Sara Sofia Baltazar Martins

U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO
DEPARTAMENTO DE ENGENHARIA E GESTÃO INDUSTRIAL

Retail Distribution Planning to Brick-and-Mortar Stores

Submitted to Faculdade de Engenharia da Universidade do Porto in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Industrial Engineering and Management, supervised by Bernardo Almada-Lobo, Associate Professor of Faculdade de Engenharia da Universidade do Porto, and Pedro Amorim, Assistant Professor of Faculdade de Engenharia da Universidade do Porto

Department of Industrial Engineering and Management
Faculdade de Engenharia da Universidade do Porto
2018

This research was partially supported by the PhD grant SFRH/BD/102013/2014 awarded by the Portuguese Foundation for Science and Technology.

“Põe quanto és no mínimo que fazes”
Ricardo Reis

Acknowledgments

First and foremost, I would like to thank my advisers Bernardo Almada-Lobo and Pedro Amorim. I always could count on their support during this big journey. Bernardo always knew the right words and the right time to push me further. He has the characteristic of challenging me to be better, always believing in me. Pedro is an inspiration to work with. He is patient and always tries to be available for anything that we need. His efficiency and commitment with high quality standards inspire me everyday to be a better professional. I will be forever grateful to both of them for the opportunity they gave me and for everything I learned with them.

I would like to express my gratitude to all the co-authors of the articles related to this thesis for the valuable suggestions, advises and for helping me to explore this research topic. I am also grateful to all my research colleagues from the CEGI open-space for the very nice and friendly atmosphere. I was lucky to be part of this amazing research group, where we discuss our work and support each other when necessary. A special thanks to Marias, Elsa, Teresa, Beatriz and Alvaro for our talks and for all the strength they gave me during this journey. Also, a warm thanks to my dear friend Carolina, who was always available to listen to me and support me when I needed.

Finally, I would like to acknowledge the commitment of my family to my well-being. Mãe, Avó and Carlos thank you for all your patience with me, particularly when I arrived home in a bad mood and you always tried to comfort me. Thank you Pai and Mãe for always being there for me, believing in me and pushing me to be better. Thank you Helder, my love, for your unconditional support and patience.

Abstract

An efficient supply chain is essential for the success of any retailer and distribution planning, in particular, plays a central role to that end. The retail industry has suffered changes throughout the years, along with those in the market dynamics. These dynamics are turning the distribution process more and more challenging. In such a context, with multiple possible strategies and resources available to perform distribution, defining the most efficient way to supply each store is a critical task.

This research intends to give new insights and contribute to the current literature regarding the distribution process of retailers, specifically for the pharmaceutical and grocery retail sectors. The objective of this applied research is threefold: (1) frame the multiple ways of efficiently supplying retail stores; (2) formulate mathematical models for different challenges that retailers face in the distribution process; and (3) develop suitable solution approaches to solve the problems efficiently.

Firstly, the planning decision that selects the different distribution strategies of supplying the stores is explored. The definition of this problem, named delivery mode planning, is not clear in the literature, and a novel definition for the problem is proposed. The features and (dis)advantages of the different product flows are described, together with the practical implications between different distribution decisions.

Secondly, the network redesign of pharmaceutical wholesalers is analyzed. The goal is to respond to more frequent and smaller orders, while staying competitive and not jeopardizing the current customer service level. While the network design has been extensively studied in the literature, few works consider the redesign variant, i.e. changing facilities location, and take into consideration the response time indicator, besides costs. Nevertheless, the response time is a key driver for wholesalers. To tackle the problem a two-phase optimization-simulation approach that incorporates the trade-off between facility and transportation costs and customer service level is developed.

Finally, the incorporation of multi-compartment vehicles in the grocery distribution planning is studied. In the last years, multi-compartment vehicles, which enable the physical separation of products/segments during transportation have been deployed, and increasingly used by grocery retailers to perform mixed-product distributions. In the literature, there are already different works analyzing the impact of performing routing plans with multi-compartment vehicles instead of single-compartment vehicles. However, most of the research is focused on fuel and waste applications, which impose different business requirements from those of the grocery distribution. The complexity of the loading/unloading process with these vehicles has never been studied. Therefore, an extension of the routing problem considering the impact of loading constraints is proposed. Furthermore, as grocery stores usually define preferable time-windows to indicate when deliveries should occur, and this decision impacts the possible mixed distributions, the routing problem is also extended to incorporate time-window assignment decisions. For both problems, extensions of metaheuristics proposed for the routing problem are developed.

Resumo

Uma cadeia de abastecimento eficiente é essencial para o sucesso de qualquer retalhista e o planeamento da distribuição desempenha um papel importante nesse sentido. Ao longo do tempo, o retalho tem sofrido mudanças e presenciado novas dinâmicas de mercado que aumentam os desafios do processo de distribuição. Neste contexto, onde existem múltiplas estratégias e recursos que podem ser usados na distribuição, a decisão de como abastecer cada loja de forma mais eficiente torna-se uma tarefa crítica.

Esta dissertação pretende explorar e contribuir para a literatura relacionada com o processo de distribuição dos retalhistas, em particular nos sectores farmacêutico e alimentar. Esta investigação aplicada tem três principais objectivos: (1) enquadrar as múltiplas formas de abastecer as lojas eficientemente; (2) formular modelos matemáticos para diferentes desafios que os retalhistas enfrentam no processo de distribuição e (3) desenvolver métodos de solução capazes de resolver os modelos eficientemente.

Primeiro, exploram-se as estratégias de distribuição a usar no abastecimento às lojas. A definição deste problema, designado de planeamento do modo de entrega, não é clara na literatura e, consequentemente, é proposta uma definição formal. As características e (des)vantagens dos diferentes fluxos de distribuição são descritas, assim como as implicações práticas relacionadas com outras decisões de distribuição.

Segundo, o redesenho da cadeia de abastecimento de um grossista farmacêutico é analisado. O objectivo é adaptar a distribuição para responder a pedidos mais frequentes e de menor dimensão que as farmácias fazem, sem prejudicar o nível de serviço e mantendo a competitividade. Enquanto o desenho da rede tem sido muito estudado na literatura, poucos trabalhos consideram a variante do problema relativa ao redesenho, isto é, a alteração da localização das instalações. Adicionalmente, é rara a consideração do tempo de resposta como indicador. Para os grossistas, para além dos custos, este é um factor-chave. Uma metodologia de optimização-simulação é proposta para resolver este problema, incorporando o *trade-off* entre os custos relacionados com as instalações e transporte e o nível de serviço ao cliente.

Finalmente, o uso de veículos multi-compartimento na distribuição alimentar é estudado. Nos últimos anos, este tipo de veículos, que permite a separação física dos produtos/segmentos durante o transporte, têm vindo a ser desenvolvidos e usados por retalhistas para distribuirem múltiplos produtos. Já existem na literatura alguns trabalhos que analisam o impacto de utilizar estes veículos no planeamento das rotas, em deterimento dos veículos simples. No entanto, a maioria da investigação é focada nos sectores do petróleo e resíduos urbanos, cujos requisitos de negócio são diferentes da distribuição alimentar. A complexidade dos processos de (des)carregamento com veículos multi-compartimento nunca foi estudada. Assim, é proposta uma extensão do problema de roteamento para incorporar restrições de carregamento. Adicionalmente, as lojas alimentares normalmente definem as horas do dia em que preferem receber cada tipo de produto. Como estas decisões afectam as possíveis combinações de distribuição mista de produtos, o problema de roteamento é também extendido para incorporar decisões sobre a definição das horas de entrega. Para ambos os problemas são desenvolvidas extensões de meta-heurísticas propostas para o roteamento.

Contents

Part I

Introduction

Chapter 1

Motivation and Framework

Supply chains (SCs) are networks of organizations linked to provide value to final customers. Depending on the industry, a SC can be characterized by distinct suppliers, manufacturers, distributors and retailers, as shown in Figure 1.1. Due to their downstream position in the SCs, retailers connect with final customers, providing them the products and services they need and desire (Sullivan and Adcock, 2002).

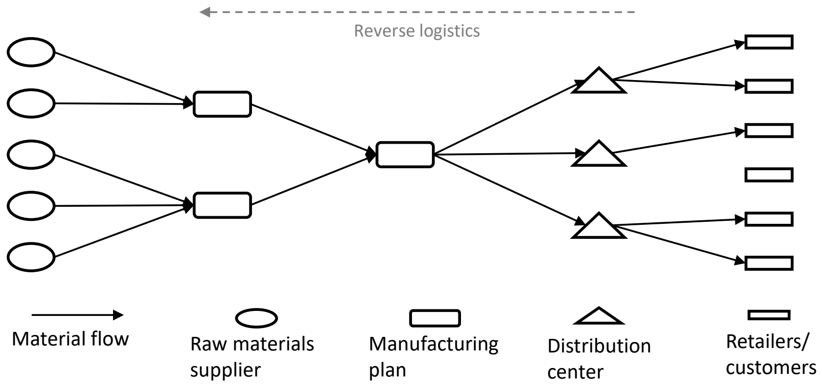


Figure 1.1: Illustrative supply chain

In the past, retailers were only simple stores receiving passively products from manufacturers according to their demand forecast. However, since the late 1990s the retail sector has been evolving in order to control, organize and manage the flow of products. Fernie et al. (2010) describe six trends that led to these transformations: (i) the first is the expansion to have their own distribution centers (DCs) in order to better control the distribution to the stores. (ii) the second is the use of different distribution strategies such as the joint distribution of different products and the centralization of slow movers. (iii) another growing trend is the adoption of cross-docking strategies with the aim of reducing inventory levels and response times. Moreover, to achieve new competitive edges some retailers are (iv) integrating their distribution process with the production phase, as well as (v) fostering reverse logistics activities. (vi) More efficient SCs have also been developed due to the growing collaboration between retailers and suppliers. Rushton et al. (2014) recognize that the retail sector has made some of the most advanced and innovative developments in distribution and SC thinking.

Due to its competitive nature, retailers have always to be aware of the customers needs and rethink their strategies in order to respond efficiently. In fact, retailers are turning their attention to customer experience (Bäckström and Johansson, 2006; Grewal et al., 2009).

For instance, nowadays they offer them service through different channels (e.g., traditional stores, catalogue, online shopping, click and collect) to respond to customers needs and desires (Sullivan and Adcock, 2002). This access to different channels is altering the customer behavior, but as mentioned by Mou et al. (2018) the brick-and-mortar (B&M) stores remain the main channel for shoppers. For this reason, the operational activities at B&M stores must be as efficient and effective as possible. In this traditional channel, retailers manage B&M stores with different formats, sizes, assortments and sales volume. The in-store planning decisions, such as assortment planning, space management and inventory management, are heavily influenced by the customer buying behavior. These decisions are directly perceived by the customer, as they define the type and amount of products available at the retail stores and how they are displayed for shopping. Nevertheless, the distribution management also plays an important role in the retailers SC efficiency and efficacy as it is responsible for supply and support the stores. The stores cannot operate efficiently if the products are not delivered on the right time, on the right amount and on the right conditions.

Distribution management deals with the physical flow and storage of products throughout the retail network until the customer (Rushton et al., 2014). It can be divided in three levels regarding the planning horizon: strategic, tactical and operational. The strategic level concerns long-term decisions regarding the SC network design (SCND), which defines the physical distribution structure. The works on network design are mostly focused on storage facilities, defining the number, size and function of the DCs. Higher number of facilities allows the reduction of the transportation costs of the network, but at the same time it increases inventory costs. However, the increase in the usage of cross-docking strategies called the attention of researchers (Van Belle et al., 2012). Therefore, SCND should integrate network configuration decisions with inventory and routing problems, as neglecting these tactical decisions leads to sub-optimality. Mid-term decisions are tackled at the tactical level and determine the main rules for order fulfillment, such as the delivery mode, the delivery pattern and the selection of transportation means (Hübner et al., 2013). The delivery mode planning determines how the products should flow from the DCs to stores. Besides direct-shipment distribution, cross-docking and milk-run modes can be used to achieve economies of scale in transportation. The frequency and the specific days of delivery to a store are defined by the delivery pattern, which is related to an inventory-routing problem, as more frequent deliveries allow to reduce the inventory level at the stores. The transport between the retailers DCs and the stores is mainly performed by trucks with different characteristics and technologies. Finally, the operational level tackles short-term decisions related to the transportation planning, such as the loading and routing of vehicles. Distinct variants have been proposed in the literature for these problems (Toth and Vigo, 2014; Pollaris et al., 2015).

This research is focused on some of the new challenges the pharmaceutical and grocery retail sectors face and aims to understand how retailers can adapt their distribution activities to remain competitive. These sectors have the peculiarity of managing a cold supply chain, which entails additional refrigerated costs, and having to adhere to tight legal regulations. However, each sector has different characteristics and markets that impose different challenges. Therefore, this research aims at addressing few of the new problems that are emerging in each sector, for which very few works have been proposed. These problems include the redesign of the supply chain network and the distribution planning with multi-compartment vehicles.

Regarding the pharmaceutical sector, due to the economic crisis many pharmacies are struggling to survive and are changing their purchasing behavior towards smaller and more frequent orders, increasing the burden of the distribution system. To adapt to this new shift of demand, the wholesalers, who are responsible for the distribution to the retail stores, need to rethink their distribution strategy in order to stay competitive while securing their service level. The service level can be measured not only in terms of the products availability, but also regarding the response time. The problem of ensuring a certain level of product availability has been extensively studied in the inventory management literature stream, and the level of response time is related to the network design problems. However, although different works have been proposed for the network design problem, few of them consider the redesign variant, i.e. changing the location and function of the DCs, and the impact it yields on the response times. This is relevant in the pharmaceutical industry as different wholesalers can respond to the pharmacies demand due to the null changeover cost for switching wholesaler, being the response time one important key driver for selection.

Concerning the grocery sector, retailers have to manage simultaneously products with different temperature requirements (e.g., frozen, chilled and ambient). In the last years, vehicles with new technologies that enable the physical separation of products during transportation have been deployed. Such vehicles are called multi-compartment vehicles (MCVs) and are being increasingly used by grocery retailers to perform mixed distributions. The use of MCVs poses new challenges on distribution planning. Most of the literature regarding distribution with MCVs focuses on its applications to the fuel distribution and waste collection. Few works address the food distribution and are mainly targeting the routing problem, with few extensions developed up to date. Due to the complexity of a grocery retailer SC, with products requiring specific temperature conditions, the loading and unloading of the products in an MCV is not as simple as in the fuel and waste application. The rear-loading of the vehicles poses great difficulty to access each compartment of the vehicle. Moreover, while previously the flow of products with distinct transportation requirements had to be distributed independently, now with the use of MCVs mixed distributions can be planned. This means that the full range of products can jointly be delivered to a store in a single visit. This is an advantage that should be taken into account when planning the distribution flow. Additionally, the definition of delivery patterns and time-windows can be heavily influenced by the products flow. If products are not assigned to the same day and time of delivery, they cannot benefit from a joint distribution. While the definition of delivery patterns for food distribution is tackled by some works, no work on time-window assignment with MCVs has been developed.

Summarizing, this research intends to give new insights into the planning decisions of the distribution process to retail B&M stores, focusing on the challenges previously described for the pharmaceutical and grocery sectors. However, before tackling those specific problems, the different ways of supplying retailers B&M stores need to be identified and understood. Regarding the pharmaceutical sector we focus on the tactical-strategic planning problem of wholesalers network redesign. Finally, regarding the grocery retailing, we study the use of MCVs, focusing on the operational routing problem with loading constraints and the operational-tactical problem of routing and time-window assignment. Figure 1.2 summarizes the scope of this thesis.

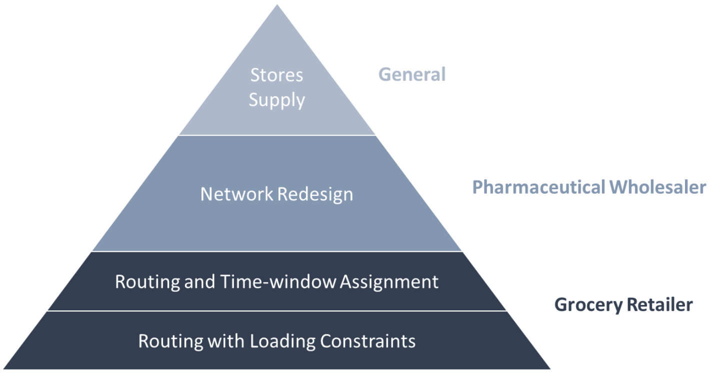

Stores Supply
General
Network Redesign
Pharmaceutical Wholesaler
Routing and Time-window Assignment
Grocery Retailer
Routing with Loading Constraints

Figure 1.2: Scope of the thesis by sector perspective

The remainder of this chapter is divided in three sections. Section 1.1 presents the research questions and objectives of this thesis. Then, Section 1.2 gives guidance to the reader on the organization and subjects of the following chapters. The main contributions of this work and future research directions are discussed in Section 1.3.

From hereafter B&M stores are described just by the term *stores*, since online stores are not within the scope of this research.

1.1. Research Objectives

This research intends to give new insights and contribute to the current literature regarding the distribution process of retailers, specifically for the pharmaceutical and grocery retail sectors. The objective of this applied research is therefore threefold: (1) frame the multiple ways of efficiently supply retail stores and understand the interdependencies between different planning decisions related to the distribution process (namely, network design, means of transportation and delivery pattern); (2) formulate mathematical models for few challenges retailers face regarding the distribution process and (3) develop suitable solution approaches to solve the models efficiently. These objectives are linked with the following research questions.

Research question 1:

How can products flow through the supply chain network of a retailer?

The distribution between DCs and stores can be planned according to different strategies. The planning decision that tackles this problem is not clearly defined in the literature, with different terms used with the same purpose and different works tackling the same problem with different definitions. To answer this research question a comprehensive review needs to be conducted to frame the multiple ways retail stores can be supplied. However, besides reviewing the works on the problem, it is imperative to identify and compare the characteristics and (dis)advantages of the different product flows, together with the practical implications between different distribution decisions.

After understanding the multiple ways retail stores can be supplied through the SC network, this research focuses on two main topics motivated by the six trends identified before: (1) the redesign of a SC network in order to respond to the pressure of more frequent and smaller deliveries and (2) the extension of the distribution planning decisions to incorporate the use of MCVs.

Research question 2:

How can pharmaceutical wholesalers analyze the redesign of their network in order to respond to more frequent and smaller deliveries, while staying competitive and not jeopardizing the current customer service level?

The SCND is a strategic decision that usually lasts for a long term. However, when the market needs change, the network design might become disadjusted and may not be able to respond efficiently to the business needs. Hence, in some occasions, redesigning the SC sooner than predicted is the most cost-effective solution.

When designing a network, the location of the facilities and how the products and customers are assigned to them have to be defined in order to secure a certain customer level at the minimum cost. However, when a network redesign is required, the company already has facilities operating and serving customers. In this case, the decisions fall on which facilities should be closed and were to position new ones, without jeopardizing the current service of the customers. Although the SCND problem is extensively studied in the literature, few works consider the redesign variant, specially from a wholesaler point of view for which the response time is important. Different works refer the importance of integrating the inventory and routing problems in the SCND, but few take into consideration the response time indicator. In the pharmaceutical sector, pharmacies place multiple orders along the day and place them to the wholesaler they know will deliver them in the shortest time.

In order to be able to answer this research question we first need to understand the distribution process of a pharmaceutical wholesaler. Afterwards, a mathematical model that frames the process has to be formulated and a solution approach developed. The approach has to consider different distribution strategies, the trade-off between inventory and transportation costs and the impact on the response time.

Research question 3:

MCVs are new technological vehicles that are able to split flexibly their loading area into different compartments. Grocery retailers are starting to use this type of vehicles, which allow them to fully supply a store in a single visit instead of using separate single-compartment vehicles to transport products at different temperatures.

In the literature, there are already different works analyzing the impact of performing routing plans with MCVs instead of single-compartment vehicles. However, most of the research is focused on the fuel and waste applications, which impose different challenges from those of the grocery distribution. There are some works on grocery distribution with MCVs, but they only target the routing decisions and analyze the operational impact of their use on DCs and stores. The complexity of the loading/unloading process with these vehicles has never been studied. Furthermore, few extensions of the correspondent routing problem, the multi-compartment vehicle routing problem (MCVRP), have been proposed. Therefore, this third research question concentrates on these two gaps and is thus divided in the following research questions.

Research question 3.1:
What is the impact of considering loading constraints in the multi-compartment vehicle routing problem of grocery retailers?

Due to legal regulations, the grocery retail sector has to comply with a careful control of the products temperature throughout the whole SC. Therefore, it is important to understand the distribution challenges that arise when switching from single-compartment to MCVs. Because compartments can be flexibly built in front of each other and multiple compartments might need to be accessed in one stop, loading constraints have to be identified and analyzed. The routing problem must secure an efficient loading and unloading process, respecting the regulations and not jeopardizing the products quality. Distinct routing problems with loading constraints have been proposed in the literature, but none of them was applied to MCVs in the grocery distribution. By answering this research question, we expect to identify the situations where the distribution planning is more critical, developing a model and solution approach that helps defining how the vehicles can be loaded efficiently when defining the routing plan with MCVs.

Research question 3.2:
What is the impact of planning consistent deliveries to grocery stores, using multi-compartment vehicles?

When products are delivered in a store they can be either immediately dispatched to the sales area for replenishment of the shelves or they can be moved to the backroom to be stored according to the required temperature. As the stores personal perform different in-store operations, the receiving process should be scheduled in advance. Grocery stores usually define preferable time-windows to indicate when deliveries should happen. The stores have to know when the products will be delivered in order to coordinate and smooth their in-store operations. The days in which the store receives goods are defined by the delivery pattern problem, which is an inventory-routing problem variant that has been studied in the literature. However, and to the best of our knowledge, few works have been proposed regarding the time-window assignment, and none with an assignment oriented to products instead of just to the store. To answer this research question, a new model and solution approach able to assign product-oriented time-windows to stores, incorporating the use of MCVs that leverages the joint deliveries of products, needs to be developed.

1.2. Thesis Structure and Synopsis

The main chapters of this thesis consist of a collection of papers aligned with the research objectives previously described. Each chapter is focused on one research question. In the remainder of this section we overview the main subjects covered in the following chapters and their main conclusions.

Chapter 2 frames the multiple ways to supply retail stores, targeting at research question 1. It also identifies the delivery mode planning as the decision that defines how products should flow throughout the SC and provides a discussion and literature review on the topic. As this planning decision is not well defined in the literature, a formal definition is proposed and the distinct delivery mode types are described (direct-shipment, cross-docking and milk-run). The main interdependencies between the delivery mode planning and the SCND, the selection of the means of transportation and the definition of delivery patterns are identified and analyzed. These problems have great impact on the selection of delivery mode. In this chapter, we conclude that most of the works on delivery mode planning considering hybrid networks, i.e. the execution of more than one type of delivery mode, are mainly focused on direct-shipments and cross-docking deliveries. In addition, the majority of the literature does not incorporate the routing problem, as well as the operational costs and constraints related with facilities. However, this is still a growing research area, which has recently received more attention with the development of cross-docking strategies and the deployment of new technologies.

