem decomposition generically assumes away many constraint sets when generating the master assembly plan, which becomes an input to the other sub problems, according to a hierarchical planning strategy. Supply and personnel planning become subordinates of the master assembly plan resulting from the production planning optimization. There is heavy reliance on hierarchical decomposition in practice. Such decomposition makes sense when there is indeed a dominance of the operational and changeover variables in terms of feasibility and cost optimization. However the personnel and supply feasibility issues and costs are often significant.

As described in the introduction to personnel constraints sets (12.12) to (12.17) and supply constraints sets (12.27) to (12.30) and (12.34), there are intricate relationships between these and the operational and changeover constraints sets (12.2) to (12.11). An example stemming from the fact that such enterprises are often among the biggest employers in their region, is the signature of social or union contracts guaranteeing that the assembly center is to maintain as stable a workforce as possible, with a fixed bottom level and penalties for not achieving its engagements. This imposes constraints which affect the feasibility and profitability of production plans. Lack of in-depth knowledge of these relationships and their impact, coupled with the lack of adequate planning optimization technology to support the effective integrated treatment of the integrated problem formulation, are the key factors inhibiting their consciously integrated treatment.

3.2 Collaborative yet rigid demand and supply chain

When considering the more dealer-collaborative variant of the rigid demand and supply chain of Figure 12.1, then the problem statement needs to take explicit consideration of constraint sets (12.18) to (12.26). On the one hand, it transforms the problem from a cost minimization perspective to a profit maximization perspective since the timing and constitution of shipments to dealer zones have direct impact on the revenue stream. On the other hand, it forces to understand the implications of not satisfying dealer preferences on the profitability of the enterprise. These involve the following:

- Each dealer aims to have early on in his possession the set of products maximizing his expected showcasing effect, thus maximizing his customer attraction potential and his revenue expectations.
- Each dealer aims not to have product over stock in order to avoid both product financing and storage costs and minimizing security risks.
- Each dealer, when feeling badly treated by the enterprise, has higher probability to lose brand loyalty and switch to sell products of competitors, bringing with them a significant percentage of their customers, thus having long term impact on sales by the enterprise.
- Each dealer is a business unit that is a flagship of the enterprise in the region he deserves. The final customers interact with the enterprise

through the dealer. The relative prosperity of the dealer generally reflects on the perception of the enterprise by the final customer, thus affecting purchasing probability.

In practice, integration of dealer network considerations in the production planning optimization, when attempted, is generally limited to insuring a first release of top products for early showcasing effect and dealing with the most important dealers (or dealer groups) which have significant weight on the enterprise sales. Again lack of in-depth knowledge, resulting on misperceptions relative to the potential gains for the committed efforts, as well as lack of planning optimization technology, restrains enterprises to exploit more collaboration with dealers.

Even in the pushy variant where products are shipped to dealers as soon as produced, subject to localized vehicle transport optimization, enterprises should understand the impact of delivery schedule to dealers on the revenue stream. They can profit from integration of constraints (12.18) to (12.26), setting all delivery preferences as soon as possible and the deviation costs to zero. This would help them optimize revenues and minimize transport costs.

Collaboration can be established on both the demand and the supply sides (Montreuil et al., 2000). Exchange of status information with suppliers and subcontractors, coupled with the establishment and modeling of their key constraints and smoothing objectives, lead to generate constraint sets (12.31) to (12.33) and to refine cost constraint (12.34). This collaboration allows suppliers to not protect themselves with extensive lead times and high minimal quantities, knowing that the enterprise takes explicit care of its constraints, costs and objectives in its production planning optimization. By systematically developing the collaboration with all critical suppliers and subcontractors, the production plan has the potential to reach levels of profitability not attainable without collaboration.

3.3 Customer-centric, agile and collaborative demand and supply chain

In between the demand and supply chains of Figures 12.1 and 12.2 lies the potential for increasing the agility and customer-centricity of a collaborative demand and supply chain organized such as in Figure 12.1, yet offering continuous ordering throughout the selling season.

In the context studied in this chapter, extreme agility involves:

- No significant changeover times and costs;
- Highly scalable operations allowing up and down swing in production level with negligible costs and constraints;

- Highly polyvalent workforce with negligible constraints and costs on manpower level modification;
- Insignificant supply lead time and constraints from suppliers and subcontractors;
- Non constraining transportation capabilities.

When extreme agility is achieved, most of the constraints melt away and production can be operated with minimal planning, mostly in a sense and respond model, producing just-in-time in pull mode, keeping minimal safety stocks to allow instant delivery when required. In such a case the operations are in a perfect position to enable the enterprise to be customer-centric, delighting both dealers and customers.

The problem lies in the fact that in most cases some degree of agility is reached, yet many constraints remain. Therefore all constraint sets (12.2) to (12.34) may be potentially needed, but applied to a smaller number of entities (e.g., lower number of modules and suppliers) and with more limited impact (e.g., smaller number of worker types, each capable of performing a wider scope of tasks).