Chapter 3 focuses on the network redesign problem of pharmaceutical wholesalers. To tackle the problem a two-phase optimization-simulation approach that incorporates the trade-off between facility and transportation costs and customer service level is developed. In the first-phase, at a strategic-tactical level, the network redesign is optimized by means of a mixed integer programming model. In this model, the number, location, function and capacity of the warehouses, as well as customer allocation, are optimized in order to minimize the total costs. In the second-phase, the operation of the SC is simulated using a discrete event simulation model to analyze the impact of the new network design from an operational perspective, specifically the service level. The analysis is performed based on a case-study with an European pharmaceutical wholesaler. The computational experiments expose the benefits of the hybrid approach, demonstrating that the optimization model provides insights on the best network redesign. The practicability of the selective implementation is grasped by the simulation model. This Chapter helps to answer research question 2.

In Chapters 4 and 5 the distribution planning of grocery retailers that use MCVs is analyzed, attempting to answer research question 3.1 and 3.2, respectively. In both chapters, a description of the distribution process of grocery retailing is provided. In order to evaluate individually the impact of the loading constraints and the consistent deliveries in the grocery distribution, two extensions of the MCVRP are tackled independently.

Chapter 4 studies the impact of the loading constraints when using MCVs. First, a description of the loading and unloading processes in grocery retailing is provided, identifying the complexity of the processes and the reason why the loading constraints should be analyzed. A mathematical model formulation for this new problem is proposed. Afterwards, a specialized packing problem is developed to define how the compartments should be built inside the vehicles in order to perform efficient loading and unloading processes, with no deterioration of product quality. With the purpose of evaluating the impact of loading constraints, both a Branch-and-Cut algorithm (B&C) and a Large Neighborhood Search (LNS) solution approach are developed, considering the packing problem within the routing problem to check and repair its loading feasibility. It is shown that loading constraints matter even for small instances problems and the number of unfeasible solutions increases as the problem sizes increase, being highly dependent on the available temperature setups and the corresponding order structure. However, a feasible loading can often be achieved by only minor changes on the routing solution and therefore with limited additional costs. Results shown that an ex-post repair of the MCVRP solution could already provide a good feasible solution in short-time. Nevertheless, performing feasibility checks during the search can in some cases lead to improved solutions, at the cost of higher computational effort. Computational experiments with an European grocery retailer case-study are also analyzed, showing that this is a relevant problem in practice and confirming the previous conclusions.

In Chapter 5, the characteristics of a consistent distribution planning to grocery retail stores (including practical considerations) are investigated. A new problem is proposed, extending the MCVRP by considering a multi-period setting with a product-oriented time-window assignment. Here, time-windows are defined for each product segment/store, taking into account the possibility given by MCVs of delivering the full product range jointly or separated. An Adaptive Large Neighborhood Search (ALNS) is proposed to solve the problem with daily operators, focusing on the improvement of routing aspects of the problem on each day, and weekly operators developed to align the time-window assignment consistently throughout the planning horizon. The effectiveness of the approach is tested on benchmark instances from literature. Numerical experiments demonstrated that planning a consistent distribution leads to better solutions than an ex-post time-window assignment of daily plans. Although the overall cost improvement can be small, a consistent planning provides more on-time deliveries. The impact analysis of performing deliveries outside the time-windows bounds is important in this context, because it can result in higher cost related to spoilage or stock-out, among others. Furthermore, if products are not assigned to the same time-window, they cannot be assigned to a joint distribution. Therefore, in this case, either the routing costs or the penalty costs increase.

1.3. Contributions and Future Research

In this section, the literature contribution and the references where the research of each chapter was originally published or submitted are identified, along with the main contribution of the PhD candidate to each of the works. Further research directions on each problem are also pinpointed.

Chapter 2 contributes to the literature by providing a formal definition for the delivery mode planning decision and a framework of the distinct delivery mode types that can be used to distribute products from retailers DCs to the stores. The main interdependencies between the delivery mode planning and other SC planning decisions, such as the network design, the delivery pattern and the selection of means of transportation are described. A comprehensive review of the works related with the delivery mode planning problem is provided, highlighting the different terms used in literature. The PhD candidate was the main author of the chapter, being responsible for the reviewing process and the development of the research paper:

- Sara Martins, Pedro Amorim and Bernardo Almada-Lobo. Delivery mode planning for distribution to brick-and-mortar retail stores: discussion and literature review. *Flexible Services and Manufacturing Journal*, 2017. DOI 10.1007/s10696-017-9290-x

This review shows that most of the works that study the delivery mode planning derive from a manufacturer context and are mainly focused on the decision between direct-shipment or cross-docking deliveries, considering simple cost structures for point-to-point deliveries. Few works explore the synergies between routing decisions and the delivery mode planning, which define different consolidations strategies. Therefore, further research can be performed by including these two planning decisions and incorporating the different types of delivery mode (direct-shipment, cross-docking and milk-run). Furthermore, the characteristics of the facilities, either origins, destinations or intermediate facilities of deliveries, are often overlooked. Although this is a tactical problem, the delivery mode planning should still analyze the constraints and costs involved with the facilities, at least at an aggregated level. Moreover, it is important to define consistent distributions plans in order to smooth the operations management and allow for more efficient procedures. In addition, the delivery mode planning has several interdependencies with other planning decisions. The facilities location and the means of transportation used can leverage different types of delivery modes, enabling distinct consolidation strategies, which are restricted by the delivery patterns. Studies on the impact of these interdependencies and on how to incorporate them in the overall distribution planning are further research opportunities.

The main contributions of Chapter 3 are two-fold. Firstly, the proposal of a mathematical model for the network redesign problem of pharmaceutical wholesalers. The model considers intermediate cross-docking facilities, besides the traditional storage DCs, incorporating the trade-off between facilities and transportation costs. This problem tackles the interdependency between the network design and the delivery planning, deciding between direct-shipments and cross-docking. Secondly, the development of a hybrid solution approach that combines optimization and simulation in order to evaluate the operational impact of the solutions. This solution approach allows to incorporate and analyze the constraints and costs involved in the facilities, heading one of the gaps raised in Chapter 2. This work originated the following paper, in which the PhD candidate was the main author, developing the mathematical formulation and the simulation model, the computational experiments and the analysis of the results.

- Sara Martins, Pedro Amorim, Gonçalo Figueira and Bernardo Almada-Lobo. An optimization-simulation approach to the network redesign problem of pharmaceutical wholesalers. *Computers & Industrial Engineering*, 106, pages 315-328, 2017. DOI 10.1016/j.cie.2017.01.026

This research paper has filled one of the gaps identified in the review of Chapter 2. Although it is focused on the wholesalers activity, the solution approach could still be used to analyze operational constraints and costs of a retailer that performs its own distribution to stores. Additionally, the solution approach may be extended to include the routing problem in order to use more realistic distribution costs, instead of an approximation, and evaluate in more detail the customer's reallocation decisions. Furthermore, a deeper interaction between the simulation and the optimization model, which would refine the optimization parameters according to the simulator solution and guide the search, may improve the final results. Finally, an extension of the problem to incorporate product allocation as a decision variable, differentiating product characteristics, would make it suitable for different contexts, where not all DCs are full-liners and some products can be centralized.Chapters 4 and 5 contribute to the growing research efforts on MCVRPs, specifically on the grocery distribution application. Chapter 4 recognizes for the first time the loading constraints inherent to the use of MCVs in grocery retailing and proposes a specialized packing problem that defines the arrangement of vehicles compartments and correspondent orders, avoiding loading issues. Besides the development of the packing problem, the main contribution of this work is the analysis of the impact of loading constraints in the routing and how they should be considered in the planning. The numerical experiments made it possible to derive general rules for the influence of loading constraints. This work resulted in the paper presented below. The first two (main) authors contributed equally to the work. The PhD candidate tackled the description of the overall problem, the development of the specialized packing problem and the B&C algorithm, the incorporation of the packing problem as a repair mechanism within the LNS algorithm and the execution of the numerical experiments.

- Manuel Ostermeier, Sara Martins, Pedro Amorim and Alexander Hübner. Loading Constraints for a Multi-Compartment Vehicle Routing Problem. *Submitted to OR Spectrum*, 2017.

Chapter 5 is the first research work to consider the time-window assignment problem in an MCVRP environment and propose a product-oriented assignment, instead of a customer-oriented one. The definition and formulation of this new problem, which despite its practical relevance has never been studied before, is one of the main contributions of this chapter, along with the ALNS algorithm to generate good quality solutions for the problem. Although the proposed solution approach considers different operators often used in the ALNS literature for different vehicle routing problems variants, new operators capable of coping with time-window assignment and soft constraints needed to be deployed. We called daily operators to the first group of operators from the literature, which focus on the routing aspects of the problem, and weekly operators to the new operators developed to align the time-windows consistently throughout the planning horizon. This work highlights the operational impact that the delivery plans have on the stores operation and reiterates the importance of a consistent planning. The PhD candidate was the main author of the correspondent research article, mainly contributing to the problem definition, the development of the ALNS algorithm, the computational tests and respective results analysis.

- Sara Martins, Manuel Ostermeier, Pedro Amorim and Alexander Hübner, Bernardo Almada-Lobo. Product-Oriented Time-Window Assignment for MCVRP. *To be submitted to European Journal of Operational Research*, 2017.

These last two works also helped filling some of the gaps identified in Chapter 2. In the grocery distribution the retailers DCs are segregated in different temperature zones, which are considered in the distribution planning as separate facilities that need to be visited although they are often within the same location. This is one of the cases where more consolidation effects can be achieved by performing milk-run deliveries, i.e. performing multiple pick-ups before the deliveries. The application of MCVRP to the grocery distribution allows to decide in which cases is advantageous to perform direct-shipments with products from only one temperature range or in which cases is better to perform milk-run deliveries with products from multiple temperature ranges. Furthermore, different types of MCVs can be used for transportation. These works considered MCVs with flexible compartmentalization, but others can be used with more strict sets of compartments. This last type of vehicles, as the single-compartment ones, are less flexible but also less expensive than the flexible MCVs. Therefore, further extensions to the MCVRP would integrate the selection of the optimal fleet mix. Moreover, the packing problem proposed in Chapter 4 could be adopted for less flexible MCVs, but a higher number of infeasible solutions could be reached, impacting more the routing cost. Additionally, the integration of routes from farther cross-docking facilities would help closing other gaps described in Chapter 2, as few works tackle the three types of delivery modes jointly, namely the direct-shipment, cross-docking and milk-run. Furthermore, there is a lack of literature concerning MCVRPs across multiple periods. Chapter 5 goes in this direction, tackling a MCVRP with a multi-period setting and including consistent delivery plans in terms of time-window assignment. Regarding consistent deliveries, further studies could be conducted to better understand whether a consistent delivery mode type influences the DCs and store operations. For the stores a consistent time delivery might suffice, regardless of the delivery mode performed. However, for the DCs a daily change of the delivery modes performed (e.g. one day it has to consolidate products from other DC and on the other day it does not) might trigger inefficiencies in the operation. Other researches on MCVRP with multi-period settings could also be fostered, such as the definition of the delivery patterns. With MCVs, more frequent deliveries of smaller size from different temperature ranges can be performed, thus enabling a reduction of inventory at the stores. Besides extending the MCVRP, another field for future research is the development and comparison of alternative solution approaches for MCVRPs, such as branch-and-price or genetic algorithms.

Overall, this thesis contributes to the current literature on retail distribution planning. It covers some of the opportunities found in the literature, bridging gaps between research and practice. The four research papers developed during the PhD are presented in the following chapters.

Bibliography

Bäckström, K. and Johansson, U. (2006). Creating and consuming experiences in retail store environments: Comparing retailer and consumer perspectives. *Journal of Retailing and Consumer Services*, 13(6):417–430.

Fernie, J., Fernie, J., Sparks, L., and McKinnon, A. C. (2010). Retail logistics in the UK: past, present and future. *International Journal of Retail & Distribution Management*, 38(11/12):894–914.

Grewal, D., Levy, M., and Kumar, V. (2009). Customer experience management in retailing: an organizing framework. *Journal of Retailing*, 85(1):1–14.

Hübner, A. H., Kuhn, H., and Sternbeck, M. G. (2013). Demand and supply chain planning in grocery retail: an operations planning framework. *International Journal of Retail & Distribution Management*, 41(7):512–530.

Mou, S., Robb, D. J., and DeHoratius, N. (2018). Retail store operations: Literature review and research directions. *European Journal of Operational Research*, 265(2):399–422.

Pollaris, H., Braekers, K., Caris, A., Janssens, G. K., and Limbourg, S. (2015). Vehicle routing problems with loading constraints: state-of-the-art and future directions. *OR Spectrum*, 37(2):297–330.

Rushton, A., Croucher, P., and Baker, P. (2014). *The handbook of Logistics & Distribution Management*. Kogan Page Publishers, London, UK, 5th edition.

Sullivan, M. and Adcock, D. (2002). *Retail marketing*. Thomson Learning, London.

Toth, P. and Vigo, D. (2014). *Vehicle Routing: Problems, Methods, and Applications, Second Edition*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics.

Van Belle, J., Valckenaers, P., and Cattrysse, D. (2012). Cross-docking: State of the art. *Omega*, 40(6):827–846.

Part II

Scientific Papers

Chapter 2

Distribution to Retail Brick-and-Mortar Stores

Delivery mode planning for distribution to brick-and-mortar retail stores: discussion and literature review

Sara Martins[*] · Pedro Amorim[*] · Bernardo Almada-Lobo[*]

Published in *Flexible Services and Manufacturing Journal*, 2017
https://doi.org/10.1007/s10696-017-9290-x

Abstract In the retail industry, there are multiple products flowing from different distribution centers to brick-and-mortar stores with distinct characteristics. This industry has been suffering radical changes along the years and new market dynamics are making distribution more and more challenging. Consequently, there is a pressure to reduce shipment sizes and increase the delivery frequency. In such a context, defining the most efficient way to supply each store is a critical task. However, the supply chain planning decision that tackles this type of problem, delivery mode planning, is not well defined in the literature. This paper proposes a definition for delivery mode planning and analyzes multiple ways retailers can efficiently supply their brick-and-mortar stores from their distribution centers. The literature addressing this planning problem is reviewed and the main interdependencies with other supply chain planning decisions are discussed.

Keywords Retail · Delivery Mode Planning · Distribution · Consolidation

2.1. Introduction

This research focuses on retailers' supply chains (SCs). These SCs have processes that are more controlled by the manufacturers, related to the production phase, and other processes more controlled by the retailers concerning the positioning of the products close to the final customers (de Jong and Ben-Akiva, 2007).

In the past, retailers were merely simple stores that passively received products from manufacturers according to their demand forecast. Nowadays, larger retailers have their own distribution centers (DCs) (Kuhn and Sternbeck, 2013) and control, organize and

[*] INESC TEC and Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

manage the flow of products through the SC (Fernie et al., 2010). In addition, retailers are adopting multi-channel strategies that present diverse alternatives to buy and take their products to final customers. These strategies make the product mix to be distributed through the traditional channel (brick-and-mortar stores) more and more heterogeneous and dynamic. Despite the increasing online shopping trend, brick-and-mortar (B&M) stores still account for the majority of the retailers sales. Moreover, to be closer to their customers, every year retailers open new stores with distinct characteristics (such as store size, product assortment and net sales) in new locations that need to be incorporated into their overall distribution planning. For these reasons, since the distribution in the different channels is very distinct, this paper focuses on the retail distribution to B&M stores.

In many companies, the transportation of products accounts for one to two thirds of the global logistics cost (Hosseini et al., 2014b). Although most of the retailers contract common carriers to perform their transportation activity, they still want to manage how distribution is planned and executed. The supply of B&M retail stores greatly impacts the final customer service level. For instance, if a vehicle delays its arrival at the store, there may be a stock-out due to insufficient time to refill the shelves (Corsten and Gruen, 2003). Moreover, in grocery retail, it is also very important to manage the distribution to ensure the quality of the product, which may be perishable (Akkerman et al., 2010; Rong et al., 2011; Amorim and Almada-Lobo, 2014). Therefore, retailers might have their own fleet or, most commonly, contract transportation services from a common carrier that will transport the products, similarly to a full-truckload (FTL) contract. Less-than-truckload (LTL) contracts are more commonly used by retailers for online distribution of small deliveries.

In such context, how should retailers use their resources to plan the flow of products from different DCs to B&M stores with distinct needs? This is the research question that this paper wants to analyze. Delivery mode planning (DMP) is the planning decision used in this study to address this problem. The DMP defines how the products should flow throughout the SC, specifying which delivery modes should be used (for instance, direct-shipment, cross-docking or milk-runs). In other words, it prescribes through which facilities products should follow to reach a certain store.

The literature is inconsistent in defining the DMP and the different types of delivery modes. Besides the distinct terminologies associated with this decision, the different terms used can also be related to other types of problems. Therefore, the aim of this paper is threefold: (i) discussing and defining DMP, (ii) analyzing the multiple ways retailers can efficiently plan the supply of their B&M stores from their DCs, and (iii) identifying the main SC planning decisions that impact or are impacted by the DMP.

The remainder of this paper is organized as follows: Section 2.2 exposes the inconsistency found in the literature regarding the DMP and proposes a formal definition for this planning decision. The delivery modes available are also described. Section 2.3 addresses the main interdependencies between the DMP and other supply chain planning (SCP) decisions, while section 2.4 organizes the papers that address the DMP problem. Finally, the last section concludes the discussion and identifies some research opportunities.

2.2. Defining Delivery Mode Planning

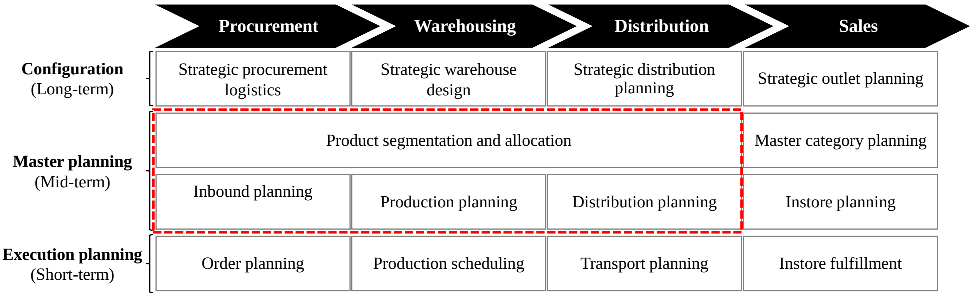
The expression “delivery mode" is not consistent in the literature to describe the distinct ways products flow from their origin to their destination. Analyzing only a direct delivery from a origin to a destination, which is the simplest type of delivery mode to be performed and the most frequently mentioned in the literature, delivery mode can be referred to in different ways:

- delivery mode (Liu et al., 2003; Kuhn and Sternbeck, 2013; Holzapfel et al., 2016);
- transport mode (Eskigun et al., 2005);
- shipment mode (Wang et al., 2016; Bortolini et al., 2016);
- transportation network (Chopra and Meindl, 2001; Du et al., 2007; Hosseini et al., 2014b);
- distribution path (Fleischmann, 2008).

However, different authors also refer and relate some of these terms to other types of decisions. For example, the expressions “transportation and shipment mode" are very often used in transportation problems to describe the decision of using a private fleet, FTL, LTL or a parcel carrier (Chu, 2005; Caputo et al., 2006; Nguyen et al., 2014) or if the transport should be done by road, rail, air or water (de Jong and Ben-Akiva, 2007; Waller et al., 2008; Ahumada and Villalobos, 2011; Manzini et al., 2014). Nevertheless, the designation “delivery mode” seems to be the most consistent in the literature and only a few times associated with other types of decisions.

2.2.1 Decision levels of Delivery Mode Planning

The DMP is mainly related to Master Planning (cf. Figure 1 in which the DMP scope is highlighted by a red dashed box). Nevertheless, depending on the level of analysis, the delivery modes available are different and in different scale.



	Procurement	Warehousing	Distribution	Sales
Configuration (Long-term)	Strategic procurement logistics	Strategic warehouse design	Strategic distribution planning	Strategic outlet planning
Master planning (Mid-term)	Product segmentation and allocation			Master category planning
	Inbound planning	Production planning	Distribution planning	Instore planning
Execution planning (Short-term)	Order planning	Production scheduling	Transport planning	Instore fulfillment

Figure 2.1: Retail Supply Chain Planning matrix adapted from Hübner et al. (2013)

If the DMP analysis goes end-to-end from suppliers to customers, considering the links between distinct echelons of the SC, then the DMP should be positioned at the top of master planning as proposed by Hübner et al. (2013). This planning problem decides whether the products should flow directly from suppliers to customers or delivered to a facility of the company (for instance, storage or cross-docking DC) for further delivery to the customers, as illustrated in the Figure 2.2 (higher tactical level). Therefore, in order to solve the resulting planning problem, it is necessary to analyze the trade-offs between procurement, storage, inventory handling and transportation costs.

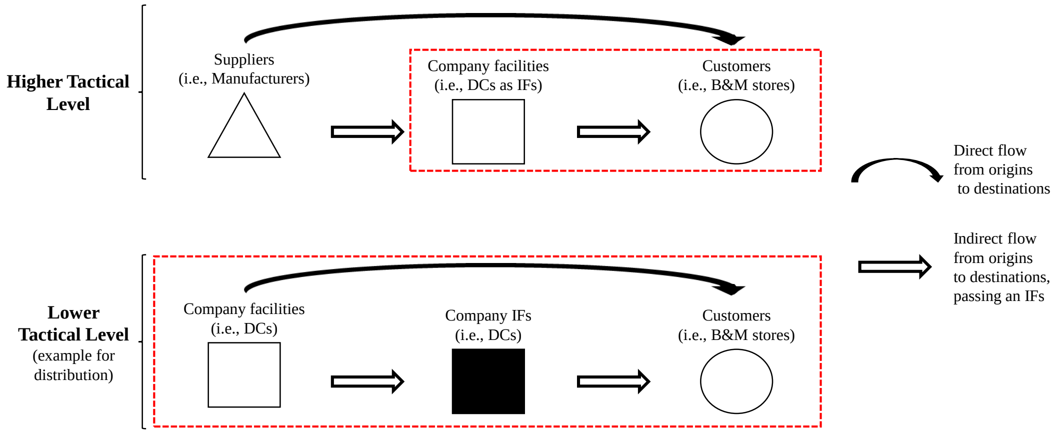

Higher Tactical Level
Suppliers (i.e., Manufacturers)
Company facilities (i.e., DCs as IFs)
Customers (i.e., B&M stores)
Direct flow from origins to destinations
Indirect flow from origins to destinations, passing an IFs
Lower Tactical Level (example for distribution)
Company facilities (i.e., DCs)
Company IFs (i.e., DCs)
Customers (i.e., B&M stores)

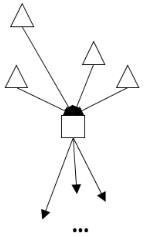
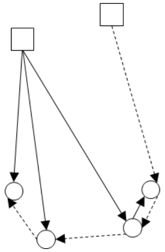
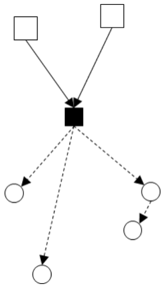
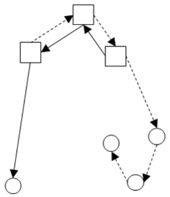
Figure 2.2: Flow analyzed in the DMP at different levels of master planning

When the focus of the analysis is the flow of products between suppliers and the company facilities (procurement) or between those facilities and the B&M stores (distribution), the DMP is positioned at the lower level of the master planning. In this case, the analysis to be conducted requires more detail and is more associated with operational activities. At this level, focusing in the distribution process, the DMP is tackled in the distribution planning module (Hübner et al., 2013). As illustrated in Figure 2.2 - lower tactical level, products can flow directly from an origin (here already a company facility) towards the customers or pass by an intermediate facility (IF). In this context, the origins are the retailers' DCs, the customers are the B&M stores and the intermediate facilities are additional DCs that can be used for (de)consolidation. Both origins and IFs are controlled by the retailer. For this DMP decision level, the products that should be delivered to the DCs, for storage or cross-docking, and afterwards sent to the customers, are already defined.