Customer centricity involves putting the goal of delighting customers at the forefront of the business preoccupations, meeting or exceeding their expectations in terms of product offering and availability, delivery speed and price, anytime prior or during the main selling season. The most direct implication is the importance of allowing dealers to order throughout the selling season to respond to customer demand, rendering impossible the termination of the production season prior to the main selling season. High availability of all products requires the maintnance of a safety stock and to be rapidly able to replenish stock in light of consumed demand by dealers. It also may involve the building of anticipatory stocks ahead of time in order to maintain high availability and fast delivery throughout the season peaks.

Such a demand and supply chain has to dynamically update its production plan, perhaps every day depending on the market dynamics. It also has to recognize explicitly that it deals with uncertain demand. These facts lead it to have to face such constraints as those in sets (12.35) to (12.43). In such a setting, the production planning horizon may, for example, be set to a few weeks. During these weeks, the production plan is precisely optimized, setting the production assignment to each time slot in the assembly center. The production plan optimization attempts to best decide on what products to produce at each time or yet when not to produce anything or start a changeover to another product. As stated in Section 2.7, it must choose between fulfilling actual and forecast orders during the planning horizon, amplifying the safety stock of selected products, or building anticipatory stocks to deal with future demand beyond the production planning horizon.

3.4 Personalized, customer-centric, collaborative and agile demand and supply chain

The demand and supply chain outlined in Figure 12.2 adds personalization to the characteristics described in Section 3.3 above. This means that instead of imposing a fixed product mix to the customers, often composed of a few hundred products, the enterprise lets customers personalize products according to their needs and tastes. Section 1 explicitly states the personalization offer in the illustrative example. According to personalization framework introduced by Montreuil and Poulin (2004), it combines popularizing, accessorizing and parametering. The popularizing option aims to offer a limited set of popular products to be highly available on the shelves so as to satisfy the needs of customers wanting their product *right now*. The accessorizing option offers a series of ready-to-personalize products coupled with a variety of accessory modules allowing each customer to personalize his product. The parametering option offers personalized products defined by the customers through parameter and option settings.

In order to deliver the personalization offer, the demand and supply chain of Figure 12.2 is transformed by the addition of fulfillment centers and the specialization of the main assembly center. The assembly center is to produce popular products, ready-to-personalize products and parametered products. The fulfillment centers assemble accessorized products from ready-to-personalize products and sets of accessories. Both types of centers are agile and highly polyvalent personnel, requiring no changeover time when switching from a product to the next, yet they both have limited capacities. Both types collaborate with both their demand side and supply chain partners.

From a production planning perspective, the fulfillment centers cannot produce an accessorized product without a customer having actually ordered that product. Therefore their planning horizon corresponds to the time required to go through the order booking. The key decisions involve the sequencing of products to be accessorized, mostly to best compromise between overall manpower smoothing, supply availability and cost from suppliers and from the assembly center, and delivery promises and transport to dealers.

Relative to the fulfillment centers, the production planning optimization problem has the following characteristics:

- Operational constraint sets (12.2) to (12.4), (12.9) and (12.10) are enforced, with no consideration for changeovers. All changeover constraints are not to be enforced.
- Personnel constraints (12.12) to (12.17) are imposed, yet with a lower number of worker types.
- Dealer network constraints (12.18) to (12.26) are still necessary, yet the proximity to market of the fulfillment centers shortens the trip times to dealer zones. Also, dealer zones for fulfillment centers are to be much smaller than in the case of Figure 12.1.
- Supply network constraints (12.27) – (12.34) are to be imposed, maximally exploiting collaboration with partners, on one side with accessory suppliers and on the other side with the assembly center.
- Even though demand is uncertain and planning is to be dynamically updated through a rolling horizon, constraint sets (12.35) to (12.39) can be discarded since there is no possibility to stock not yet ordered products. Constraint sets (12.40) to (12.43) may be imposed to ensure that leaving an empty production slot is to be cost adequately through the production planning optimization.

In this personalized setting, the clients of the assembly center have become the fulfillment centers rather than the dealers themselves. Even though it has to wait for an actual order from a fulfillment center to produce a parametered product, it can produce popular products as well as ready-to-personalize products prior to getting actual orders for them from fulfillment centers. Therefore its planning horizon is to be longer than the planning horizon of fulfillment centers. The production planning optimization has the following characteristics:

- Operational and personnel constraints (12.2) to (12.4), (12.9), (12.10) and (12.12) to (12.17) are to be imposed as in the fulfillment centers.
- The dealer network constraints (12.18) to (12.26) become the fulfillment center network constraints. The zones correspond to a single fulfillment center.
- Supply network constraints (12.27) – (12.34) are to be imposed, maximally exploiting collaboration with suppliers.
- Constraint sets (12.35) – (12.43) are to be imposed as described in Section 2.7, replacing dealers by fulfillment centers.

For both the fulfillment centers and the assembly center, collaboration with very agile suppliers and subcontractors can permit to take them out of the production planning optimization model since they are not constraining. Fast and accurate information transfer with them is then sufficient for them to supply the centers with the required modules in time for their assembly into the products being manufactured. Similarly, in the case of suppliers having very limited influence on the production plan optimality, the supply relationship can be decoupled, operated in pull mode using a dynamically updated kanban system insuring sufficient stock to avoid shortages.

4. Conclusion

First the chapter has introduced a spectrum of demand and supply chain alternatives for high value consumer products, varying in terms of collaboration, customer centricity, agility and personalization. Second it has introduced a comprehensive production planning optimization enabling to adequately model these alternatives. Third it has provided a thorough analysis of production planning optimization modeling as a demand and supply chain is transformed to incorporate more collaboration, customer centricity, agility and personalization. It has also discussed the impact of production planning optimization knowledge and technology.

It is shown that some of the transformations increase the complexity of the model to be solved while some others decrease this complexity.

- Collaboration increases model complexity, yet this added complexity allows to achieve better global optimization.
- Agility generally decreases model complexity by removing constraints and imposing less restrictive parameters. It generally leads to improved global optimization.
- Customer centricity has a double effect. On one hand it decreases model size by switching from rare optimization using long planning horizon to frequent optimization using a shorter planning horizon. On the other hand it increases model complexity in order to deal adequately with the inherent market uncertainty involved in attempting to delight customers through the dealers.
- Personalization generically increases modeling complexity both due to the explosive product scope and the required structural transformation of the demand and supply chain. However some centers end up with more simple models due to the fact that they end up producing strictly to order.
- Lack of production planning optimizaтion knowledge generally results in lower model complexity, and mostly in model inadequacy, through the ignorance or inadequate representation of constraints and cost factors. Entire panes of modeling can end up ignored or delegated to being imposed production planning decisions. It results in potential for lower global optimization.
- Lack of available adequate production planning optimization technology results in problem decomposition in order to avoid having to deal

with the overall complexity of the global problem. This inherently leads to global sub-optimization.

The chapter opens many avenues for further research. Below are listed a few promising avenues for the research community:

- The introduced comprehensive production planning model for assembly centers in demand and supply chains is not yet solved in an integrated manner, either to global optimization for small cases or heuristically for larger cases.
- The impact of networked collaborative and decomposition approaches to dynamically address production planning of assembly centers in demand and supply chains is not empirically studied.
- The impact of agility, customer centricity and personalization on production planning of assembly centers in demand and supply chains is not empirically studied.
- Simulation technologies are needed to enable the efficient realization of empirical studies as stated above, coupling optimization and stochastic system dynamics, and enabling adequate representation of all stakeholders in the demand and supply chain, from customers and dealers to production planners and suppliers.
- The current model should be expanded to integrate engineering issues related to new product introduction, yearly product evolution and on going engineering changes.
- The type of research reported in this chapter should be performed for other important demand and supply chain contexts.

The chapter brings to light important insights for the professional community:

- Production planning of assembly centers in demand and supply chains is a complex optimization problem having major financial and feasibility impacts. It lies at the core of a center contribution to the enterprise performance and should be addressed accordingly.
- Transformations in the demand and supply chain have significant impacts on the production planning problem formulation. They alter constraints, cost and operational parameters, the degrees of freedom, and the essence of the objective function. Overall they affect its complexity, its size and its profitability potential.
- Emphasis should be put on adequate training of managers and planners, making sure that they have the knowledge and in-depth understanding required to adequately address production planning in such contexts. This is *not* a minimal impact problem to be solved by untrained, unprepared, under equipped administrative personnel.

- Current and proposed production planning optimization technology should be carefully audited as to the degree with which it supports an adequate comprehensive modeling and solution. The technology should fit the needs. Gaps may have significant impacts on global optimization and enterprise performance.

Acknowledgements This research has been supported by the Canada Research Chair in Enterprise Engineering. Special thanks to Caroline Cloutier and Marc Poulin for valuable editorial help.

References

Hoover, W.E., Jr., Eloranta, E., Holmström J., and Huttunen, K. (2001). *Managing the Demand-Supply Chain: Value Innovations for Customer Satisfaction.* Wiley.

Montreuil B., Frayret J.-M., and D'Amours, S. (2000). A strategic framework for networked manufacturing. *Computers for Industry*, 42:299 – 317.

Montreuil, B. and Poulin M. (2004). Demand and supply network design scope for personalised manufacturing. Forthcoming in *International Journal of Production Planning and Control.*

Poulin, M., Montreuil, B. and Martel, A. (2004), Implications of personalization offers on demand and supply network design: A case from the golf club industry. Forthcoming in*European Journal of Operational Research.*