2.2.2 Types of Delivery Modes

The demand of a B&M store can be satisfied through different delivery modes. When a network uses more than one type of delivery mode it is called a hybrid or tailored network (Chopra and Meindl, 2001; Guastaroba et al., 2016). In the literature, delivery modes can be categorized as: direct-shipment, warehousing, cross-docking and milk-runs. Once again, the definition of the delivery modes is not coherent in the literature. The distinct characterizations used are presented next, and the advantages and disadvantages are outlined. This is summarized in Table 2.1.

Table 2.1: Characteristics of the delivery modes

Delivery mode type	Scheme	Description	Advantages	Disadvantages
Warehousing (Flow from supplier sites)		The suppliers send the products to a storage DC, from where they are afterwards distributed.	Shorter response lead time that allows shorter inventory levels at the stores (Tompkins et al., 1996; Waller et al., 2006) Works like a buffer between supply and demand (Tompkins et al., 1996) Allows the consolidation of products (Tompkins et al., 1996)	Total delivery lead time increases (Van Belle et al., 2012) Higher inventory and handling costs (Van Belle et al., 2012)
Direct-shipment (Flow from retailer DC sites)		Stores are supplied by each zone/location independently	Beneficial for high value or high demand products (Mokhtarinejad et al., 2015) Economic when an FTL is achieved (Hosseini et al., 2014b) Shorter transit time (Min, 1998) Reduces the risk of delays (Min, 1998; Mokhtarinejad et al., 2015)	Stores receive different product categories separated (Sternbeck and Kuhn, 2014) FTL might not be achieved for a single store (Buijs et al., 2014) Longer routes might be performed to achieve FTL (Fleischmann, 2008)
Cross-docking (Flow from retailer DC sites)		Products are received in an intermediate facility and sent as early as possible to the destination, with no or low storage time.	Improves truck utilization (Liu et al., 2003; Van der Vlist and Broekmeulen, 2006; Van Belle et al., 2012) Allows for the consolidation of shipments (Ma et al., 2011; Van Belle et al., 2012; Sternbeck and Kuhn, 2014) Reduces transportation costs (Van Belle et al., 2012; Hosseini et al., 2014b) Reduces the number of truck arrivals to the stores (Sternbeck and Kuhn, 2014)	Involves additional material handling (Van Belle et al., 2012; Hosseini et al., 2014b; Ladier and Alpan, 2016) Requires coordination between the inbound and outbound shipments (Van Belle et al., 2012; Ladier and Alpan, 2016; Guastaroba et al., 2016) Requires investments in information systems (Guastaroba et al., 2016) Higher transit time and distance for each shipment (Lapierre et al., 2004)
Milk-run (Flow from retailer DC sites)		A vehicle picks up products from different facilities and then delivers them to stores.	No need to send multiple independent LTL to stores (Hosseini et al., 2014b) Allows for more frequently deliveries (Hosseini et al., 2014b) Reduces the number of truck arrivals to stores (Sternbeck and Kuhn, 2014)	Is only beneficial for closer facilities (Hosseini et al., 2014b)

Legend: △ Supplier □ Origin (DC) ○ Destination (B&M Store) ■ Intermediate Facility (DC)

2.2.2.1 Warehousing

Warehousing delivery indicates that instead of the products being sent from the suppliers to the stores, they are delivered from the retailer DCs from their stock. The DCs are responsible for receiving products from suppliers, storing them and afterwards picking them up according to the customers' orders, building up a shipment to be delivered (Van Belle et al., 2012; Buijs et al., 2014).

By resorting to warehousing, the retailer has the products stored closer to the stores and therefore, the response lead time to an order is shortened (Tompkins et al., 1996). This allows the stores to reduce their inventory levels (Waller et al., 2006), but it increases the inventory and handling costs at the DCs (Van Belle et al., 2012). Additionally, since products from different suppliers are stored in the same place, distinct order consolidations can be performed (Tompkins et al., 1996). With this delivery mode, the in-transit time of a product (from suppliers to stores) is increased (Van Belle et al., 2012) and therefore, it is only beneficial for products with a long shelf life.

2.2.2.2 Direct-Shipment

Direct-shipment deliveries are the simplest delivery mode. Each origin independently distributes the products to their destinations. According to Liu et al. (2003), the vehicles can visit one or more destinations in the same route. However, some authors consider that the vehicles can only visit one destination directly and then return (Gaur and Fisher, 2004; Eskigun et al., 2005; Galbreth et al., 2008; Hosseini et al., 2014b). This paper considers the definition used by Liu et al. (2003), being the main point of this delivery mode that no intermediate facilities are visited.

This is the delivery mode with less risk of delays because it is not necessary to wait for the arrival of other orders to be aggregated. Therefore, it is more beneficial for high value products (Mokhtarinejad et al., 2015). The disadvantage is that products allocated to distinct facilities are delivered at the destination at different points in time, which increases the number of vehicle arrivals (Sternbeck and Kuhn, 2014). Additionally, in case the shipment size to each destination is small, it might be difficult to achieve full-truckloads (FTL). This leads to the consolidation of different destinations in the same vehicle and results in longer routes (Lapierre et al., 2004).

2.2.2.3 Cross-docking

In a delivery via cross-docking, shipments from different origins are consolidated with others for the same destination, similar to warehousing, but with no or low storage time (Van Belle et al., 2012; Ma et al., 2011). Other terminologies are found in the literature, such as hub-and-spoke (Liu et al., 2003) and transshipment (Langevin and Riopel, 2005). According to Van Belle et al. (2012), there are different types of cross-docking based on how the process is performed, namely: (i) the number of touches made to the shipment, (ii) the number of stages and (iii) the interchangeability of orders.

A cross-docking delivery can also be used for deconsolidating shipments. In this case, one origin sends large shipments to an intermediate facility where they are broken into smaller ones and sent separately to their destinations. This procedure is performed when the origins are far from the destinations and the order size of the shipments for each destination is small. In this case, the intermediate facility is located in the last miles of the distribution.

This delivery mode makes it possible to improve vehicle utilization and to reduce transportation costs (Van der Vlist and Broekmeulen, 2006; Van Belle et al., 2012; Hosseini et al., 2014b). To accomplish that and to enable a successful implementation, the inbound and outbound vehicles need to be coordinated. This requires significant investments in an information system (Guastaroba et al., 2016). Compared to direct-shipment delivery, it requires additional material handling because the shipment needs to be moved from the inbound to the outbound vehicles, with a possible short storage in between. The product in-transit time increases in comparison with direct-shipments (Hosseini et al., 2014b). However, because economies of scale are allowed both in the inbound and outbound vehicles, it is possible to make deliveries more frequently (Liu et al., 2003).

2.2.2.4 Milk-Run

The milk-run delivery mode got its name from the traditional system of milk distribution and sales in Western culture, in which the milkman delivered bottles of milk to each customer and collected the empty bottles. This method became very popular in the manufacturing industry and the manufacturers started creating a collection of routes to the suppliers (Hosseini et al., 2014b; Lin et al., 2015). A vehicle is sent from the manufacturer's facility to visit different suppliers, collect the products and return to the origin location (Brar and Saini, 2011). Therefore, the milk-run term often refers to pickup routes (Brar and Saini, 2011; Hosseini et al., 2014b; Lin et al., 2015).

de Jong and Ben-Akiva (2007) and Günther and Seiler (2009) distinguish inbound and outbound milk-runs. The inbound milk-run consolidates shipments from different origins, which are collected in sequence, to be delivered to the same destination. On the other hand, the outbound milk-run consolidates shipments at one origin to be delivered to different destinations. Basically, they define the milk-run as a route that can be used for pickups (inbound) or deliveries (outbound). However, in some studies the milk-run term is simply referred to as a route (Liu et al., 2003; Galbreth et al., 2008; Agustina et al., 2014) and in others as routes that visit multiple origins and destinations (Berman and Wang, 2006; Buijs et al., 2014). In this paper, milk-run is defined as a special variant of a pickup and delivery routing problem where all collections have to be performed before starting the deliveries, making it possible to smooth the operations management. Pickup and delivery routes may also be used by some retailers, but normally as a way to collect products to be returned to the distribution center and not to be delivered on that route. This type of problem is addressed more at the operational level and is not in the scope of this work.

The milk-run is normally used when the origins are close to each other, otherwise the transportation costs would increase (Hosseini et al., 2014b). Moreover, it can be beneficial for destinations far away from the origins (Günther and Seiler, 2009). This indicates that the design of the SC network has a great impact on the benefit of the delivery mode types (later discussed in Section 2.3.1). With a milk-run delivery, the origins do not need to wait until an FTL is achieved to perform the delivery or send an LTL separately when the shipments are small. Therefore, it makes it possible more frequent deliveries (Hosseini et al., 2014b). As in the cross-docking delivery, it allows to consolidate shipments from different deliveries. However, while in the cross-docking process the shipments from distinct origins are sent to an IF and can be rearranged and loaded into the vehicle according to the sequence of deliveries. This is not possible in the milk-run process. The execution of the milk-run does not make it possible to rearrange shipments. For this reason, the sequence of the pickups and how the shipments are loaded into the vehicle need to take into consideration how the deliveries will be performed (later discussed in Section 2.3.2).

For both the cross-docking and milk-run delivery modes, if the products have distinct transportation requirements, the vehicles used must ensure them. For instance, in the grocery retailing, it is necessary to take into account different temperatures. This suggests that the type of vehicles used might restrict the type of consolidation performed and, consequently, the delivery mode. This interdependency will be discussed in Section 2.3.

Note that the delivery modes presented are not mutually exclusive. A retailer can, for example, resort to warehousing in the connection with the suppliers and perform direct-shipments from their DCs to their stores. Warehousing is the only delivery mode that is only analyzed at the higher tactical level of DMP, since it is related to the flow between suppliers and stores. The remaining delivery modes can be analyzed from both levels, depending on the flow/connection under analysis (see Figure 2.2).

2.2.3 Impact of consolidation rules on Delivery Mode Planning definition

Based on the previous discussion about the types of delivery modes and the decision levels of DMP, the following definition is proposed for this planning problem:

"Delivery mode planning (DMP) is a decision that outlines the rules for (de)consolidation and transportation between origins and destinations, taking into consideration the bundling effects of the distinct destinations and products supplied, the facilities locations established and the resources available."

Distinct rules for (de)consolidation can be applied in DMP and they will dictate which delivery mode types are more advantageous to use. The replenishment and shipment consolidation are some of the most important issues in today's SCP (Gürbüz et al., 2007; Waller et al., 2008).

Shipment consolidation is a procedure that combines different orders in a single shipment. It promotes FTLs, but the vehicles can only depart when all orders associated are prepared and loaded. Gümüş and Bookbinder (2004) describe different types of consolidation. A time-based approach is when all the orders processed in a given period are aggregated in a shipment. This is the case of pharmaceutical distribution, for instance, where the pharmacies place many orders to the distributors throughout the day. Afterwards, these orders are aggregated and released for picking before the vehicles leave (Martins et al., 2017). Another approach is defining the shipment quantity that triggers the dispatch of the order preparation. Higginson and Bookbinder (1994) and Higginson and Bookbinder (1995) analyze these two types of consolidation in more detail.

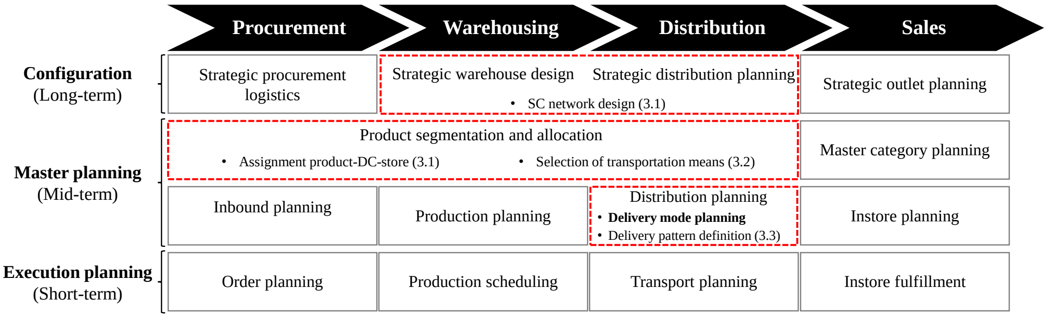
Each of the previous consolidation types can additionally be associated with a spatial consolidation and/or a product consolidation (Gümüş and Bookbinder, 2004). The spatial consolidation aggregates orders from customers located close to each other and, subsequently, requires the definition of a cost-minimizing routing (Fleischmann, 2008). When customer order sizes do not make it possible to achieve FTL, economies of scale can be reached by combining orders from different customers in the same vehicle. Spatial consolidation can be executed in all the delivery mode types, especially for the direct-shipment case where achieving FTL for a single destination is more difficult. Additionally, different products can also be aggregated in the same shipment, which is designed as product consolidation. If the products are assigned to the same facility, it is already common to aggregate them in a single order for a customer. However, the challenge is deciding if they should be aggregated when they are not in the same location. For this reason, this consolidation type needs to be analyzed more carefully when cross-docking and milk-run delivery modes are in question. In these cases, consolidation costs regarding additional material handling and transportation need to be considered.

After a consolidation, a deconsolidation process might occur. The deconsolidation breaks down large shipments into smaller ones to be sent to separate destinations. While consolidation can be performed both in the origins and in the IFs to combine shipments, the deconsolidation process is only performed in the IFs. Therefore, it is associated with a cross-docking delivery mode and the deconsolidation process is normally performed in the last miles of the distribution. Note that a deconsolidation is not mandatory every time a consolidation process occurs.

The consolidation possibilities, and consequently the delivery mode possibilities, are limited by the delivery patterns defined. The delivery pattern decision is defined in the distribution planning (see Figure 2.1) and determines the frequency of supplies and the days of replenishment to the B&M stores. Additionally, the location of the retailer's facilities in the network can leverage distinct delivery modes. As previously discussed (Section 2.2.2), some delivery modes are more suitable for closer DCs and others for distant stores. Moreover, depending on the products' transportation requirements, distinct resources might be needed to process them. These conditions lead to the main interdependencies the DMP has with other SCP decisions, which will be discussed in the next Section.

2.3. Delivery Mode Planning main interdependencies

The definition proposed for DMP states that this planning decision has to be defined taking into consideration the bundling effects of the distinct destinations and products supplied, the facilities locations and the resources available (see Section 2.2.3). The bundling effects of the distinct destinations and products supplied are affected by the allocation of the products and stores to the DCs and the delivery patterns defined. Additionally, the locations of the DCs can leverage different delivery mode types, which can also be restricted by the type of vehicles available to perform the transportation (see Section 2.2). These interdependencies link the DMP mainly to three other decisions of SCP that restrict or are restricted directly by it: (i) the SC network design and product-DC-store assignment, (ii) the selection of transportation means and (iii) the definition of delivery patterns.



	Procurement	Warehousing	Distribution	Sales
Configuration (Long-term)	Strategic procurement logistics	Strategic warehouse design • SC network design (3.1)	Strategic distribution planning	Strategic outlet planning
Master planning (Mid-term)	Product segmentation and allocation • Assignment product-DC-store (3.1) • Selection of transportation means (3.2)			Master category planning
	Inbound planning	Production planning	Distribution planning • **Delivery mode planning** • Delivery pattern definition (3.3)	Instore planning
Execution planning (Short-term)	Order planning	Production scheduling	Transport planning	Instore fulfillment

Figure 2.3: Supply Chain Planning matrix adapted from Hübner et al. (2013). DMP main interdependencies are highlighted with a red dashed box.

2.3.1 SC Network Design and product-DC-store assignment

The SC network design can leverage distinct types of delivery modes depending on how it is defined and, therefore, impact the DMP. The location of the retailer facilities is a long-term decision that it is planned at strategic level and is connected to the allocation of stores to the DCs defined in the master planning (see Figure 2.3). If all the products are allocated in each DC, only direct-shipments or delivery via cross-docking can be used, with the latter being used for deconsolidation in the last miles. On the other hand, if the products are allocated to distinct DCs but they are close to each other, a milk-run delivery mode can also be an option, as well as a cross-docking for consolidation.

There are various research areas that address location and distribution decisions together. One example is the location-routing problems that combine the location and routing decisions (Aykin, 1995; Nagy and Salhi, 2007; Drexl and Schneider, 2015). However, this type of problem does not consider IFs. The aim is to locate the origin facilities taking into consideration the routing requirements to cover all destinations. Another example is the location and routing scheduling problems with cross-docking (Mousavi and Tavakkoli-Moghaddam, 2013; Mousavi et al., 2014). In this case, the location decisions focus on cross-docking facilities. However, the flow must pass through all these facilities, resulting in a pure network.

Gümüş and Bookbinder (2004) were the first to incorporate consolidation effects into a location-distribution model with cross-docking facilities. They analyze a hybrid network with distinct structures, considering one to multiple suppliers and products and multiple seed customers. Each seed customer is a group of customers that the authors route outside the model. Their goal is to decide how to supply each product to each seed customer and the number and location of the cross-docking facilities to be used. The products have distinct costs and can be supplied by different delivery modes. The problem is solved with an exact method.

More recently, Mokhtarinejad et al. (2015) tackled the location decisions jointly with the DMP for a hybrid network with a one-to-one structure. A scheduling sub-problem is included and the cross-docking facilities only have an inbound door. This condition makes the vehicles stay in queue until the door is free for unloading. Mokhtarinejad et al. (2015) propose a three-stage method to solve the problem. Their goal is to minimize the transportation costs, which differ from direct-shipment to cross-docking, and traveling time.

2.3.2 Selection of transportation means

The selection of transportation means is a tactical decision, from master planning (see Figure 2.3), and is most commonly referred to as the decision of selecting a transportation mode amongst truck, rail, air and water. It is associated with global supply chains that have to flow products worldwide (Vidal and Goetschalckx, 2001; Ahumada and Villalobos, 2011). In the retail industry the transportation of products between DCs and stores is normally ensured by truck, since the DCs are positioned strategically close to the B&M stores.

Considering the truck transportation as a given, the selection of transportation means chooses the type of truck to be used. Distinct types of vehicles impose different limitations to the way the distribution is performed. Naturally, the type of vehicle can restrict the types of delivery modes that can be chosen. For instance, if multi-compartment vehicles are not used, products that need to be physically separated cannot be consolidated in the same delivery, whether it is by cross-docking or milk-run. Therefore, the decision regarding the means of transportation used impacts the range of possible delivery modes.

The truck attribute that is most commonly used as a distinctive factor in the literature is the loading capacity, particularly in vehicle routing problems with heterogeneous fleets. Another distinction is multi-compartment versus single compartment vehicles. This characteristic often appears in problems regarding waste or glass collection, where the material has to be separated according to its characteristics (Henke et al., 2015; Oliveira et al., 2015). In these problems, the truck can load different materials at each pickup point. Because loading is made from the top of the truck, the sequence by which it is made is not important. However, there are cases in which it is. In grocery retailing, to transport products with distinct temperature requirements, it is necessary to use a multi-compartment vehicle that secures those temperatures in the different compartments. However, in this case the vehicles are often loaded and unloaded by the rear of the truck. Therefore, the sequencing of the loading will restrict the sequencing of the unloading, depending on the compartments' division. There are distinct types of trucks that allow for more flexibility in the arrangement of compartments than others. In trucks with longitudinal divisions (Figure 2.4a), the products cannot be unloaded from inside compartment A unless the others are emptied first. On the other hand, the trucks with transversal divisions (Figure 2.4b) are more flexible, as products from different compartments can be unloaded simultaneously. Note that more flexibility can be added if the truck can be divided into more compartments of distinct sizes and shapes.

Pollaris et al. (2015) review vehicle routing problems with loading constraints. The authors describe the various types of loading constraints that can be found and underline

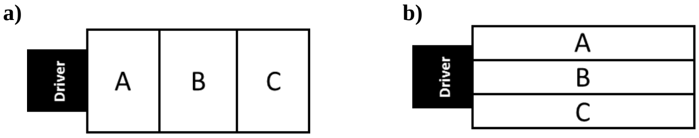


Figure 2.4: Example of multi-compartment vehicle types. **a** Longitudinal divisions and **b** transversal divisions

the importance of considering them in routing problems. Sequence-based loading is one of the loading constraints associated with the location of the items inside the vehicle. The authors state that it is commonly used in vehicle routing problems. However, all the papers that consider it rely on single-compartment vehicles. Moreover, whether it is a single or multi-compartment vehicle, if the products to be loaded are in different locations, it is necessary to analyze both the sequence-based loading and unloading (Ostermeier et al., 2016).

2.3.3 Definition of delivery patterns

The bundling effects that can be achieved in distribution planning, whether related to the consolidation of stores or products, is limited by the delivery patterns defined. The delivery pattern is a decision of distribution planning (see Figure 2.3) and determines the replenishment frequency of the stores, for instance once or twice a week, and the days of replenishment, such as every Monday and Thursday (Kuhn and Sternbeck, 2013). It is defined for each store and can be different for distinct product segments.

The selection of a delivery pattern allows the retailers to organize their operational activities more smoothly, since it defines exactly the days when the deliveries will be made. Therefore, the coordination between flows and operational planning, such as workforce scheduling at the DCs and stores, is simplified (Gaur and Fisher, 2004; Schöneberg et al., 2010).

The more frequent the deliveries, the more fractionated the shipments to distribute will be. As a result, it is more difficult to achieve an FTL for a given store and spatial consolidation counteracts that. However, only stores with deliveries scheduled for the same day can be consolidated. A similar reasoning is applied for product consolidation. Note that the patterns might not be exactly equal but can match in some days. Hence, the delivery patterns defined for each store-product limit the consolidation possibilities in a given day and consequently the DMP (Schöneberg et al., 2010).

Kuhn and Sternbeck (2013) conducts an exploratory study on the delivery pattern strategies used by the retailers. Most retailers interviewed by the authors implement a strategy in which each store and product segment has a specific delivery pattern. In such case, different characteristics have to be taken into consideration regarding the products, such as perishability. In the case of stores, it is necessary to take into consideration other aspects such as sales volume and backroom size (Kuhn and Sternbeck, 2013). Most retailers recognize that this planning decision is based on rules of thumb or only focuses on the transportation aspect.

The delivery pattern decision can be seen as an inventory routing problem (IRP). Research on IRPs has been growing since the 1980s, especially because of the increasing trend of the vendor management inventories (VMI) model. In the VMI, the suppliers have the power to decide when and by how much to supply their customers, ensuring that they have the agreed stock-outs.

Bertazzi and Speranza (2012) present an introduction to the IRP and describe its main characteristics in terms of: (i) shipping time and planning horizon, (ii) structure and objective of the distribution policy and (iii) decision space. They classify the decisions of an IRP in two ways. Decisions over time concern only the problems where the routes are fixed and the aim is to decide when and by how much to supply the customer. In case the routing is also included, the problem incorporates decisions over time and space. In their review, Bertazzi and Speranza (2012) focus on the first case. The previous papers by Bertazzi et al. (2000) and Bertazzi and Speranza (2002) tackle the problem of the IRP with decisions over time considering delivery frequencies only.

Gaur and Fisher (2004) classify the IRP into strategic, tactical and infinite horizon. The purpose of the strategic IRP to define the fleet required to perform the supplies at the minimum cost and the tactical IRP considers the use of a fixed fleet over a finite time horizon. The tactical IRP decides at which period the customers should be supplied considering transportation costs, target service levels, inventory levels, randomness of demand and resource constraints. The infinite time horizon IRP aims at minimizing the transportation, ordering and inventory holding costs, taking into consideration constant demand rates and unrestricted fleet. In their work, Gaur and Fisher (2004) address the vehicle routing and delivery scheduling problem of a retailer operating in the Netherlands. Their goal is to define of a weekly delivery schedule for each store and to determine routes at a minimum cost.

There are few papers on IRP that explicitly consider delivery frequency as a decision variable. Moreover, the ones that do consider it focus on inventory and transportation costs without taking into account other interdependencies in the SC. The study by Kuhn and Sternbeck (2013) identifies the main interdependencies between the delivery patterns, transportation, DC operation and instore logistics planning.

Sternbeck and Kuhn (2014) address the delivery pattern problem for grocery retailers, incorporating the interdependencies previously discussed. Their model considers a pallet as unit of analysis and that each store can only be supplied once a day. It is assumed that distinct product segments delivered on the same day are delivered together. The authors propose a decision model that selects the stores' delivery patterns from a set of pre-defined patterns. It is assumed that the transportation is made by a 3PL and, therefore, the related costs are only dependent on the distance between the DCs and the stores and the volume transported.

Holzapfel et al. (2016) extend the work by Sternbeck and Kuhn (2014) to include spatial consolidation effects in their model. They consider the bundling effects of supplying distinct stores in the same route if they have the same day of delivery. The aim is to minimize the total costs of the entire retail distribution chain. It is considered that the retailer is responsible for the entire distribution process and, therefore, optimizing routes is relevant. The authors propose a sequential solution procedure.

2.4. Delivery Mode Planning Problems

2.4.1 Characterization of the problems reviewed

As mentioned before, this work focuses on the distribution of products from the retailers' DCs, whether they are used as warehousing or cross-docking, to the B&M stores. Therefore, the literature review presented below focuses on the DMP problems that address the distribution process in particular, disregarding those that address the integration with procurement and warehousing processes. Moreover, only problems that make it possible to choose more than one delivery mode (hybrid networks) are analyzed, contrary to pure networks where only one delivery mode can be chosen.

The DMP can be analyzed for different SCs. Buijs et al. (2014) distinguish three SC configurations: (i) few-to-many, (ii) many-to-few and (iii) many-to-many. The first focus on the retail industry where a reduced number of DCs supply a high number of stores. In the second configuration, the emphasis is on the manufacturer context where multiple suppliers satisfy a smaller number of manufacturing facilities. Most research on DMP deals with this context. The last is related to the parcel delivery industry where orders are sent from multiple origins to multiple destinations. This section reviews all the literature concerning the DMP problem regardless of the industry, since inferences can be made to the retail context.

The way the retail B&M stores are supplied has a great impact on their service level. Due to the control retailers want to have over their SC, a private fleet of vehicles or a common carrier under the contract of dedicated FTL is chosen often to execute their transportation. Min (1998) discusses some of the advantages and disadvantages of using a private fleet or a normal contract with a carrier. A private fleet allows more control over transport operations and the vehicles can be used for temporary storage and advertisement. Additionally, using a private fleet makes the operation more responsive and flexible (Min, 1998). Hiring a dedicated service from a carrier to perform the transportation also entails the advantages of a private fleet.

Caputo et al. (2006) describe the difference between an FTL and an LTL service and identify the cost structures normally used for each case. While the cost of an LTL service is only dependent on the quantity sent and the carrier fees, the cost of an FTL depends on the consolidations (spatial) made in each vehicle. The cost functions used in a problem analysis should give an accurate representation of the real life costs (Günther and Seiler, 2009). However, the real carrier rates can be very complex and difficult to analyze (Crainic and Laporte, 1997). The rates represent a contract between two parties taking into consideration a given set of specifications that try to cover possible deviations from the agreement (Günther and Seiler, 2009). For this reason, most papers reviewed simplify the calculation of transportation costs. Most cost structures used are very simple, with fixed costs per connection origin-destination or variable costs related either to the quantity transported or kilometers made. The last case is only present in the papers that include the routing problem (see Section 2.4.2). Most of the models reviewed assume that the transportation is made by a carrier or third party logistics provider (3PL) and, therefore, do not consider the routing problem (see Section 2.4.3).

Guastaroba et al. (2016) review papers concerning transportation planning with intermediate facilities (IFs). Without doing it explicitly, the authors' review addressed the DMP for pure and hybrid networks. Therefore, the literature presented in this section extends their work, putting more emphasis on this planning problem. As in their work, the SC network structure analyzed in the papers reviewed is characterized as one-to-one (1-1), when the demand is defined for each pair origin-destination, one-to-many (1-M), which is similar to the previous but the distribution network considered has many destinations, and many-to-many (M-M), when distinct origins can supply the same destination with the same product. Table 2.2 compiles the abbreviations used in this review to characterize the problems.

Table 2.2: Summary of the abbreviations used in the review

Delivery mode	**Fleet**	**Costs**
DS: Direct-shipment CD: Cross-docking MR: Milk-run	HT: Heterogeneous HM: Homogeneous L: Limited UL: Unlimited	FTC: Fixed transportation cost VTC: Variable transportation cost HC: Holding cost C: Consolidation cost PCTW: Penalization cost for time-windows

The review is divided into the studies that include the routing problem and those that do not (Section 2.4.2 and 2.4.3, respectively). Table 2.3 categorizes the papers reviewed according to the different attributes. Information about the solution methods (exact or heuristic approaches) used in the different problems and if the problem is modeled with time dynamism or not are also stated.

As it can be observed in Table 2.3, the DMP of hybrid networks has received more attention recently and most literature focus on cross-docking networks (Van Belle et al., 2012).

2.4.2 DMP with routing problem

The papers that incorporate the routing problem in the DMP problem normally use a transportation cost based on the distance traveled, and sometimes add a fixed cost for each vehicle used.

Liu et al. (2003) were the first to consider the planning of hybrid networks with the routing problem. They consider a network where a set of suppliers have to visit a set of customers and the decision is whether to make the deliveries directly or indirectly, passing through an IF. The objective is to minimize the transportation costs associated with the distance and the authors propose a heuristic procedure to solve the problem. It starts by solving the problem considering a pure network and afterwards it makes iterative improvements by changing pairs supplier-customer from one delivery mode type to the other.

Van der Vlist and Broekmeulen (2006) do not explicitly include the routing problem, but incorporate in their analysis a parameter that expresses the efficiency of the routing and the cost per stop. They use a formula that identifies the break-even point between direct-shipment and cross-docking deliveries, considering the cost of an FTL between facilities

Table 2.3: Summarization of the attributes of the papers reviewed

Paper	Delivery mode	Routing	Network structure	Authors' problem definition	Fleet	Operational considerations	Costs	Time dynamism	Solution method
Liu et al. (2003)	DS ; CD	Yes	1-1	Mixed truck delivery system	HM ; UL	-	VTC	No	Heuristic
Lapierre et al. (2004)	DS ; CD	No	1-1	Distribution network design	-	-	-	No	Meta-Heuristic
Berman and Wang (2006)	DS ; CD	No	1-1	Distribution strategy	HM ; UL	Consolidation time	FTC ; HC	No	Exact Method
Van der Vlist and Broekmeulen (2006)	DS ; CD	No	1-1	Transport consolidation	HM ; UL	-	FTC	No	Heuristic
Li et al. (2006)	DS ; CD	No	M-M	Shipment consolidation problem in distribution networks	HM ; UL	Capacitated origins ; TW for origins and destinations	VTC ; HC FTC	No	Meta-Heuristic
Galbreth et al. (2008)	DS ; CD	No	1-M	-	HT ; UL	Capacitated IF ; Earlier supplies allowed ; Limited shipments per customer	FTC ; VTC HC	Yes	Exact Method
Günther and Seiler (2009)	DS ; MR	No	1-1	Operational transportation planning problem	HT ; UL	Allows pickups and deliveries ; TW for origins and destinations	VTC	No	Heuristic
Musa et al. (2010)	DS ; CD	No	1-1	Transportation problem of cross-docking networks	HM ; UL	-	FTC	No	Meta-Heuristic
Üster and Agrahari (2010)	DS ; CD	No	1-1	Load-planning problem	HT ; UL+L	CD with two IFs	VTC	No	Heuristic
Ma et al. (2011)	DS ; CD	No	M-M	Shipment consolidation and transportation problem	HM ; UL	Capacitated origins ; TW for origins and destinations	FTC; VTC HC	No	Meta-Heuristic
Dondo et al. (2011)	DS ; CD	Yes	M-M	Vehicle routing problem with cross-docking in supply chain management	HT ; UL	Allows pickups and deliveries ; TW for destinations ; service time	FTC ; VTC PCTW	Yes	Exact Method
Jewpanya and Kachitvichyanukul (2012)	DS ; CD	No	M-M	Cross-docking distribution planning	HM ; UL	Capacitated origins ; Consolidation time ; TW for origins and destinations	FTC ; VTC HC ; C	Yes	Meta-Heuristic
Santos et al. (2013)	CD ; MR	No	1-1	Pickup and delivery problem with cross-docking	HM ; L	-	FTC	No	Exact Method
Hosseini et al. (2014b)	DS ; CD MR	Yes (MR)	1-1	Transportation problem of a consolidation network	HM ; UL	-	FTC	No	Meta-Heuristic
Holla (2015)	DS ; CD	No	1-1	Cross-docking problem	HT ; L	Capacitated IF ; Possibility to lease or rent trucks	VTC ; FTC HC ; C	Yes	Exact Method

and the load to be sent directly and consolidated. The consolidated load is updated in an iterative process.

Dondo et al. (2011) include a pickup and delivery routing problem in their model. The network considered has one origin, a set of IFs and a set of customers. The customers can be supplied from the origin or from an IF. The IFs are used as warehouses with target inventory levels, but they can also be used as cross-docking facilities. A route that starts at the origin can perform deliveries to the customers and pickups or deliveries to the IFs. The routes starting at the IFs can only perform deliveries to the customers. A heterogeneous fleet of vehicles is considered with depots at different locations and each vehicle can perform more than one route. The vehicles' capacity is verified both in terms of weight and volume. The work by Dondo et al. (2011) is an extension and generalization of the model proposed in Dondo et al. (2009), where cross-docking is not allowed and the IFs are considered demand points. Both problems are formulated and solved with an exact method.

2.4.3 DMP without routing problem

The papers that do not include the routing decisions most commonly assume either a cost structure similar to an LTL rate, with quantity discount in some cases, or an FTL rate per vehicle sent between two given locations.

Lapierre et al. (2004) analyze a hybrid network with a one-to-one structure. Their model decides the best delivery mode to supply the customers taking into account the consolidations that can be performed. These consolidations are determined considering the distinct attributes of the load (weight and volume) and the distinct transportation costs of LTL, FTL and parcel. Hence, the decisions regarding the consolidation and carrier selection on each delivery are dynamic and connected. Two meta-heuristics are used to solve the problem.

Üster and Agrahari (2010) also consider an FTL and an LTL cost structure in a network with a one-to-one structure. However, in this case the products can flow directly from an origin to a destination or pass by two IFs. In the latter, the products are sent in an LTL carrier to a consolidation center, where they are combined and sent by an FTL carrier to the deconsolidation center. At this point the shipments are divided again between LTLs and sent to their destinations. The vehicles used for an FTL are limited and assumed homogeneous. To solve the problem, the authors develop three heuristics.

Galbreth et al. (2008) analyze a different network structure, namely a one-to-many. Their goal is to decide whether the supplier should send an FTL to customers or an FTL to a cross-docking facility, with limited capacity, where the load will be divided between several LTLs. The FTL shipments are restricted and have a fixed cost associated per pair origin-destination. In contrast, the LTL shipments are incapacitated and associated with a modified all-unit discount (MAUD) cost function (Chan et al., 2002). Galbreth et al. (2008) assume that the customers allow earlier supplies with a penalty cost and that the number of shipments received per period is limited. They use an exact method to solve a small instance and have insights regarding the cross-docking value in a network. An analysis is also conducted on the effects of the demand and holding costs.

The capacitated cross-docking problem is also tackled by Holla (2015) in a one-to-one structure network. The author considers a cross-docking facility with a small storage area limited per product, which has distinct consolidation costs. The suppliers and cross-docking facilities own a limited fleet of vehicles, having a fixed cost regardless of the route and a variable cost depending on the quantity transported. If required, the vehicles not used can be rented or more vehicles can be leased. Holla (2015) solves the problem with an exact method and performs a sensitivity analysis.

Petersen and Ropke (2011) also incorporate consolidation costs at the cross-docking facilities. The authors analyze a Danish transporter of flowers in which containers of flowers have to be picked up at an origin and then delivered to a destination. Hard time-windows are considered for both origins and destinations. The service time for loading and unloading the containers in the three echelons is determined using a staircase linear function. A meta-heuristic is proposed to solve the problem.

A cost structure different from the previous ones is considered by Li et al. (2006). In their work, the transportation cost includes a setup cost for sending a vehicle between a given pair origin-destination and a cost related to the time between the two points and the size of the shipment. The problem has several homogeneous suppliers with limited capacity. The decision is to determine how each customer should be supplied, considering the possibility of using an IF for (de)consolidation at a given holding cost, which depends on the number of units handled and the time spent. Time-windows are imposed for the suppliers, the IFs and the customers. Li et al. (2006) propose a two-step approach to solve the problem. This is also analyzed by Ma et al. (2011), who considers the same problem structure, but uses distinct heuristics to solve it. Jewpanya and Kachitvichyanukul (2012) further extend this work by including the cost of the consolidation operations in the cross-docking.

Musa et al. (2010) develop a model to define the best consolidation plan and fleet dispatching, adapted from Donaldson et al. (1998). The network contains a one-to-one structure where the objective is to minimize the transportation costs. Musa et al. (2010) propose meta-heuristic to solve the problem.

Hosseini et al. (2014b) extend the work of Musa et al. (2010) by including the possibility of supplying the customers through milk-run deliveries. The network and costs considered are the same, but the paper also comprises the transportation costs between distinct suppliers. In their model, the authors consider a routing sub-problem only for milk-run deliveries. It is assumed that whatever the delivery mode used, the customers are always supplied in a single-drop route. To solve the problem, Hosseini et al. (2014b) propose two meta-heuristics. This work is an extension of their previous research where only direct-shipment and milk-run deliveries were allowed (Hosseini et al., 2014a). To the best of our knowledge, the paper by Hosseini et al. (2014b) is one of the few to consider the three types of delivery modes as decision variables in the distribution process.

In Berman and Wang (2006), the milk-run delivery mode is also identified as one of the possible ways of supplying the customers, but it is not included as a decision. They propose a pre-processing step before running their model to group suppliers in milk-runs, which are afterwards considered one origin. However, this approach does not evaluate the trade-off between using a milk-run or cross-docking delivery mode, which imposes distinct consolidation strategies and resources.

Günther and Seiler (2009) design an operational transportation planning to decide the orders that should be consolidated in one shipment according to the savings that could be accomplished. The authors use four distinct consolidation schemes: bundling, inbound and outbound milk-run and pickup and delivery. The combination of the first three schemes lead to a direct-shipment or milk-run delivery mode with single or multiple drop routes. The schemes are run in sequence and generate the different order combinations. The consolidations respect a set of constraints regarding time-windows and truck characteristics and only the ones that generate savings compared to their individual distribution are considered. In the end, Günther and Seiler (2009) propose an exact method to select the order combinations that make it possible to achieve the minimum global cost. Additionally, they developed a heuristic that can be integrated with commercial Transportation Management Systems.

The pickup and delivery problem with cross-docking addressed in Santos et al. (2013), although not referred as such, also considers the milk-run delivery mode besides cross-docking. It relies on a hybrid network with a one-to-one structure where two types of routes can be performed. Firstly, the vehicles can visit multiple origins to pick up the products, return to the IF for possible shipment rearrangement and afterwards go to the destinations for deliveries. Moreover, after visiting multiple origins the vehicles can start immediately delivering to the destinations, without passing by the IF. This last case can be defined as a milk-run, where one delivery or multiple deliveries can be performed. In both cases, the fleet is considered homogeneous and limited. Santos et al. (2013) assume a set of routes of each type and their respective cost. The authors use an exact method to identify which routes to perform. Although not explicitly, this work also makes it possible to execute the three delivery mode types with given particularities, such as the depot being the cross-docking facility.

2.5. Conclusions and research opportunities

In the retail industry there are multiple products flowing from distinct DCs to B&M stores with heterogeneous needs. In such a context, defining the most efficient way to supply each store is a complex task. Delivery mode planning (DMP) is a tactical decision that defines how the products should flow throughout the SC and which facilities the products should pass by on the way to a given store.

This paper identifies and analyzes the inconsistency found in literature in the definition of DMP and the delivery mode types available and, therefore, our own definition is proposed. The distinct delivery mode types (warehousing, direct-shipment, cross-docking and milk-run) are characterized and their advantages and disadvantages discussed.

The main interdependencies between the DMP and the SC network design and product-DC-store assignment decisions, the selection of the means of transportation and the definition of delivery patterns are also discussed. Moreover, a review is presented focusing on research related to hybrid network distribution, where at least two distinct delivery mode types can be used in the same network.

Based on the discussion of the previous sections four research directions are identified to improve DMP.

Include the different delivery mode types in the same model

Hosseini et al. (2014b) is the only work that analyzes the three delivery modes simultaneously. The milk-run delivery mode is only beneficial in specific cases, when the facilities are close to each other, which may justify its absence in most research on DMP research. However, as the products require distinct temperatures in grocery retailing, it is normal to have DCs with different temperatures in the same location, fostering the milk-run deliveries to leverage the retailer distribution.

Introduce consistency

Most research on DMP focuses on a single period planning horizon. The authors analyze a given SC network and the goal is to optimize the flow of goods between origins and destinations using IFs, if needed. Dondo et al. (2011) and Jewpanya and Kachitvichyanukul (2012) consider a multi period planning horizon in their models because of the time-window constraint, but assume only one day of distribution. Galbreth et al. (2008) and Holla (2015) are the only ones to consider multiple periods but only to account the operational characteristics of their problems. However, it is very important for both the retailer DCs and the B&M stores to have a planned consistent distribution throughout time to facilitate the operations management. Therefore, as the different delivery mode types yield different requirements and impact the DCs and the stores in distinct ways, performing the same type of delivery for a given period allows more efficient procedures.

Consider operational constraints related to the IFs and the B&M stores

Few of the papers surveyed consider the operational costs of IFs and their capacity constraints. This is related to the fact that the majority of the papers do not consider time dynamism. This topic is addressed more frequently in the cross-docking problems with pure networks where synchronization issues need to be analyzed. More information on this type of problems is available in Van Belle et al. (2012) and Ladier and Alpan (2016). Besides the IFs, the B&M stores may also have operational constraints in terms of the delivery time-windows that are considered in some of the papers surveyed, and in receiving capacity. However, none of the papers reviewed incorporated this constraint.

Incorporate the interdependencies with other SCP decisions

The DMP has several interdependencies with other SCP decisions. The products' allocation to the DCs, whether they are centralized or not, as well as the distance between them, leverage different types of delivery modes that can be used by the retailer. Moreover, the allocation of stores to the DCs influences the bundling effects for spatial consolidation.

Only two papers analyze both the DMP with location and allocation decisions, considering a hybrid network (Gümüş and Bookbinder, 2004; Mokhtarinejad et al., 2015).

Furthermore, the bundling effects that can be achieved are limited by the delivery patterns defined. There is one work that tackles the delivery pattern decision in a retail context, considering product and spatial consolidation of stores, but it assumes that different products delivered in the same day flow together with the same cost (Holzapfel et al., 2016). However, if the products require distinct transportation requirements the resources necessary to perform the joint delivery might not be available by the retailer or yield distinct transportation costs. This leads to the decision regarding the selection of transportation means, which, as discussed previously, can restrict the types of deliveries performed. In addition, in case the distinct product segments are located in different DCs, it is necessary to consolidate the segments prior to the delivery to the stores.

Acknowledgements This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project «POCI-01-0145-FEDER-006961», and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The first author was also supported by Grant SFRH/BD/102013/2014 from FCT.

Bibliography

Agustina, D., Lee, C., and Piplani, R. (2014). Vehicle scheduling and routing at a cross docking center for food supply chains. *International Journal of Production Economics*, 152:29–41.

Ahumada, O. and Villalobos, J. R. (2011). A tactical model for planning the production and distribution of fresh produce. *Annals of Operations Research*, 190(1):339–358.

Akkerman, R., Farahani, P., and Grunow, M. (2010). Quality, safety and sustainability in food distribution: a review of quantitative operations management approaches and challenges. *OR Spectrum*, 32(4):863–904.

Amorim, P. and Almada-Lobo, B. (2014). The impact of food perishability issues in the vehicle routing problem. *Computers & Industrial Engineering*, 67:223–233.

Aykin, T. (1995). The hub location and routing problem. *European Journal of Operational Research*, 83(1):200–219.

Berman, O. and Wang, Q. (2006). Inbound logistic planning: minimizing transportation and inventory cost. *Transportation Science*, 40(3):287–299.

Bertazzi, L. and Speranza, M. G. (2002). Continuous and discrete shipping strategies for the single link problem. *Transportation Science*, 36(3):314–325.

Bertazzi, L. and Speranza, M. G. (2012). Inventory routing problems: an introduction. *EURO Journal on Transportation and Logistics*, 1(4):307–326.

Bertazzi, L., Speranza, M. G., and Ukovich, W. (2000). Exact and heuristic solutions for a shipment problem with given frequencies. *Management Science*, 46(7):973–988.

Bortolini, M., Faccio, M., Ferrari, E., Gamberi, M., and Pilati, F. (2016). Fresh food sustainable distribution: cost, delivery time and carbon footprint three-objective optimization. *Journal of Food Engineering*, 174:56–67.

Brar, G. S. and Saini, G. (2011). Milk run logistics: literature review and directions. In *Proceedings of the World Congress on Engineering*, volume 1, pages 6–8.

Buijs, P., Vis, I. F., and Carlo, H. J. (2014). Synchronization in cross-docking networks: A research classification and framework. *European Journal of Operational Research*, 239(3):593–608.

Caputo, A. C., Fratocchi, L., and Pelagagge, P. M. (2006). A genetic approach for freight transportation planning. *Industrial Management & Data Systems*, 106(5):719–738.

Chan, L. M. A., Muriel, A., Shen, Z.-J. M., Simchi-Levi, D., and Teo, C.-P. (2002). Effective zero-inventory-ordering policies for the single-warehouse multiretailer problem with piecewise linear cost structures. *Management Science*, 48(11):1446–1460.

Chopra, S. and Meindl, P. (2001). Supplier chain management–strategies, planning, and operation.

Chu, C.-W. (2005). A heuristic algorithm for the truckload and less-than-truckload problem. *European Journal of Operational Research*, 165(3):657–667.

Corsten, D. and Gruen, T. (2003). Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks. *International Journal of Retail & Distribution Management*, 31(12):605–617.

Crainic, T. G. and Laporte, G. (1997). Planning models for freight transportation. *European Journal of Operational Research*, 97(3):409–438.

de Jong, G. and Ben-Akiva, M. (2007). A micro-simulation model of shipment size and transport chain choice. *Transportation Research Part B: Methodological*, 41(9):950–965.

Donaldson, H., Johnson, E. L., Ratliff, H. D., and Zhang, M. (1998). Schedule-driven cross-docking networks. *Georgia Tech TLI report, The Logistics Institute, Georgia Tech, Atlanta*.

Dondo, R., Méndez, C. A., and Cerdá, J. (2009). Managing distribution in supply chain networks. *Industrial & Engineering Chemistry Research*, 48(22):9961–9978.

Dondo, R., Méndez, C. A., and Cerdá, J. (2011). The multi-echelon vehicle routing problem with cross docking in supply chain management. *Computers & Chemical Engineering*, 35(12):3002–3024.

Drexl, M. and Schneider, M. (2015). A survey of variants and extensions of the location-routing problem. *European Journal of Operational Research*, 241(2):283–308.

Du, T., Wang, F., and Lu, P.-Y. (2007). A real-time vehicle-dispatching system for consolidating milk runs. *Transportation Research Part E: Logistics and Transportation Review*, 43(5):565–577.

Eskigun, E., Uzsoy, R., Preckel, P. V., Beaujon, G., Krishnan, S., and Tew, J. D. (2005). Outbound supply chain network design with mode selection, lead times and capacitated vehicle distribution centers. *European Journal of Operational Research*, 165(1):182–206.

Fernie, J., Fernie, J., Sparks, L., and McKinnon, A. C. (2010). Retail logistics in the uk: past, present and future. *International Journal of Retail & Distribution Management*, 38(11/12):894–914.

Fleischmann, B. (2008). Distribution and transport planning. In *Supply Chain Management and Advanced Planning*, pages 231–246. Springer.

Galbreth, M. R., Hill, J. A., and Handley, S. (2008). An investigation of the value of cross-docking for supply chain management. *Journal of Business Logistics*, 29(1):225–239.

Gaur, V. and Fisher, M. L. (2004). A periodic inventory routing problem at a supermarket chain. *Operations Research*, 52(6):813–822.

Guastaroba, G., Speranza, M. G., and Vigo, D. (2016). Intermediate facilities in freight transportation planning: A survey. *Transportation Science*, 50(3):763–789.

Gümüş, M. and Bookbinder, J. H. (2004). Cross-docking and its implications in location-distribution systems. *Journal of Business Logistics*, 25(2):199–228.

Günther, H.-O. and Seiler, T. (2009). Operative transportation planning in consumer goods supply chains. *Flexible Services and Manufacturing Journal*, 21(1-2):51–74.

Gürbüz, M. Ç., Moinzadeh, K., and Zhou, Y.-P. (2007). Coordinated replenishment strategies in inventory/distribution systems. *Management Science*, 53(2):293–307.

Henke, T., Speranza, M. G., and Wäscher, G. (2015). The multi-compartment vehicle routing problem with flexible compartment sizes. *European Journal of Operational Research*, 246(3):730–743.

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

Higginson, J. K. and Bookbinder, J. H. (1995). Markovian decision processes in shipment consolidation. *Transportation Science*, 29(3):242–255.

Holla, S. (2015). *Dynamic Hybrid Cross-Docking Model with Multiple Truck Types*. PhD thesis, American University of Sharjah.

Holzapfel, A., Hübner, A., Kuhn, H., and Sternbeck, M. G. (2016). Delivery pattern and transportation planning in grocery retailing. *European Journal of Operational Research*, 252(1):354–68.

Hosseini, S. D., Shirazi, M. A., and Ghomi, S. M. T. F. (2014a). Harmony search optimization algorithm for a novel transportation problem in a consolidation network. *Engineering Optimization*, 46(11):1538–1552.

Hosseini, S. D., Shirazi, M. A., and Karimi, B. (2014b). Cross-docking and milk run logistics in a consolidation network: A hybrid of harmony search and simulated annealing approach. *Journal of Manufacturing Systems*, 33(4):567–577.

Hübner, A. H., Kuhn, H., and Sternbeck, M. G. (2013). Demand and supply chain planning in grocery retail: an operations planning framework. *International Journal of Retail & Distribution Management*, 41(7):512–530.

Jewpanya, P. and Kachitvichyanukul, V. (2012). A Particle Swarm Optimization for Cross-docking Distribution Planning Problem. In *Proceedings of the Asia Pacific Industrial Engineering & Management Systems Conference 2012*, pages 464–472.

Kuhn, H. and Sternbeck, M. G. (2013). Integrative retail logistics: an exploratory study. *Operations Management Research*, 6(1-2):2–18.

Ladier, A.-L. and Alpan, G. (2016). Cross-docking operations: Current research versus industry practice. *Omega*, 62:145–162.

Langevin, A. and Riopel, D. (2005). *Logistics systems: design and optimization*. Springer Science & Business Media.

Lapierre, S. D., Ruiz, A. B., and Soriano, P. (2004). Designing distribution networks: Formulations and solution heuristic. *Transportation Science*, 38(2):174–187.

Li, X., Lim, A., Miao, Z., and Rodrigues, B. (2006). Reducing transportation costs in distribution networks. In *Advances in Applied Artificial Intelligence*, pages 1138–1148. Springer.

Lin, Y., Xu, T., and Bian, Z. (2015). A two-phase heuristic algorithm for the common frequency routing problem with vehicle type choice in the milk run. *Mathematical Problems in Engineering*, 2015.

Liu, J., Li, C.-L., and Chan, C.-Y. (2003). Mixed truck delivery systems with both hub-and-spoke and direct shipment. *Transportation Research Part E: Logistics and Transportation Review*, 39(4):325–339.

Ma, H., Miao, Z., Lim, A., and Rodrigues, B. (2011). Crossdocking distribution networks with setup cost and time window constraint. *Omega*, 39(1):64–72.

Manzini, R., Accorsi, R., and Bortolini, M. (2014). Operational planning models for distribution networks. *International Journal of Production Research*, 52(1):89–116.

Martins, S., Amorim, P., Figueira, G., and Almada-Lobo, B. (2017). An optimization-simulation approach to the network redesign problem of pharmaceutical wholesalers. *Computers & Industrial Engineering*, 106:315–328.

Min, H. (1998). A personal-computer assisted decision support system for private versus common carrier selection. *Transportation Research Part E: Logistics and Transportation Review*, 34(3):229–241.

Mokhtarinejad, M., Ahmadi, A., Karimi, B., and Rahmati, S. H. A. (2015). A novel learning based approach for a new integrated location-routing and scheduling problem within cross-docking considering direct shipment. *Applied Soft Computing*, 34:274–285.

Mousavi, S. M. and Tavakkoli-Moghaddam, R. (2013). A hybrid simulated annealing algorithm for location and routing scheduling problems with cross-docking in the supply chain. *Journal of Manufacturing Systems*, 32(2):335–347.

Mousavi, S. M., Vahdani, B., Tavakkoli-Moghaddam, R., and Hashemi, H. (2014). Location of cross-docking centers and vehicle routing scheduling under uncertainty: a fuzzy possibilistic–stochastic programming model. *Applied Mathematical Modelling*, 38(7):2249–2264.

Musa, R., Arnaout, J.-P., and Jung, H. (2010). Ant colony optimization algorithm to solve for the transportation problem of cross-docking network. *Computers & Industrial Engineering*, 59(1):85–92.

Nagy, G. and Salhi, S. (2007). Location-routing: Issues, models and methods. *European Journal of Operational Research*, 177(2):649–672.

Nguyen, C., Dessouky, M., and Toriello, A. (2014). Consolidation strategies for the delivery of perishable products. *Transportation Research Part E: Logistics and Transportation Review*, 69:108–121.

Oliveira, A. D., Ramos, T. R. P., and Martins, A. L. (2015). Planning collection routes with multi-compartment vehicles. In *Operations Research and Big Data*, pages 131–138. Springer.

Ostermeier, M., Martins, S., Amorim, P., and Huebner, A. (2016). Loading constraints for a multi-compartment vehicle routing problem in grocery distribution (working paper).

Petersen, H. L. and Ropke, S. (2011). The pickup and delivery problem with cross-docking opportunity. In *Computational Logistics*, pages 101–113. Springer.

Pollaris, H., Braekers, K., Caris, A., Janssens, G. K., and Limbourg, S. (2015). Vehicle routing problems with loading constraints: state-of-the-art and future directions. *OR Spectrum*, 37(2):297–330.

Rong, A., Akkerman, R., and Grunow, M. (2011). An optimization approach for managing fresh food quality throughout the supply chain. *International Journal of Production Economics*, 131(1):421–429.

Santos, F. A., Mateus, G. R., and Da Cunha, A. S. (2013). The pickup and delivery problem with cross-docking. *Computers & Operations Research*, 40(4):1085–1093.

Schöneberg, T., Koberstein, A., and Suhl, L. (2010). An optimization model for automated selection of economic and ecologic delivery profiles in area forwarding based inbound logistics networks. *Flexible Services and Manufacturing Journal*, 22(3-4):214–235.

Sternbeck, M. G. and Kuhn, H. (2014). An integrative approach to determine store delivery patterns in grocery retailing. *Transportation Research Part E: Logistics and Transportation Review*, 70:205–224.

Tompkins, J., White, J., Bozer, Y., Frazelle, E., Tanchoco, J., and Trevino, J. (1996). Facilities planning. *John Wiley& Sons Inc. 2nd edition. USA*, pages 36–47.

Üster, H. and Agrahari, H. (2010). An integrated load-planning problem with intermediate consolidated truckload assignments. *IIE Transactions*, 42(7):490–513.

Van Belle, J., Valckenaers, P., and Cattrysse, D. (2012). Cross-docking: State of the art. *Omega*, 40(6):827–846.

Van der Vlist, P. and Broekmeulen, R. A. (2006). Retail consolidation in synchronized supply chains. *Zeitschrift für Betriebswirtschaft*, 76(2):165–176.

Vidal, C. J. and Goetschalckx, M. (2001). A global supply chain model with transfer pricing and transportation cost allocation. *European Journal of Operational Research*, 129(1):134–158.

Waller, M., Meixell, M. J., and Norbis, M. (2008). A review of the transportation mode choice and carrier selection literature. *The International Journal of Logistics Management*, 19(2):183–211.

Waller, M. A., Cassady, C. R., and Ozment, J. (2006). Impact of cross-docking on inventory in a decentralized retail supply chain. *Transportation Research Part E: Logistics and Transportation Review*, 42(5):359–382.

Wang, L.-C., Cheng, C.-Y., and Wang, W.-K. (2016). Flexible supply network planning for hybrid shipment: a case study of memory module industry. *International Journal of Production Research*, 54(2):444–458.

CHAPTER 3

Network Redesign for Pharmaceutical Wholesalers

An optimization-simulation approach to the network redesign problem of pharmaceutical wholesalers

Sara Martins[*] · **Pedro Amorim**[*] · **Gonçalo Figueira**[*] · **Bernardo Almada-Lobo**[*]

Published in *Computer & Industrial Engineering*, 2017
http://dx.doi.org/10.1016/j.cie.2017.01.026

Abstract The pharmaceutical industry operates in a very competitive and regulated market. The increased pressure of pharmacies to order fewer products and to receive them more frequently is overcharging the pharmaceutical's distribution network. Furthermore, the tight margins and the continuous growth of generic drugs consumption are pressing wholesalers to optimize their supply chains. In order to survive, wholesalers are rethinking their strategies to increase competitiveness. This paper proposes an optimization-simulation approach to address the wholesalers network redesign problem, trading off the operational costs and customer service level. Firstly, at a strategic-tactical level, the supply chain network redesign decisions are optimized via a mixed integer programming model. Here, the number, location, function and capacity of the warehouses, the allocation of customers to the warehouses and the capacity and function of the distribution channels are defined. Secondly, at an operation level, the solution found is evaluated by means of a discrete event simulation model to assess the impact of the redesign in the wholesaler's daily activities. Computational results on a pharmaceutical wholesaler case-study are discussed and the benefits of this solution approach exposed.

Keywords Network Redesign · Pharmaceutical Industry · Optimization · Simulation

3.1. Introduction

Pharmaceutical wholesalers purchase large quantities of drugs to manufacturers (laboratories), store them in warehouses and later distribute them to retailers (pharmacies) according to their demand. Because of the high number of laboratories, products and retailers,

[*] INESC TEC and Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

wholesalers play an important role in the pharmaceutical industry. Retailers do not have to negotiate with several suppliers to purchase the products they need. Moreover, laboratories can benefit from economies of scale by not managing and distributing multiple low quantity orders to the retailers.

The pharmaceutical industry operates in a very regulated market. Regulatory entities outline distribution procedures and fix prices, imposing the margins of each player of the supply chain (SC). Wholesalers have the lowest margin, about 4%, while the laboratories' margins can reach up to 70%. These figures, together with product exclusivity, underline the bargaining power of laboratories that depreciate their delivery service with long and highly variable delivery lead times. However, this behavior is changing because of the increasing acceptance and growth in the market of generics, predicting a shift in the pharmaceutical SC management. This may result in an improvement of the suppliers' lead time, which will in turn boost their service and make them more competitive, enabling wholesalers to advance in their inventory management.

On the other stream of the SC, the economic crisis is forcing pharmacies to change their purchasing behavior, ordering more frequently and in lower volumes, reaching up to 3-4 daily deliveries, thus burdening the wholesalers' distribution system. Due to the intense competition in today's marketplace, companies need to rethink their strategies and to start operating as SC members instead of competing individually (Baghalian et al., 2013). The pharmaceutical SC is shifting towards this philosophy, where laboratories, wholesalers and retailers will collaborate in a more synchronized way.

With these transformations occurring in the pharmaceutical industry, companies are committed to optimizing their current SC. For this reason, the supply chain network redesign (SCNR) is one of the most important decisions. In this industry, the response time of wholesalers and the availability of products are their competitive edges. Therefore, the process of redesigning a wholesaler network must be carefully studied, involving the logistics and marketing departments. It is critical to analyze how a solution affects the operational procedures and the respective customer service level. A small variation in the service can result in substantial sales losses as the customers have no changeover costs for switching between competitors. This is the reason why supply chain network design and redesign need to be analyzed differently. Typically, the former is defined as trading-off the costs and the expected attraction function of customer service. In the latter, the company already has a market share and the customers are subjected to a certain service level.

There is scarce literature addressing the SCNR from the point of view of the wholesaler. This paper is a contribution in this direction, proposing a novel approach to the SCNR problem of pharmaceutical wholesalers. The aim is to optimize the strategic-tactical redesign decisions of the wholesalers' network and operationally evaluate the solutions obtained. The objective is to minimize the wholesaler total costs without jeopardizing the current customer service level. Because a wholesaler activity is very time sensitive, with multiple orders taking place at the same time and in a large scale, modeling the different operations and their relationships in one mathematical programming model would be very complex and lead to an intractable model. In these types of systems, simulation is a popular approach, since it deals with complex flows with no mathematical sophistication. On the other hand, by using solely simulation models, the number of decision scenarios is rather limited, and hence the optimization can be compromised. Therefore, in order to truly optimize the wholesaler's network and at the same time obtain a clear image of the impact of implementing a new design, both from the operational and marketing points of view, this paper develops an optimization-simulation approach. This approach simplifies the optimization model and at the same time makes it easier to understand the pressure that the redesigning process will put on the operational activities, and how that would reverberate on the customer service level. Firstly, a mixed integer linear programming (MIP) model is formulated at a macro level to optimize the main strategic-tactical decisions involved in the network redesign. Then, the real-world operational conditions of the SC are simulated by means of a discrete event simulation model. This paper thus contributes with a comprehensive approach to the SCNR problem, as well as its instantiation in a real-world case study.

The remainder of this paper is organized as follows. Section 3.2 presents a literature review on network design and redesign problems. The problem characteristics are defined in Section 3.3 and the optimization-simulation methodology adopted is described in Section 3.4. The evaluation of computational results obtained for a pharmaceutical wholesaler case-study, which is described in Section 3.5, is provided in Section 3.6. Section 3.7 concludes with some remarks on the methodology, results and future work.

3.2. Literature review

An SC is composed of a set of facilities from suppliers to customers and processes related to purchasing, inventory control and distribution (Sabri and Beamon, 2000). The number and location of facilities and the allocation of customers to one or more locations are extensively studied in the literature as facility location problems. The facility location problem can be modeled for various real-world situations, for instance supporting the strategic decision of a company's supply chain network design (SCND).

The SCND is considered one of the most important strategic decisions for businesses because it influences all the other SC decisions. There is vast literature on SCND models with different considerations, such as the SC structure, the objective function, the incorporation of uncertainty and the integration of different decisions.

Badri et al. (2013) propose a mathematical model and a Lagrangean relaxation approach for the SCND problem considering different time periods for strategic and tactical decisions. The model incorporates investment decisions throughout time, indicating when a facility should be opened or enlarged. Moreover, a distinction is considered between subcontracted and private facilities, where the former can be closed at a given period and the latter cannot. This can also be applied in a company SCNR, where there are open facilities that are owned and cannot be closed in the redesigning phase.

Eskigun et al. (2005) develop an SCND model for an outbound automotive SC, where vehicles are transported from the assembly plants to dealers, with the possibility of passing through distribution centers (DC) to consolidate the merchandise. To evaluate customer satisfaction, the authors used indicators such as lead times, incorporating in the model the waiting time of the vehicles in the facilities, considering that a vehicle only departures from the DC when the last unit arrives. A similar analogy can be made to the pharmaceutical wholesaler case, where the vehicles only depart when all orders placed by the pharmacies in that route are prepared. To avoid congestion in the DC, Eskigun et al. (2005) limit its capacity to a given number of vehicles. To solve this problem, the authors use a Lagrangean heuristic for a deterministic model, and afterwards address the uncertainty involved in the problem through scenario analysis.

In Baghalian et al. (2013) a robust SCND model is presented that considers uncertainty both from the demand and supply sides. The first type of uncertainty is addressed using a demand distribution function, and the second using the probability of disruption occurrence in the upstream stages of the SC. These considerations help to incorporate costs related to customer service level. Due to the nonlinear relations, the problem is solved using a piecewise linear approximation in order to obtain a linear model. The authors also provide a review of stochastic SCND research, concluding that most of the research only considers demand-side uncertainty.

Sabri and Beamon (2000) propose an approach for a four echelon SCND where strategic and operational levels are integrated simultaneously. Their model incorporates production, delivery and demand uncertainty, as well as a performance vector to evaluate the entire SCND. They employed an ε-constraint method because it is simple and capable of solving non-linear models.

The importance of integrating different decision levels in the SCND problem has been extensively studied and different models and approaches are proposed in the literature. Due to the influence that strategic location decisions have on inventory and distribution planning, ignoring these tactical decisions at network stage leads to sub-optimality (Shen and Qi, 2007; Stadtler and Kilger, 2008; Baghalian et al., 2013; Guerrero et al., 2013). For instance, a higher number of facilities make it possible to reduce the distribution costs, but at the same time it increases inventory costs (Hübner et al., 2013; Zinn et al., 1989). Consequently, models combining the facility location problem (FLP) with the inventory control problem (ICP) and the routing problems (RP) have been proposed.

Location-inventory problems (LIPs) analyze the inventory control policy by adding to the location problem decisions on the facilities' ordering process, both regarding the size and time of order fulfillment. Naturally, integrating these problems increases the models' complexity. Diabat et al. (2013) propose a Lagrangean relaxation-based heuristic to solve the LIP. A set of DCs and retailers that hold working inventory which has not yet been requested by retailers or end-customers, respectively, are considered. The goal is to determine the optimal number and location of DCs, the customer-DC allocation and the respective inventory strategies. A reformulation and piecewise linearization is proposed for this problem by Diabat and Theodorou (2015). Gzara et al. (2014) present a model for designing a responsive service parts logistic system with backorders. The authors consider a time-based service level to guarantee deliveries of service parts within a specific time window.

Location-routing problems (LRPs) include, besides the location decisions, the definition of the best delivery routes from the selected facilities to the assigned customers. Drexl and Schneider (2014) review different variants and extensions of the LRP and conclude that the main difficulty of LRP is managing the location-allocation and routing sub-problems.

One of the most commonly used solution methods is a heuristic approach that decomposes the problem in an iterative sequence of location-allocation and routing sub-problems. By incorporating the routing problem, the distribution costs can be more precisely estimated.

The integration of these three problems, in what is called inventory-location-routing problem (ILRP), was first tackled by Javid and Azad (2010). The authors propose a methodology based on tabu search and simulated annealing for a single period, single product, two echelon SC structure with stochastic demand and (Q,R) inventory policy. Guerrero et al. (2013) propose a matheuristic approach to the ILRP, using an exact method with estimated distribution costs to obtain the SC design, which is then used in a heuristic to optimize the routing decisions.

Due to the influence that these strategic-tactical decisions have on the supply chain network redesign, the approach presented in this paper explores the main interdependencies of the distinct dimensions to better restrict and evaluate different solutions. In the proposed approach, the inventory control policy decision is not incorporated in the optimization model; instead, a (R,s,S) policy (Silver et al., 2009) is embedded in the simulation model, involving operational uncertainty. Moreover, since the stock levels are influenced by the customers' allocation decisions, the benefits of aggregating customers and of resorting to cross-docking facilities (Zinn et al., 1989) are included in the optimization model when determining the safety stock of the warehouses. In addition, because there are times of the day where most pharmacies place orders, generating demand peaks, it is also important to consider the warehouses' capacity, the congestion that might occur at different points in time of the day and their impact on the distribution.

With regard to the pharmaceutical industry, the research mostly focuses on the manufacturers, mainly because of the high costs involved in the R&D of new drugs and their time-to-market-side. Sousa et al. (2011) address an allocation / dynamic planning problem that optimizes the SC of a pharmaceutical company, considering primary and secondary sites with their respective warehouses and final product market areas, neglecting the distribution network inside the market. The objective of the model is to maximize the profit.

Izadi and Kimiagari (2014) present a study on a pharmaceutical distributor in Iran where a Monte Carlo simulation is used to evaluate the demand risk factor. Buil et al. (2010) implement a discrete event simulation model to optimize the configuration of an SC, while analyzing the implementation of innovative processes in the operation. Using this type of simulation models is advantageous for analyzing the possible impacts of implementing a strategy, because it makes it possible to visualize how the system can react to changes. Various companies are already using these models to represent the systems, to detect improvement opportunities on the flow of materials and information, and to test strategies to be adopted (Brown and Sturrock, 2009). Nevertheless, by using simulation models instead of optimization models, the number of decision scenarios is quite limited.

In recent years, the increase in computing performance has made it possible to use methodologies incorporating optimization and simulation models simultaneously (Bartolacci et al., 2012). According to Figueira and Almada-Lobo (2014), there are several methods to hybridize optimization-simulation, which can be classified according to the interaction between the two models and the search algorithm. In this paper, the optimization-simulation methodology used can be classified as Solution Generator (SG). The simulator is only required to evaluate the solution obtained in the optimizer, instead of evaluating each solution found along the search.

3.3. Pharmaceutical Wholesaler Supply Chain Network Redesign

A wholesaler's SC is defined by $\mathcal{W}$ warehouses that distribute $\mathcal{P}$ products to C customers, produced by various suppliers, as illustrated in Figure 3.1.

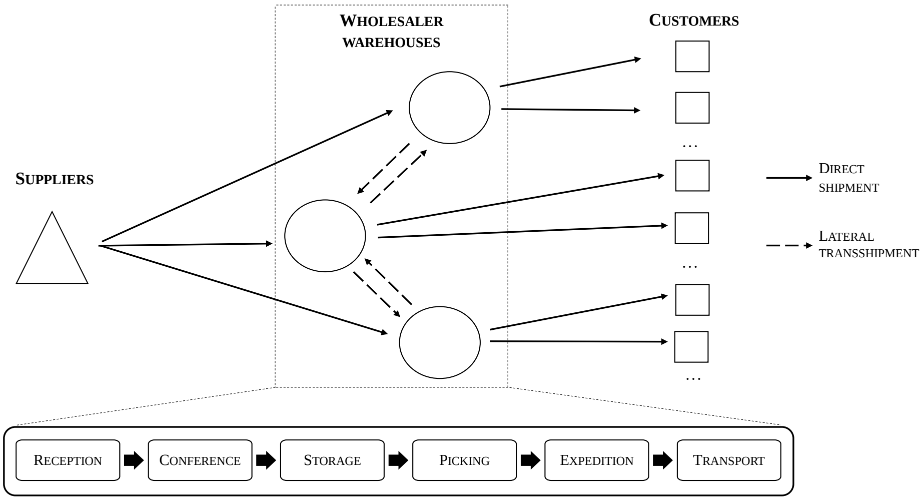


Figure 3.1: Supply chain of a wholesaler

The warehouses can have different functions, $\mathcal{F}$. They can operate as stockists, holding inventory of products, as cross-docking, having the incoming materials directly loaded onto outbound transportation with low or no storage in between, or with both functions, where some products are stored and others are not. Customer demand is satisfied by the inventory held by the wholesalers. The number of stockist warehouses influences the total inventory level. The higher the number of warehouses, the higher the company's total inventory. To reduce the number of stockist warehouses and achieve economies of scale in the transportation, wholesalers can use cross-docking warehouses. These facilities make it possible to send larger vehicles to a cross-docking facility, instead of using small vehicles to each region. In this intermediate facility, the load is divided into smaller vehicles, which perform the last miles of the distribution. When cross-docking warehouses are used, a lateral transshipment channel has to be defined, representing the link between the company facilities, through which internal material is transported.

As shown in Figure 3.1, warehouses perform many activities. The supplier delivers the product and its condition is verified. If the product's quality is ensured, it is stored in accordance with its specifications. When a customer order arrives, the products are picked and forwarded to the expedition zone. The orders are then transported and delivered to customers, following a previously agreed schedule.

Each warehouse performs its activities with an operational level $\mathcal{L}$, which indicates the processing capacity of the activity (similarly to a plant's production capacity). Since the distribution can be made using heterogeneous vehicles with different capacities, an operational capacity level must be defined for each vehicle typology, $\mathcal{V}$.

The network design of a wholesaler is defined by the number, location, function and capacity of its warehouses, the allocation of products and customers to those warehouses, and the capacity and definition of the transportation resources.

The pharmaceutical sector has specific characteristics to be taken into consideration when analyzing the SCNR of a wholesaler. Although the product allocation to warehouses is one of the main decisions when defining a network design, it is not addressed in this paper. As most of the warehouses in this industry are full-liners, they have to cover the complete assortment of products. Therefore, the main strategic-tactical decisions analyzed here are as follows:

- Warehouse location and function
- Allocation of customers to warehouses
- Lateral transshipment definition
- Level of activity of each operation at each warehouse
- Inventory level at each warehouse

Some pharmaceutical wholesalers perform additional services as third-party logistic providers to capitalize and facilitate the utilization of their resources. Considering this material flow in the analysis is important because it consumes distribution capacity. Therefore, two types of material flow will be differentiated, which result in two types of merchandise transportation containers ($\mathcal{M}$):

- Products owned by the wholesaler that are bought, stored, picked and transported to their customers.
- External merchandise that is simply transported (cross-docking operation).

The products owned by the company are picked according to customer orders and packed into containers called tubs, which have standard volumes. This means that even if the customer only orders one product, one tub must be used. External merchandise is managed as a cross-docking operation, only passing through the wholesaler's expedition zone. They are distributed as delivered by the suppliers and transportation units do not have a standard measurement.

3.4. Hybrid Optimization-Simulation Methodology

This paper describes a hybrid optimization-simulation methodology for analyzing and redesigning the pharmaceutical wholesalers' SC. This methodology includes a mathematical programming model to generate solutions according to the parameters defined initially, which are latter analyzed using a discrete event simulation model.

The mathematical programming model optimizes the main decisions of the SCNR through a macro analysis of the wholesaler's operation. The objective of the model is to find a network design solution that reduces the total costs for the wholesaler. In this phase, the problem is analyzed from a strategic perspective and aggregated data are used.

The operational characteristics of a wholesaler's activities are detailed in the discrete event simulation model. The operational activities of the new network design of the wholesaler's SC are evaluated by means of operational indicators, such as vehicle delays.

The purpose of combining simulation and optimization in a single approach is to enhance the solutions obtained. Since in SCNR the company is already established in the market and possesses historical data on the business, this approach can provide valuable indicators about a solution, because it is possible to establish comparisons with the current state. In this Solution Generator (SG) scheme, the simulator is only required to evaluate the solution obtained in the optimizer (Figueira and Almada-Lobo, 2014). The interaction between the two models is illustrated in Figure 3.2.

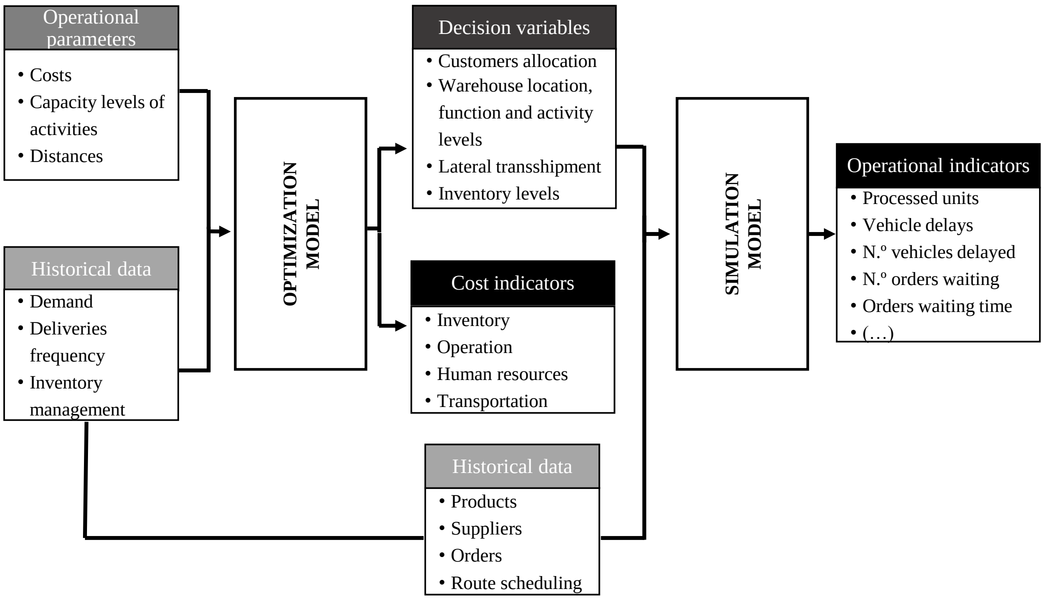


Figure 3.2: Interaction scheme between the optimization and simulation model.

The optimization model is fed with high level data related to the wholesaler's activity, such as customer demand and operational costs. According to the pharmaceutical wholesaler restrictions, a solution is obtained for redesigning the company's network, corresponding to the main strategic decisions incorporated in the model. Moreover, cost indicators are estimated for the respective configuration. Subsequently, the network redesign is imposed in the simulation model. The number and function of warehouses, activity levels and customer allocations are defined according to the optimization model results and the new configuration is simulated, making it possible to analyze the expected operation in a real application. At this stage, a more detailed data set is incorporated regarding customer orders, routing schedules and specifications of products and suppliers. These inputs make it possible to mimic the operation of the warehouses using the same past conditions, such as the time when the customers placed their orders and the suppliers' actual lead time, but applying the new network design. This way, for instance, a customer order might be processed in a different warehouse than in the past, and thus prepared at a different point in time because of the congestion at the new warehouse. At the end of the run, the simulator provides operational indicators regarding the activities performance.

3.4.1 Network Redesign Optimization Model

The network redesign problem is formulated as a mathematical mixed-integer programming (MIP) model, where the main strategic and tactical decisions regarding the SCNR are optimized.

As in traditional facility location problems, the complexity of this model increases with the number of warehouses and customers considered. Due to the high number and density of customers (mostly pharmacies in some regions, normally cities) of a pharmaceutical wholesaler, a clustering of customers it is suggested according to their geographical location. This makes it possible to reduce the number of customer-related variables since in reality these customers will most probably be served by the same route.

The notation used in the formulation is presented below.

Indices and sets

$t \in \mathcal{T}$	time period
$p \in \mathcal{P}$	product
$w, d \in \mathcal{W}$	warehouse
$f \in \mathcal{F}$	warehouse functions
$c \in \mathcal{C}$	clusters of customers
$v \in \mathcal{V}$	vehicle typologies
$l \in \mathcal{L}$	operational levels (of the warehouses)
$m \in \mathcal{M}$	merchandise transportation containers
$n \in \mathcal{N}$	number of break points of linear pieces
P_w	set of products p allocated to warehouse w

Parameters

IS_w	initial stock of warehouse w (units)
D^t_{cp}	demand for product p of customers of cluster c in period t (units)
ASD	average standard deviation of clusters demand in the warehouse average replenishment cycle time (units)
CF_{wp}	capacity consumption factor of product p in conference and picking activities of warehouse w
CC_{wl}	conference activity capacity with operational level l at warehouse w (units)
CP_{wl}	picking activity capacity with operational level l at warehouse w (units)
CT_{wvl}	transport activity capacity of vehicle typology v with operational level l at warehouse w (cubic meters)
FCC_{wl}	fixed cost of conference activity with operational level l at warehouse w (euros)
FCP_{wl}	fixed cost of picking activity with operational level l at warehouse w (euros)
FCT_{wvl}	fixed cost of transport activity for vehicle typology v with operational level l at warehouse w (euros)

MTQ Minimum transportation quantity to serve a cluster (percentage of demand)

CV_{pmw} Conversion factor of product p transported in a container m at warehouse w to volume (cubic meters/unit)

$TVCC_{wc}$ transport activity variable cost for serving cluster c through warehouse w (euros/km)

$TFCW_{wd}$ transport activity fixed cost for serving warehouse d through warehouse w (euros)

OC_{wf} monthly operation cost of warehouse w with function f (euros)

IC_p inventory cost of each unit of product p stocked per time period (euros)

CUD unit cost of unsatisfied demand (euros/unit)

Decision variables

X_{wf} equals 1 if warehouse w is operational with function f, 0 otherwise

Y^f_{wd} equals 1 if warehouse w supplies warehouse d with function f, 0 otherwise

Z_{wc} equals 1 if warehouse w supplies directly cluster c, 0 otherwise

ZA_{wc} equals 1 if warehouse w supplies indirectly cluster c, 0 otherwise

A_{wdc} equals 1 if warehouse w supplies indirectly cluster c through cross-docking on warehouse d, 0 otherwise

NC_{wn} equals 1 if warehouse w supplies n number of clusters, 0 otherwise

QTC^t_{wcpv} quantity of product p transported on vehicle typology v from warehouse w to cluster c at time period t (units)

QTW^t_{wdpv} quantity of product p transported on vehicle typology v from warehouse w to warehouse d at time period t (units)

I^t_{wp} inventory level of product p at warehouse w in time period t (units)

QR^t_{wp} quantity of product p received at warehouse w in time period t (units)

CA_{wl} equals 1 if warehouse w performs the conference activity at operational level l, 0 otherwise

PA_{wl} equals 1 if warehouse w performs the picking activity at operational level l, 0 otherwise

TA_{wvl} equals 1 if warehouse w performs the transport activity with vehicle typology v at operational level l, 0 otherwise

CFC_w conference activity cost at warehouse w (euros/time period)

PFC_w picking activity cost at warehouse w (euros/time period)

TFC_w transport activity cost at warehouse w (euros/time period)

UD unsatisfied demand (units)

This model considers products owned by the wholesaler that are stocked, represented by $p = 1$, and merchandise from the additional (external) services, represented by $p = 2$. It is assumed that the warehouses functioning as cross-dockers ($f = 1$) can receive and send both types of products, while warehouses with stock function ($f = 2$) must have inventory of the first type of products and perform cross-docking with the second. The parameter CV_{pmw}, regarding the conversion of product quantities in transportation volume, is calculated based on the percentage of products that are transported in each container type at each warehouse, the average number of products in each container type and the volume of the container.

For ease of notation, $|T|$, $|W|$ and $|C|$ refer to the cardinalities of the respective sets. The model for finding the optimal network redesign for a wholesaler reads:

$$\begin{aligned}\text{Minimize}\quad & |T|\sum_{w\in W}\sum_{f\in F}X_{wf}\cdot OC_{wf}+\sum_{w\in W}\sum_{p\in P}\sum_{t\in T}I^{t}_{wp}\cdot IC_{p}+|T|\sum_{w\in W}(CFC_{w}+PFC_{w}+TFC_{w})\\ & +\sum_{w\in W}\sum_{c\in C}(TVCC_{wc}\cdot\sum_{t\in T}\sum_{p\in P}\sum_{v\in V}QTC^{t}_{wcpv}/D^{t}_{cp})+|T|\sum_{w\in W}\sum_{d\in W:d\neq w}\sum_{f\in F}Y^{f}_{wd}\cdot TFCW_{wd}\\ & +UD\cdot CUD \end{aligned}\tag{3.1}$$

Objective function (3.1) minimizes the global costs of the SC. The first three terms refer to the operation costs of the warehouses, the inventory holding costs and the fixed costs associated with the warehouses activity levels, respectively. The following two terms relate to the distribution costs. The first is related to the distribution to the customers, considering a variable cost influenced by the distance, the cost per kilometer, the frequency of supplies and the position of the customers in relation to the other customers in the cluster (the more customers there are close to a given customer, the better). Because a cluster can be served by two warehouses, for instance depending on the time of the day, the distribution costs of the cluster must be split by the warehouses. This division is made based on the proportion of the customer demand each warehouse supplies. The second term of the distribution costs refers to the lateral transshipment, where a fixed cost is applied to each connection. The cost of unmet demand is incorporated in the last term.

$$\sum_{f\in F}X_{wf}\leq 1\quad \forall w\in W\tag{3.2}$$

$$\sum_{w\in W:w\neq d}Y^{1}_{wd}\geq X_{d1}\quad \forall d\in W\tag{3.3}$$

$$\sum_{d\in W:d\neq w}\sum_{f\in F}Y^{f}_{wd}\leq(|W|-1)\cdot X_{w2}\quad \forall w\in W\tag{3.4}$$

$$\sum_{f\in F}Y^{f}_{wd}\leq 1\quad \forall w,d\in W:w\neq d\tag{3.5}$$

$$\sum_{w\in W:w\neq d}Y^{f}_{wd}\leq(|W|-1)\cdot X_{df}\quad \forall d\in W,f\in F\tag{3.6}$$

$$\sum_{p\in P}\sum_{v\in V}\sum_{t\in T}QTW^{t}_{wdpv}\leq M\cdot\sum_{f\in F}Y^{f}_{wd}\quad \forall w,d\in W:w\neq d\tag{3.7}$$

Constraints (3.2)-(3.7) are related to the operation of the warehouses. Constraints (3.2) and (3.3) assure that a warehouse can only have one function (or be closed) and that when it operates as cross-docking it must be supplied by another warehouse, respectively. Constraint (3.4) ensures that only a stock warehouse can transfer products to other warehouses. The lateral transshipments are required not only to supply the cross-docking warehouses, but also to transfer orders between stock warehouses, according to the company policies. The frequency of deliveries is different depending on the type of supply. Therefore, the lateral transshipment can have one of two different functions, and can only supply an active warehouse (constraints (3.5) and (3.6)). The transfer of goods between warehouses can only be performed in case there is a connection between them, as established in constraint (3.7). Note that M denotes a big number.

$$\sum_{c\in C}Z_{wc}\leq|C|\cdot\sum_{f\in F}X_{wf}\quad \forall w\in W\tag{3.8}$$

$$\sum_{p\in P}\sum_{v\in V}\sum_{t\in T}QTC^{t}_{wcpv}\leq M\cdot Z_{wc}\quad \forall w\in W,c\in C\tag{3.9}$$

$$\sum_{w \in W} \sum_{v \in V} QTC^t_{wcpv} \leq D^t_{cp} \quad \forall c \in C, p \in P, t \in T \tag{3.10}$$

Constraints (3.8) - (3.10) impose that clusters can only be allocated and receive goods from operational warehouses, limiting the quantities transported in each time period to the cluster's demand.

$$QR^t_{wp} + \sum_{d \in W} \sum_{v \in V} QTW^t_{dwpv} + I^{t-1}_{wp} = \sum_{d \in W} \sum_{v \in V} QTW^t_{wdpv} + \sum_{c \in C} \sum_{v \in V} QTC^t_{wdpv} + I^t_{wp} \quad \forall w \in W, p \in P, t \in T \tag{3.11}$$

$$\sum_{p \in P} \sum_{t \in T} I^t_{wp} \leq M \cdot X_{w2} \quad \forall w \in W \tag{3.12}$$

$$I^0_{w1} = IS_w \cdot X_{w2} \quad \forall w \in W \tag{3.13}$$

$$I^0_{w2} = 0 \quad \forall w \in W \tag{3.14}$$

Constraints (3.11) - (3.14) control the level of inventory according to the function of the warehouses and the type of product. Constraint (3.11) guarantees the material flow balance through each warehouse, at each time period. Constraints (3.12) and (3.13) state that only stock warehouses can hold inventory and set the initial inventory levels for the current active warehouses. Naturally, warehouses that are not activated have null stock at the beginning of the planning horizon. Constraint (3.14) establishes that the products related to external merchandise ($p = 2$) cannot be retained in inventory, being only subject to a cross-docking operation.

Due to the variability of customer orders and the suppliers' replenishment lead time, the stock warehouses have to maintain enough inventory (cycle and safety stock) to provide an adequate service level. The safety stock depends on the number of customers the warehouse supplies and their demand variability. By aggregating customers, it is possible to reduce the global inventory level, if it is assumed that the demands are uncorrelated (Zinn et al., 1989). Thereby, it is possible to calculate an approximation of the safety stock according to the allocation decisions, based on the number of customers a warehouse supplies, incorporating in the model the benefits of aggregating stock, and resorting to cross-docking.

Assuming the cluster demands are uncorrelated and that the variability of demand is the same at all locations, the Square Root Law (SRL) can be applied to estimate how much the safety stock varies when the clusters are aggregated in one warehouse. The SRL is defined as follows.

$$\sigma_a = \sigma_i \cdot \sqrt{n}$$

Where σ_a is the aggregated standard deviation, σ_i the standard deviation of location i and n is the number of locations aggregated. σ_i is the average standard deviation of the cluster demand. Therefore, considering the formula of the SRL, the safety stock of the warehouses can be calculated as follows.

$$I^t_{w1} \geq ASD \cdot \sqrt{\sum_c (Z_{wc} + Z_{dc} \cdot Y_{w,d}) - M \cdot (1 - X_{w2})} \quad \forall w, d \in W, t \in T \tag{3.15}$$

Note that Z_{wc} indicates the clusters that warehouse w supplies directly and $Z_{dc} \cdot Y_{w,d}$ indicates whether warehouse w supplies cluster c through cross-docking at warehouse d. Summing $Z_{wc} + Z_{dc} \cdot Y_{w,d}$ gives the number of clusters aggregated in warehouse w.

The assumption of equal variability throughout all locations is a simplification that leads to an overestimation of the advantages of aggregating stocks. However, in this problem this assumption yields an adequate approximation to the model proposed by Zinn et al. (1989), considering different demand variabilities. Figure 3.3 compares the results achieved by the SRL and the model proposed by Zinn et al. (1989), where two methods for aggregating locations are used (Zinn 1 and Zinn 2). The results show that the overestimation is more significant when the number of locations aggregated is high.

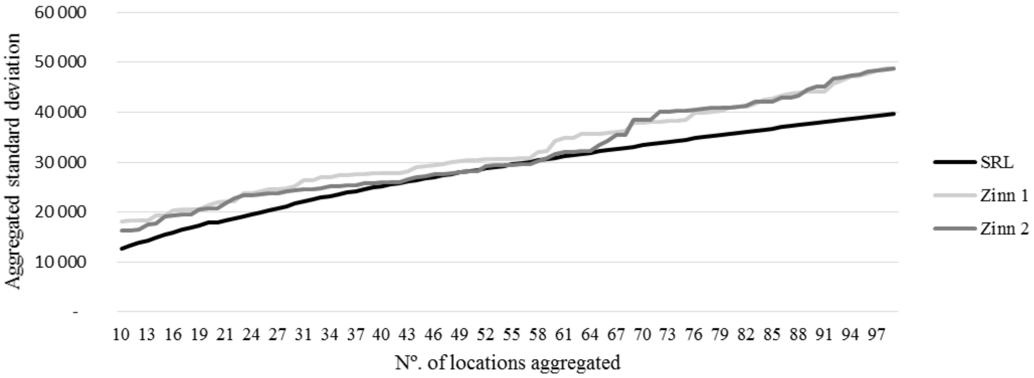


Figure 3.3: Comparison of the aggregated standard deviation using the SRL and the Zinn et al. (1989) model

$$\sum_{l \in L} CA_{wl} = 1 \quad \forall w \in W \tag{3.16}$$

$$\sum_{l \in L} PA_{wl} = 1 \quad \forall w \in W \tag{3.17}$$

$$\sum_{t \in T} QR^t_{wp} \leq M \cdot X_{w2} \quad \forall w \in W, p \in P_w \tag{3.18}$$

$$\sum_{p \in P_w} QR^t_{wp} \cdot CF_{wp} \leq \sum_{l \in L} CC_{wl} \cdot CA_{wl} \quad \forall w \in W, t \in T \tag{3.19}$$

$$\sum_{p \in P_w} \sum_{v \in V} (\sum_{d \in W} QTW^t_{wdpv} \cdot CF_{wp} + \sum_{c \in C} QTC^t_{wcpv} \cdot CF_{wp})$$

$$\leq \sum_{l \in L} CP_{wl} \cdot PA_{wl} + M \cdot X_{w1} \quad \forall w \in W, t \in T \tag{3.20}$$

Constraints (3.16) - (3.20) define the operating conditions of the conference and picking activities. These warehouse activities are modeled with levels that limit the available capacity (constraints (3.16) and (3.17)) (Guimarães et al., 2012). Constraint (3.18) imposes that only stock warehouses can receive goods from suppliers and only for products that are allocated to the respective warehouse. Constraints (3.19) and (3.20) guarantee that the total amount of units processed in the conference and picking activities at each period of time does not exceed the capacity of the activity's operational level defined.

$$\sum_{l\in L} TA_{wvl} = 1 \quad \forall w \in W, v \in V \tag{3.21}$$

$$\sum_{p\in P_w}\sum_{m\in M}\Big(\sum_{c\in C} QTC^t_{wcpv}\cdot CV_{pmw} + \sum_{d\in W} QTW^t_{wdpv}\cdot CV_{pmw}\Big) \le \sum_{l\in L} CT_{wvl}\cdot TA_{wvl} \quad \forall w \in W, v \in V, t \in T \tag{3.22}$$

When it comes to distribution (outbound logistics), constraints (3.21) and (3.22) define the level of activity for each vehicle typology in each warehouse, in a way that the volume transported at each time period meets the capacity set for that level.

$$CFC_w \ge \sum_{l\in L} FCC_{wl}\cdot CA_{wl} \quad \forall w \in W \tag{3.23}$$

$$PFC_w \ge \sum_{l\in L} FCP_{wl}\cdot PA_{wl} \quad \forall w \in W \tag{3.24}$$

$$TFC_w \ge \sum_{v\in V}\sum_{l\in L} FCT_{wvl}\cdot TA_{wvl} \quad \forall w \in W \tag{3.25}$$

$$UD \ge \sum_{w\in W}\sum_{c\in C}\sum_{p\in P}\sum_{v\in V}\sum_{t\in T} QTC^t_{wcpv} - \sum_{c\in C}\sum_{p\in P}\sum_{t\in T} D^t_{cp} \tag{3.26}$$

Constraints (3.23)-(3.26) define the cost terms of the objective function. The fixed costs per time period related to the conference, picking and distribution activities that depend on the activity level set are determined in constraints (3.23)-(3.25), respectively. Constraint (3.26) computes the unmet demand. Note that back ordering is not allowed.

Finally, the variable domains are defined as follows:

$$X_{wf}, Y^f_{wd}, Z_{wc}, NC_{wn}, CA_{wl}, PA_{wl}, TA_{wl} \in \{0,1\};$$
$$QTC^t_{wcpv}, QTW^t_{wdpv}, I^t_{wp}, QR^t_{wp}, CFC_w, PFC_w, TFC_w, UD \ge 0 \tag{3.27}$$

Remark. The model SCNR (3.1)-(3.27) is non-linear because of constraint (3.15), which contains the warehouses' safety stock (represented below). Therefore, to linearize the model the constraint has to be reformulated and new constraints must be added.

$$I^t_{w1} \ge ASD\cdot \overbrace{\sqrt{\sum_c (Z_{wc} + \underbrace{Z_{dc}\cdot Y_{w,d}}_{\text{term 1}})}}^{\text{term 2}} - M\cdot(1-X_{w2}) \quad \forall w, d \in W, t \in T$$

Similarly to Diabat and Theodorou (2015), term 1 is linearized using an auxiliary variable and term 2 by means of a piecewise linearization. Let $A_{wdc} = Z_{dc}\cdot Y_{w,d}$ and ZA_{wc} be a binary variable that equals 1 if warehouse w supplies c indirectly. Now, the number of clusters supplied (directly and indirectly) by a warehouse can be calculated as $Z_{wc} + ZA_{wc}$, subject to the following auxiliary constraints.

$$A_{wdc} \ge Z_{dc} + Y^1_{wd} - 1 \quad \forall w, d \in W, c \in C \tag{3.28}$$

$$\sum_{w} A_{wdc} \le Z_{dc} \quad \forall d \in W, c \in C \tag{3.29}$$

$$\sum_{c} A_{wdc} \le M\cdot Y^1_{wd} \quad \forall w, d \in W \tag{3.30}$$

$$\sum_{d} A_{wdc} \leq M \cdot ZA_{wc} \quad \forall w \in W, c \in C \tag{3.31}$$

$$ZA_{wc} \leq \sum_{d} A_{wdc} \quad \forall w \in W, c \in C \tag{3.32}$$

Constraints (3.28)-(3.30) define the auxiliary variable A_{wdc} as $Z_{dc} \cdot Y_{w,d}$, indicating whether cluster c is supplied indirectly by warehouse w, through cross-docking in warehouse d. Constraints (3.31) and (3.32) identify the clusters that are supplied indirectly by warehouse w, independently of the cross-docking warehouse.

Since an active warehouse can supply from one to all clusters, the term 2 is divided into $n = 1, ..., |C|$ linear pieces. Let NC_{wn} be equal to 1 in case warehouse w supplies n clusters, and 0 otherwise.

$$Z_{wc} + ZA_{wc} = \sum_{n} n \cdot NC_{wn} \quad \forall w \in W, c \in C \tag{3.33}$$

$$\sum_{n} NC_{wn} \leq 1 \quad \forall w \in W \tag{3.34}$$

$$I^{t}_{w1} \geq ASD \cdot \sqrt{n} - M \cdot (2 - X_{w2} - NC_{wn}) \quad \forall w \in W, n \in N, t \in T \tag{3.35}$$

$$Z_{wc} \leq \sum_{p \in P} \sum_{v \in V} \sum_{t \in T} QTC^{t}_{wcpv} \quad \forall w \in W, c \in C \tag{3.36}$$

$$\sum_{v \in V} QTC^{t}_{wcpv} \geq MTQ \cdot D^{t}_{cp} \cdot Z_{wc} \quad \forall w \in W, c \in C, p \in P, t \in T \tag{3.37}$$

By adding two more constraints to identify the linear piece that is activated (3.33) and (3.34), the safety stock level of a stock warehouse can be reformulated as in constraint (3.35). Constraints (3.36) and (3.37) need to be added to restrict the allocation variable even more and impose a minimum transportation quantity to serve a cluster, respectively. Otherwise, the model would allocate all clusters to the open stock warehouses, even when not supplied, in order to reduce the inventory.

Note that all clusters supplied by a cross-docking warehouse are indirectly allocated to the warehouses that supply it and that one cluster can be directly allocated to more than one warehouse. Here it is assumed that all warehouses to which the cluster is allocated have to cover its demand, as only at an operational level it is decided which of the warehouses will supply the cluster in each occasion.

The linear SCNR (denoted as SCNR^{lin}) reads:

$$\begin{aligned}
\text{SCNR}^{lin} \quad \text{Minimize} \quad & |T| \sum_{w \in W} \sum_{f \in F} X_{wf} \cdot OC_{wf} + \sum_{w \in W} \sum_{p \in P} \sum_{t \in T} I^{t}_{wp} \cdot IC_{p} \\
& + |T| \sum_{w \in W} (CFC_{w} + PFC_{w} + TFC_{w}) \\
& + \sum_{w \in W} \sum_{c \in C} (TVCC_{wc} \cdot \sum_{t \in T} \sum_{p \in P} \sum_{v \in V} \frac{QTC^{t}_{wcpv}}{D^{t}_{cp}}) \\
& + |T| \sum_{w \in W} \sum_{d \in W : d \neq w} \sum_{f \in F} Y^{f}_{wd} \cdot TFCW_{wd} \\
& + UD \cdot CUD
\end{aligned}$$

subject to: $(3.2) - (3.14)(3.16) - (3.37)$

$$X_{wf}, Y^{f}_{wd}, Z_{wc}, ZA_{wc}, A_{wdc}, NC_{wn}, CA_{wl}, PA_{wl}, TA_{wl} \in \{0,1\};$$

$$QTC^{t}_{wcpv}, QTW^{t}_{wdpv}, I^{t}_{wp}, QR^{t}_{wp}, CFC_{w}, PFC_{w}, TFC_{w}, UD \geq 0$$

In order to better understand the dynamics of the model, an illustrative example is provided below.

3.4.1.1 Illustrative Example

Figure 3.4 presents a toy instance for the network redesign problem. In this example, there are four possible locations for warehouses (w) that have to supply just one type of product (to simplify the index, p is omitted in this example) and meet the demand of five customers (c). Only one period of time is optimized (index t is eliminated from the variables). The graph shows the base capacity of the warehouse activities (CC_{wl}, CP_{wl}), customer demand (D^{t}_{c}), and the distance between adjacent entities. An average standard deviation of sixty ($\sigma = 60$) is considered for the cluster demand. Full line arrows represent a distribution line between warehouses and customers, and dashed arrows represent a distribution line between warehouses.

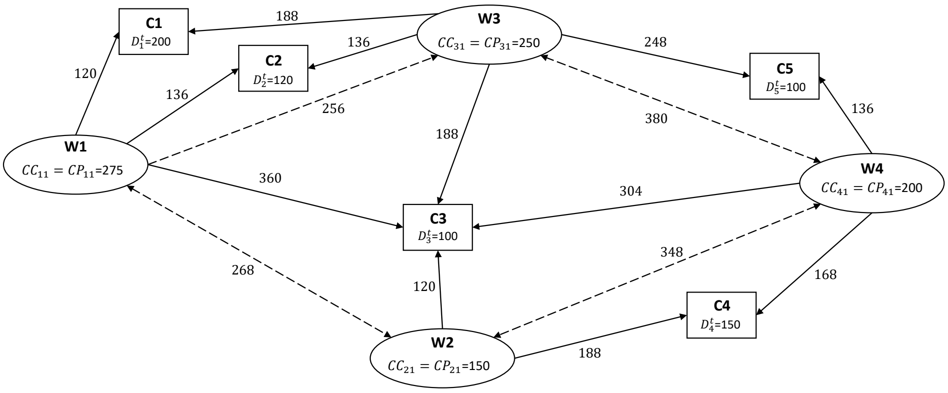


Figure 3.4: Network design illustrative example

The warehouses' activities may have 1 of 3 operational levels, being the value of their fixed costs equal to the double of their capacity ($FCC_{wl} = 2 \cdot CC_{wl}$, for the conference activity). For the conference and picking activities, besides the base level presented ($l = 1$), the level $l = 0$ indicates that the activity is not performed and the level $l = 2$ refers to the double of the base capacity. The distribution activity can be performed by two vehicle typologies, $v = 1$ and $v = 2$. Table 3.1 shows the capacity and fixed cost of each operational level. The same conditions are assumed for all warehouses.

An active stock warehouse costs \$10.000, with zero initial stock, whilst a cross-docking warehouse costs \$200. To calculate the variable distribution cost, a frequency of one and two daily deliveries are considered for cross-docking and stock warehouses, respectively, 31 days a month. The cost per unit of distance for deliveries between warehouses and customers is 0.6 and 0.3 between warehouses.

Table 3.1: Operational levels of the transport activity for all warehouses.

Vehicle typology (v)	**Activity level** (l)	**Capacity** (CT_{wvl})	**Fixed cost** (FCT_{wvl})
1	0	0	0
1	1	150	50
1	2	250	100
2	0	0	0
2	1	250	110
2	2	400	200

Figure 3.5 illustrates the optimal network design solution to model $SCNR^{lin}$ for this instance. Values are represented for the main strategic decision variables, related to the warehouse function and capacity, customer allocations and quantities transported.

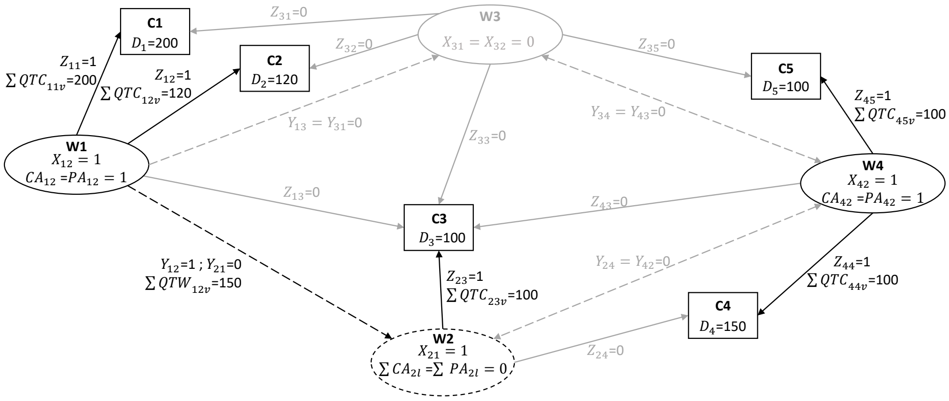


Figure 3.5: Network design illustrative example solution

Warehouses 1 and 4 are open as stock warehouses (X_{12} = 1 ; X_{42} = 1), operating both with conference and picking activities at level 2. The former supplies customers 1 and 2, and the latter customers 4 and 5. Warehouse 2 is open as a cross-docking (X_{21} = 1) with a null conference and picking capacity. It receives products from warehouse 1 (Y^1_{12} = 1) which are supplied to customer 3 indirectly (ZA_{13} = 1). Regarding the distribution, warehouses 2 and 4 only use one vehicle typology, the former uses typology 2 (v = 2) with the base capacity level and the latter uses typology 1 (v = 1) at maximum capacity (level 2). Warehouse 1 uses both vehicle typologies operating typology 1 with the maximum capacity and typology 2 with the base capacity. Note that warehouse 3 is not part of the network.

3.4.2 Operational Discrete Event Simulation

The aim of the discrete-event simulation model is to replicate in detail the wholesalers activities and add to the solution previously obtained by the optimization model operational indicators that support its evaluation.

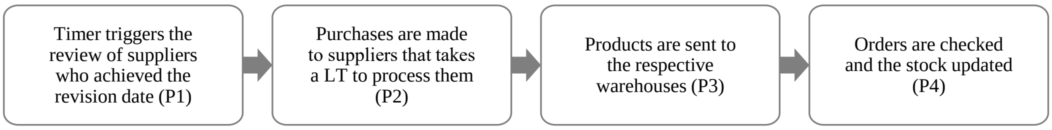
Simulating real-world operational conditions makes it possible to understand the pressure that the new design will cause on the operational activities and how it will impact the customer service level. Therefore, all activities performed by a warehouse are depicted in the simulation model, from the procurement of products to the loading and departure of vehicles from the shipping dock (see Figure 3.1).

All the warehouses that should function as stockists are analyzed in this model according to the solution provided by the optimization model. As mentioned in Section 3.3, the cross-docking considered in this problem is a simple exchange of orders between vehicles (without any rearrangement or consolidation with orders from other vehicles), which can be considered as a transshipment process. Therefore, since the operational stress takes place in the orders preparation at the stock warehouse, the cross-docking warehouses are not included in the simulation model. The cross-docking operation will only be affected in case a delay occurs upstream in the SC. Similarly, the operation of the additional services performed by the wholesaler as a third-party logistic provider is not considered in this analysis, as it does not influence the activities evaluated in this model, but solely the distribution activity capacity, which is already analyzed in the optimization model.

There are three major actions that trigger events in the wholesalers' SC: procurement, lateral transshipment and customer orders. Customer orders trigger the picking of products that reduces the inventory, which is later replaced by the procurement and lateral transshipment actions.

The simulator is designed to replicate real-world conditions so that the material flow is identical to what happens in reality, going through the several phases according to the causal relationships created by the different actions. The way the events triggered by the three defined actions (represented in Figures 3.6-3.8) are formulated in the simulator and will be explained in the remaining of this Section.

3.4.2.1 Procurement


Timer triggers the review of suppliers who achieved the revision date (P1)

Purchases are made to suppliers that takes a LT to process them (P2)

Products are sent to the respective warehouses (P3)

Orders are checked and the stock updated (P4)

Figure 3.6: Simulated procurement process

The procurement process (Figure 3.6) is triggered by the inventory policy of the wholesaler (sub-process P1). This paper considers an (R,s,S) inventory policy. This policy generates a supplier review each time the inventory position is below the minimum stock (s) at the periodic review (R). Each time a supplier is reviewed, the quantities to order for all products are calculated based on the maximum level (S) for each warehouse (sub-processes P2-P3). When the goods are received at the warehouse, the products are verified and stocked by the First-in-First-out (FIFO) rule, with stochastic processing times (sub-process P4).

The new design of the SC affects this process in the number of procurement orders made (procurement activity) and quantities ordered (conference activity). Clearly, if a warehouse accommodates more demand, it may have more orders to perform and more quantities to verify. Therefore, from this process it is possible to analyze the procurement and conference activities.

3.4.2.2 Lateral transshipment

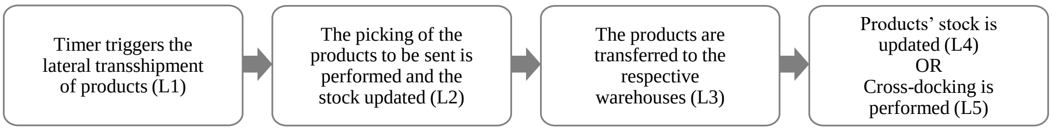

Timer triggers the lateral transshipment of products (L1)
The picking of the products to be sent is performed and the stock updated (L2)
The products are transferred to the respective warehouses (L3)
Products' stock is updated (L4) OR Cross-docking is performed (L5)

Figure 3.7: Simulated lateral transshipment process

The lateral transshipment process (Figure 3.7) is triggered every day according to the supply policy defined for serving warehouses that have a stock function and cross-docking function. The products are picked according to the destination and sent to the warehouses (sub-processes L2-L3), where either the stock is updated (sub-process L4) or a cross-docking operation is performed (sub-process L5).

3.4.2.3 Customer orders

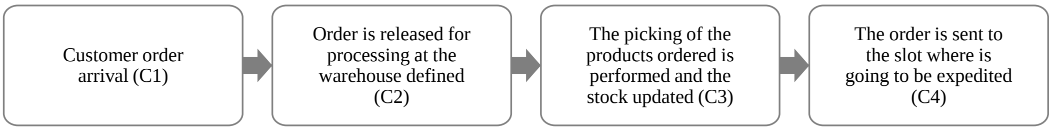

Customer order arrival (C1)
Order is released for processing at the warehouse defined (C2)
The picking of the products ordered is performed and the stock updated (C3)
The order is sent to the slot where is going to be expedited (C4)

Figure 3.8: Simulated customers orders process

The dispatching of the orders (Figure 3.8) is triggered by the arrival of a customer request, which is immediately sent for processing by the FIFO rule (sub-process C1). The orders are picked in the warehouse where the client is allocated and expedited in a predefined route (sub-processes C2-C3). Each vehicle waits for all orders made before the route departure time. The orders placed afterwards are transferred automatically to the next route of the respective client (sub-process C4).

The number of products in each order and the respective amounts are determined by probability distributions. Moreover, the products are randomly chosen according to the probability of selecting each one. These probability distributions are defined for each warehouse based on the demand history. When a vehicle arrives at the warehouse, it waits until the slot defined for its route is free to park and then starts expediting its orders. If its respective order processing is delayed due to congestion in the picking activity, the departure may also be delayed.

The two processes previously described, lateral transshipment and customer orders, affect the picking and expedition activities, which in turn will influence the distribution activity and ultimately the customer service level. The two activities analyzed in this process make it possible to define indicators related to the congestion in the picking area (processing and waiting time) and in the expedition zone (waiting time and departure delays).

3.5. Pharmaceutical Wholesaler Case-study

The case study presented in this paper focus on one of the leading distributors of pharmaceutical products operating in a European country. It commercializes approximately 17,000 products from over 300 different suppliers to more than 1,500 pharmacies.

The group mainly operates as a pharmaceutical wholesaler, buying and storing products to be sold to customers (mainly pharmacies) in the future. Despite the difficulties of the market, with several pharmacies closing down, the company has struggled to increase its market share. Due to the fragility of the sector, the company's strategy not only focuses on growth, but also on differentiating itself from the competition by the quality of its service. Currently, the group operates in the domestic market through five warehouses. Two of them are large-scale warehouses with automated operations, P1 and S1, being the first the main warehouse of the company. To be closer to the customers, three smaller secondary warehouses, S2, S3 and S4, with a manual operation, are also used by the group.

This company performs two additional businesses besides the normal wholesaler operation that may influence the different activities. As presented before in Section 3.3, these businesses can be divided into two classes of products: products related to the normal operation of the wholesaler ($p = 1$), which affects all activities, and products related to additional services ($p = 2$), which only affect the distribution activity. The second case requires the definition of lateral transshipment movements between the larger warehouses, P1 and S1, and the remaining warehouses, because only the former receives the products from the suppliers of the additional services ($p = 2$). Moreover, as warehouse P1 is the only one operating at night, all orders placed at this period are prepared there and distributed in the morning to all the other warehouses. The orders are subsequently forwarded to the customers. In order to incorporate this last specification in the $SCNR^{lin}$ model, it is necessary to include the constraints presented below.

$$\sum_{p \in P_w} \sum_{v \in V} QTW^t_{wdpv} \leq PND_d \cdot \sum_{c \in C} \sum_{p \in P_w} Z_{dc} \cdot D^t_{cp} + M \cdot X_{d1}$$
$$\forall w \in W : w = P1, d \in W : d \neq w, t \in T \quad (3.38)$$

$$\sum_{p \in P_w} \sum_{v \in V} QTW^t_{wdpv} \geq PND_d \cdot \sum_{c \in C} \sum_{p \in P_w} Z_{dc} \cdot D^t_{cp} - M \cdot X_{d1}$$
$$\forall w \in W : w = P1, d \in W : d \neq w, t \in T \quad (3.39)$$

Constraints (3.38) and (3.39) guarantee that a pre-defined percentage of demand of night orders (PND_d) is processed in warehouse P1 and transferred to the other warehouses with a stock function.

Besides analyzing the location and activity of the current warehouses, the goal with this study is also to assess whether there is another configuration of the SC that could reduce the operational costs. Therefore, three new possible locations are considered in the analysis, in addition to the five warehouses currently operated by the company. These new locations were defined together with the company's managers.

Due to the importance that the company gives to the quality of its service, it is necessary to evaluate the impact on customer service, as well as, the cost reduction. Therefore, the methodology proposed in this paper can serve as a guideline and perform what-if analysis for the company's SCNR.

The company's customers were grouped into 100 clusters using the *k*-means algorithm and the demand for the two types of products was assigned to each cluster. This company works with two types of merchandise containers for each product and three different vehicle typologies can be used for transportation. All the activities addressed in the optimization model consider three operational levels.

The computational results for this case study are presented and analyzed in the following section. The instances used in the optimization model can be made available upon request.

3.6. Computational Results

The experiments were performed on a PC with Intel(R) Core i7 CPU with 3.4 GHz and 32 GB RAM. The optimization model was implemented using ILOG OPL Studio as a modeling environment and its incorporated CPLEX 12.5 solver. The simulation model was implemented on Simio simulation software version 6.97.

Results are discussed in three subsections: (1) a validation of the models to confirm the adherence of the models to the real-world setting and to determine indicators that are used as baseline; (2) the improvements that may be achieved by optimizing the current SCND; and (3) the redesign of the SC.

The statistics of the MIP models for cases (2) and (3) are presented in Table 3.2.

Table 3.2: Model statistics

	(2) Current SCND	(3) SCNR
Continuous variables	37,696	60,025
Binary variables	575	1,048
Constraints	3,628	4,803

The optimization model considered monthly time periods and analyzed one year of operation. The model provides an optimal solution in less then twenty minutes. The cost indicators provided by the optimization model for each scenario are compared considering the following categories: inventory costs (Inventory), related to the level of inventory required and the correspondent holding costs; warehouse operation costs (Operation), associated with the opening and function of the warehouses; activity fixed costs (Activities), defined by the level of activity selected; direct-shipment (Direct shipment) and lateral transshipment (Lateral transshipment) costs, associated with the transportation to customers and between warehouses, respectively.

Due to the detail incorporated in the simulation model, with all products and orders listed, it is not feasible to simulate a long time horizon within an acceptable running time that would allow for multiple iterations (the simulator takes around three hours to simulate a month of operation). The company only had detailed and quality data for the year considered in the optimization model. Since the demand is seasonal and the company wanted to design their network for the peak of operation, the peak month was used in the simulation model. According to the company's coordinators this is the month when the operation is more in distress (month of the colds), specially in the main warehouses (P1 and S1) because the picking is automatic and the process cadence can only be increased slightly, without further investment. Therefore, it was critical to analyze how possible changes in the network would affect the operation of the warehouses. Analyzing the operation for the complete peak period makes it possible to build and validate the model. It is assumed that if a solution presents a good operational performance in the critical month it will probably also work smoothly in the remaining months. Note that in the presence of more quality data additional statistical validations should be performed in order to better define the period that should be tested (Sargent, 2014), taking into consideration the trade-off between the running time and the intended goal of the simulation - analyze the activity's performance.

3.6.1 Validation of the Models

Before testing alternative networks, the models are validated against the current scenario of the SC. This makes it possible to verify the accuracy of the models and the data, and to determine the indicators that are used as baseline.

3.6.1.1 Optimization model

Real costs (in percentage of the overall costs) for each category are compared to the ones given by the optimization model when the current network design of the SC is applied (a linear programming model was run by fixing all the binary decision variables of the model proposed). The results, shown in Table 3.3, demonstrate that the optimization model indicators have a good fit to the real-world values. Moreover, operational indicators of each activity, such as the number of units processed in the conference and picking activities and the number of kilometers made, are also compared and the deviation obtained is less than 1%.

Table 3.3: Comparison of cost indicators from the optimization model with the real-world values for baseline scenario.

	Inventory	Operation	Activities	Direct shipments	Lateral transshipments	Total
Real value	9.5%	8.4%	20.7%	57.1%	4.3%	100.0%
Model value	9.6%	8.4%	20.7%	57.2%	4.1%	100.1%
Deviation	+0.1%	0.0%	0.0%	+0.1%	-0.2%	+0.1%

3.6.1.2 Simulation model

The simulation model was validated using some of the techiques described by Sargent (2014), such as data validity, face validity and model behavior analysis. All the flows and connections between the different processes of the company were analyzed and the conceptual model of the simulator was validated by the company's coordinators. Moreover, the model behavior was analyzed using the system input/output data described in Figure 3.2 for the simulation model, where real input data from the month in analysis was used. Simple indicators, such as the number of units processed in the conference and picking activities were compared. Although small deviations to the real values were obtained, they could be explained by the lack of information regarding the in-transit inventory at the beginning of the period. Additionally, the behavior of the procurement process triggered by the inventory policy of the wholesaler was also analyzed and the level of inventory simulated is similar to the real case, as presented in Figure 3.9. Hence, the overall model behavior was validated.

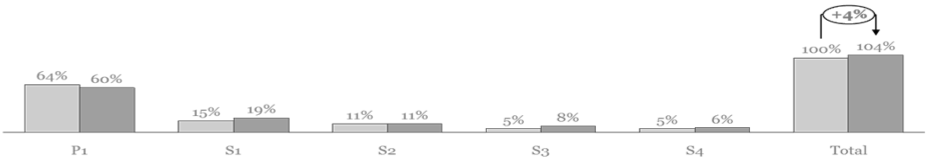


Figure 3.9: Comparison of the inventory levels for each warehouse from the simulation model (dark grey) with the real-world values (light grey)

The operational indicators, which complement the ones provided by the optimization model, are presented in Tables 3.4 and 3.5. Note that the real data for these indicators is not available, but they were validated by the respective activity coordinators of the company.

Table 3.4 indicates the current average congestion that the conference and picking activities face when orders are processed. The indicators considered in the expedition activity, Table 3.5, are used to understand in more detail the impact of the SC redesign. The first indicates the average delay of departing vehicles, and the last two refer to the percentage of vehicles that depart within a given delay. It is worth noting that delays up to 20 minutes are usually recoverable during the distribution route.

Table 3.4: Simulation indicators for conference and picking activities for the baseline scenario.

Indicator	Conference	Picking
Average number of orders waiting to be processed	71.7	0.5
Average waiting time per order	3.5 days	1.1 min

All the validations were made together with the company's coordinators and made it possible to determine that the model is suitable for the intended purpose. Therefore, the results achieved in this scenario are used as baseline, for comparison to the alternative scenarios.

Table 3.5: Simulation indicators for expedition activity for the baseline scenario.

Indicator	Expedition
Average delay time	13.1 min
% travels delayed between 10 to 20 minutes	17%
% travels delayed more than 20 minutes	8%

3.6.2 Optimization of the Current Supply Chain Network Design

Once the models are validated, the first aim is to understand whether modifications can be performed to reduce the total costs without changing the current network design (i.e. maintaining the current five warehouses), causing minimum disruptions to the current SC operation. Denoted as C-SCND, this scenario shows the improvements that may be achieved by optimizing the current SCND.

To test this scenario, the binary variables of the optimization model related to the network design (X_{wf} and Y^{f}_{wd}) are fixed to the values of the current design, only making it possible to optimize the activities operational level and customer allocation.

The results indicate that this scenario may lead to annual savings of about 3%. Table 3.6 compares the cost indicators of this solution against the ones obtained for the baseline model. Table 3.7 indicates the number of people to reallocate for the conference and picking activities, according to the operational level selected, and the difference of total kilometers made for the distribution activity.

Table 3.6: Comparison of cost indicators from the baseline and C-SCND scenarios.

Cost indicator	Inventory	Operation	Activities	Direct shipments	Lateral transshipments	Total
Baseline scenario	9.6%	8.4%	20.7%	57.2%	4.1%	100.1%
C-SCND scenario	9.2%	8.4%	21.0%	54.7%	4.1%	97.3%

Table 3.7: Difference of the operational activities state between the C-SCND and baseline scenarios.

Warehouse	P1	S1	S2	S3	S4	Overall
Conference (person)	+1	+1	0	0	0	+2
Picking (person)	-2	0	+3	-1	0	0
Transport (% kms)	-9%	+1%	+3%	-3%	+3%	-4.2%

The results indicate minor changes that can be performed in the SC operation to yield significant gains. The direct shipment is the cost category that provides the greatest savings due to the 4% decrease in kilometers traveled (Table 3.7) as a result of the customers' reallocation to other warehouses. This reallocation modifies the number of units processed in the conference and picking activities at the warehouses, causing a reallocation of staff that increases the fixed costs of these activities (Table 3.6). However, because the conference activity is reinforced (Table 3.7), the level of inventory is lower, thus reducing the inventory holding costs. In the picking activity, warehouse S2 is reinforced with three more people because of its closeness to customers, and the operational capacity of its activities is superior to other warehouses (note that this capacity is differentiated in the model by means of parameter CP_{wl}).

3.6.3 Optimization of the Supply Chain Network Redesign

With regards to the optimization of the supply chain network redesign (SCNR scenario), new potential locations are considered for warehouses, other than the five the company already has. The goal is to question whether the current location of the warehouses is appropriate or whether there are other locations that could minimize the operational costs of the SC. Moreover, the idea is to discuss the function of the current warehouses. Model $SCNR^{lin}$ has been solved to optimality. The results indicate that the configuration that minimizes the global operational costs of the company is maintaining the five warehouses. However, the function of warehouse S2 should be changed to cross-docking, and supplied by the main warehouse P1.

The variations in the cost indicators and in the operational level of the conference, picking and transportation activities are presented in Tables 3.8 and 3.9. As warehouse S2 is changed to cross-docking, no conference and picking activities are performed; therefore, their level is set to zero. Table 3.9 only shows the variation for the transport activity.

Table 3.8: Comparison of cost indicators between the SCNR and baseline scenarios.

Cost indicator	Inventory	Operation	Activities	Direct shipments	Lateral transshipments	Total
Baseline scenario	9.6%	8.4%	20.7%	57.2%	4.1%	100.1%
SCNR scenario	8.9%	7.9%	20.0%	54.3%	4.6%	95.6%

Table 3.9: Difference of the operational activities state between the SCNR and baseline scenarios.

Warehouse	P1	S1	S2	S3	S4	Overall
Conference (person)	+1	+1	-	0	+1	+1
Picking (person)	+3	0	-	+1	+1	-6
Transport (% kms)	-12%	+1%	+5%	-2%	+3%	-4.7%

Redesigning the SC allows for savings of about 4.4%, the highest savings resulting from a reorganization of the distribution to customers (see Table 3.8). As warehouse P1 supplies warehouse S2 in cross-docking, its conference and picking activities have to be reinforced with four more people (Table 3.9). Moreover, this change triggers the need to supply warehouse S2 more frequently, thus increasing the costs of lateral transshipments. However, the warehouse operation costs are reduced.

Looking into more operational indicators obtained through the simulation model, it is possible to conclude that the picking and expedition activities are the most affected activities in the SCNR scenario. Table 3.10 demonstrates the pressure that the allocation of warehouse S2 puts on the activities of warehouse P1. The average number of orders waiting to be processed in the picking activity is expected to increase 10 units, having each an average waiting time of 5 more minutes.

Table 3.10: Difference of the picking activity operational indicators between SCNR and baseline scenarios.

Warehouse	P1	S1	S2	S3	S4
Average number of orders waiting	+10.2	0	-	-0.1	-0.1
Average waiting time (min)	+4.6	-0.1	-	-0.8	-0.7

Because of the stress created in the preparation of the orders, the vehicles leave warehouse P1 on average 7 minutes later than in the baseline model (Table 3.11).

Table 3.11: Difference of the expedition activity operational indicators between SCNR and baseline scenarios.

Warehouse	P1	S1	S2	S3	S4
Average delay time (min)	+6.9	0	-	-0.8	-1.7
% travels delayed between 10 to 20 minutes	-1%	0%	-	-1%	+1%
% travels delayed more than 20 minutes	+14%	0%	-	0%	-1%

In this new scenario, the number of travels that are delayed for more than 20 minutes increases 7%. For warehouse P1, the increase is 14% and, therefore, its service level is significantly reduced. Besides the delays in the vehicles' departure from the stock warehouse, the time required to transport the goods from warehouse P1 to S2, together with the time required for deconsolidation, must be considered. Moreover, the reduction of daily deliveries in the region of warehouse S2 has to be taken into consideration as well. In the current scenario the customers can have up to 5 deliveries per day, but with the warehouse working as cross-docking this frequency is forced to decrease to 3 deliveries, according to company policies. A higher frequency jeopardizes projected savings.

3.6.3.1 What-if Analysis

In the previously discussed SCNR scenario, the delays are related to the higher congestion in the picking activity as a result of an increase in the number of orders to be processed in warehouse P1.

The pharmacies place more orders at lunch time (for afternoon replenishment), making the picking activity at this hour rather pressured even in the current design. Therefore, a new simulation is conducted with the design obtained in SCNR scenario (maintaining five warehouses, with warehouse S2 changed to cross-docking), hereafter called SCNR scenario 1, but with an adjustment in the orders' time of arrival to the warehouse. In this new simulation, SCNR scenario 2, the orders made to warehouse S2 at the lunch hour are offset one hour earlier to potentially mitigate the delays obtained in the first test.

It is clear from Table 3.12 that this change is advantageous, reducing the average delays at warehouse P1 to 16 minutes, and the percentage of delays above 20 minutes by 7%. Hence, instead of delivering the orders later to the customers of warehouse S2, customers should be convinced to use an earlier ordering window at lunch time.

Table 3.12: Comparison of warehouse P1 expedition activity operational indicators from the baseline scenario and SCNR scenarios 1 and 2.

Warehouse	Baseline scenario	SCNR Scenario 1	SCNR Scenario 2
Average delay time (min)	12	18.9	15.6
% of travels delayed between 10 to 20 minutes	17%	16%	19%
% of travels delayed more than 20 minutes	3%	17%	10%

3.7. Conclusions and Future Work

This paper focuses on the supply chain network redesign (SCNR) of pharmaceutical wholesalers. Due to the increased pressure of the market, wholesalers are rethinking their strategies in order to minimize their costs, without jeopardizing the current customer service levels.

A hybrid optimization-simulation methodology is proposed to determine the best network design for established wholesalers. In the first phase, from a strategic perspective, the main decisions related to the SCNR are optimized using a MIP model. In this model, the number, location, function and capacity of the warehouses, as well as customer allocation, are optimized in order to minimize the total costs. In the second phase, the operation of the SC is simulated using a discrete event simulation model to analyze the new network design from an operational perspective. Using historical data from the operation, the warehouses' activities are simulated to understand the stress that the changes proposed may cause on them, and subsequently on the level of customer service. To conclude, a what-if analysis is conducted to determine the robustness of the solutions obtained.

The computational results presented for a case study validated the models and exposed the benefits of integrating optimization and simulation in one approach for SCNR. The indicators obtained from both models complement each other and make it possible to evaluate a solution from strategic, tactical and operational perspectives. As demonstrated in the case study, the optimization model can provide insights on the best network redesign that together with the simulation model helps understand the practicability of that implementation. In this case, the analysis of the delays caused by the redesign are one of the most important indicators for the company, because they may impact the service level. Therefore, the results of the simulation help to understand the adjustments that can be made for the solution to work. The results demonstrated that maintaining the company's five warehouses and yet changing one to cross-docking may guarantee savings up to 4.4%. Moreover, if an earlier ordering window at lunch time is imposed, the redesign will not cause delays beyond acceptable limits.

For businesses in which the details are critical and very time sensitive, this approach of integrating optimization and simulation leads to accurate and robust solutions. Although this methodology requires a high number of parameters to feed both models, most of them are obtained from historical data that a company can easily access.

Future work should focus on the incorporation of product allocation as one of the optimization decisions, making this methodology more suitable for wholesalers of other industries that face different operational conditions. The optimization and simulation can also be extended to further integrate the delivery routes. This extension would help to evaluate in more detail the customers' reallocation decisions. Moreover, an interaction between the simulation and the optimization model, to refine the optimization parameters according to the simulator solution and guide the search, would probably improve the final results.

Acknowledgements This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project «POCI-01-0145-FEDER-006961», and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The first author was also supported by Grant SFRH/BD/102013/2014 from FCT.

Bibliography

Badri, H., Bashiri, M., and Hejazi, T. H. (2013). Integrated strategic and tactical planning in a supply chain network design with a heuristic solution method. *Computers & Operations Research*, 40(4):1143–1154.

Baghalian, A., Rezapour, S., and Farahani, R. Z. (2013). Robust supply chain network design with service level against disruptions and demand uncertainties: A real-life case. *European Journal of Operational Research*, 227(1):199–215.

Bartolacci, M. R., LeBlanc, L. J., Kayikci, Y., and Grossman, T. A. (2012). Optimization modeling for logistics: Options and implementations. *Journal of Business Logistics*, 33(2):118–127.

Brown, J. E. and Sturrock, D. (2009). Identifying cost reduction and performance improvement opportunities through simulation. *Proceedings of the 2009 Winter Simulation Conference*, pp. 2145–2153.

Buil, R., Piera, M. a., and Laserna, T. (2010). Operational and strategic supply model redesign for an optical chain company using digital simulation. *Simulation*, 87(8):668–679.

Diabat, A., Richard, J.-P., and Codrington, C. W. (2013). A lagrangian relaxation approach to simultaneous strategic and tactical planning in supply chain design. *Annals of Operations Research*, 203(1):55–80.

Diabat, A. and Theodorou, E. (2015). A location–inventory supply chain problem: Reformulation and piecewise linearization. *Computers & Industrial Engineering*, 90:381–389.

Drexl, M. and Schneider, M. (2014). A survey of variants and extensions of the location-routing problem. *European Journal of Operational Research*, 241:283–308.

Eskigun, E., Uzsoy, R., Preckel, P. V., Beaujon, G., Krishnan, S., and Tew, J. D. (2005). Outbound supply chain network design with mode selection, lead times and capacitated vehicle distribution centers. *European Journal of Operational Research*, 165(1):182–206.

Figueira, G. and Almada-Lobo, B. (2014). Hybrid simulation–optimization methods: A taxonomy and discussion. *Simulation Modelling Practice and Theory*, 46:118–134.

Guerrero, W., Prodhon, C., Velasco, N., and Amaya, C. (2013). Hybrid heuristic for the inventory location-routing problem with deterministic demand. *International Journal of Production Economics*, 146(1):359–370.

Guimarães, L., Klabjan, D., and Almada-Lobo, B. (2012). Annual production budget in the beverage industry. *Engineering Applications of Artificial Intelligence*, 25(2):229–241.

Gzara, F., Nematollahi, E., and Dasci, A. (2014). Linear location-inventory models for service parts logistics network design. *Computers & Industrial Engineering*, 69:53–63.

Hübner, A. H., Kuhn, H., and Sternbeck, M. G. (2013). Demand and supply chain planning in grocery retail: an operations planning framework. *International Journal of Retail & Distribution Management*, 41(7):512–530.

Izadi, A. and Kimiagari, A. (2014). Distribution network design under demand uncertainty using genetic algorithm and Monte Carlo simulation approach : a case study in pharmaceutical industry. *Journal of Industrial Engineering International*, 10(1).

Javid, A. A. and Azad, N. (2010). Incorporating location, routing and inventory decisions in supply chain network design. *Transportation Research Part E: Logistics and Transportation Review*, 46(5):582–597.

Sabri, E. H. and Beamon, B. M. (2000). A multi-objective approach to simultaneous strategic and operational planning in supply chain design. *Omega*, 28:581–598.

Sargent, R. G. (2014). Verifying and Validating Simulation Models. *Proceedings of the 2014 Winter Simulation Conference*, pp. 118–131.

Shen, Z. M. and Qi, L. (2007). Incorporating inventory and routing costs in strategic location models. *European Journal of Operational Research*, 179:372–389.

Silver, E. A., Naseraldin, H., and Bischak, D. P. (2009). Determining the reorder point and order-up-to-level in a periodic review system so as to achieve a desired fill rate and a desired average time between replenishments. *Journal of the Operational Research Society*, 60(9):1244–1253.

Sousa, R. T., Liu, S., Papageorgiou, L. G., and Shah, N. (2011). Global supply chain planning for pharmaceuticals. *Chemical Engineering Research and Design*, 89(11):2396–2409.

Stadtler, H. and Kilger, C. (2008). *Supply Chain Management and Advanced Planning*. Springer, Berlin Heidelberg.

Zinn, W., Levy, M., and Bowersox, D. J. (1989). Measuring the effect of inventory centralization decentralization on aggregate safety stock: The "square root law" revisited. *Journal of Business Logistics*, 10(1):1–14.

CHAPTER 4

Grocery Retail Distribution with Loading Constraints

Loading Constraints for a Multi-Compartment Vehicle Routing Problem

Manuel Ostermeier [*] Sara Martins[†] · Pedro Amorim[†] · Alexander Hübner [*]

Submitted to *OR Spectrum*, 2017

Abstract Multi-compartment vehicles (MCVs) are able to circumvent the fact that products from different segments cannot be transported together. Using flexible MCVs allows tailoring the compartment size and position for each different tour. However, compartments still need to be accessible for loading at the distribution center and orders accessible for unloading at customers. This paper addresses such loading and unloading challenges raised in the distribution planning when using MCVs with flexible compartments. Literature on multi-compartment vehicle routing problems (MCVRPs) with flexible compartments is scarce. However, practice shows that a growing number of retailers use these technically advanced vehicles in their distribution process. Besides classical routing decisions, the configuration of each vehicle for each tour becomes an essential aspect of tour planning when using MCVs. This requires definition of the temperature mix, compartment sizes, and the combination of different orders and therefore customers on the vehicles. Routing and loading layout planning are therefore interdependent for MCVs. Our work addresses the problem of obtaining feasible MCV loading with a cost-optimal routing. Hence, we extend the MCVRP to take into account loading constraints. We present a specialized packing problem to define how the compartments should be built, taking into account these loading constraints. This model is used in a branch-and-cut (B&C) and in a large neighborhood search (LNS) framework to check the loading feasibility of the tours. Numerical studies show that the adapted LNS framework reaches the optimal solution for small instances and can be applied efficiently to larger problems. Additionally, further tests on larger instances helped to derive general rules regarding the influence of loading constraints, which were validated with a case study analysis with a European retailer.

Keywords Loading constraints · Multi-temperature logistics · Vehicle routing · Large Neighborhood Search · Branch-and-Cut

[*] Catholic University Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany

[†] INESC TEC and Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

4.1. Introduction

This paper addresses loading and unloading issues raised in the distribution planning when using multi-compartment vehicles (MCVs) with flexible compartments. The main applications of MCVs are waste collection (e.g., Henke et al. (2015)), fuel distribution (e.g., Derigs et al. (2011)) and food distribution (e.g., Hübner and Ostermeier (2016)). The focus of this paper is on retail grocery distribution, where efficient logistics are essential due to narrow margins, requirements for higher product availability and extensive regulations (e.g., product traceability). Retailers need to fulfill temperature requirements when transporting grocery products (e.g., frozen, fresh, ambient). The usual way in the past was to deliver each product segment, i.e. each group of products with a particular temperature requirement, separately on single-compartment vehicles (SCVs). However, in the last 10 years, a growing number of retailers have been using the relatively new technology of MCVs (Klingler et al., 2016). The MCVs are technically able to split their loading area flexibly into different compartments, each adjusted to a particular temperature that fulfills specific product requirements. The number and size of compartments of each MCV are not predefined. It is possible to built up between one and five different compartments. This can be adjusted for each tour separately without any loss of capacity. Additionally, the position of the compartments can be chosen freely on the vehicle according to the orders assigned. These features of an MCV make it possible to perform joint deliveries of different segments within the same vehicle.

However, these conditions yields additional challenges. Stores usually order products from different segments at the same time. Retailers organize the distribution center (DC) by temperature zone. To consolidate a customer order with different segments, a truck needs to pick up the orders separately for each segment from the temperature-specific DC areas. The vehicle loading is carried out from the rear of the vehicle. This is different than for the waste or glass collection where the products are loaded from the top. Figure 4.1 illustrates a possible loading layout of an exemplary MCV tour with four different product segments. Retailers use standardized transportation units. The loading area is divided by two horizontal compartment walls (lengthwise from rear door to front) and three vertical walls (between segments 1 and 2, 2 and 3 and 2 and 4). When a delivery with orders from multiple compartments is performed, these compartments have to be easily accessed to unload all orders of the customer. Due to the rear-loading and unloading, the access to a compartment may be blocked by other compartments. As given in Figure 4.1, parts of compartment 1 and therefore orders within it, can only be accessed if the corresponding parts of compartments $2,3$ and 4 have been, at least, partially emptied before. Neglecting the (un)loading operations during the routing planning can otherwise result in cases where different orders can only be unloaded by rearranging parts of other loaded orders.

The arising problem can be classified as a multi-compartment vehicle routing problem (MCVRP) with **L**oading **C**onstraints (MCVRP_LC). It combines VRP with MCVs and a loading problem. More specifically, it defines routing, assigns orders to compartments, defines compartments sizes, and specifies how the compartments and the orders should be arranged within the vehicle to obtain feasible and cost minimal tours. Since most of the literature on MCVs focus on waste and petrol applications, where each compartment

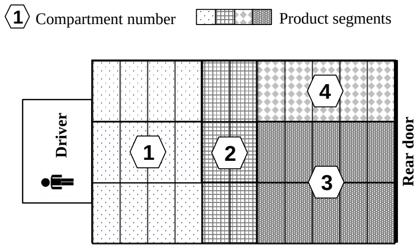


Figure 4.1: Exemplary loading layout of one MCV with four product segments

can be accessed individually (e.g., from top or by separate filler), the loading problem of MCVs has so far not been studied in literature. This work has multiple contributions. First of all, we identify the loading constraints with rear-loading. Two solution approaches are developed for the MCVRP_LC: an exact approach (branch-and-cut (B&C) algorithm) and a heuristic (extension of a large neighborhood search (LNS) framework). Both approaches are enriched with the the development of a multi-compartment packing problem (MCPP) that defines how the compartments should be built and the orders loaded into the truck. Finally, we identify the impact of loading constraints in MCVRPs by means of numerical studies with real and simulated retail data, allowing to generalize the findings.

The remainder of this paper is organized as follows. We detail the loading and distribution problem in Section 4.2 and discuss related literature in Section 4.3. The mathematical formulation for the MCVRP_LC is given in Section 4.4, together with the MCPP model. The proposed solution approaches are presented in Section 4.5. Numerical experiments evaluating the influence of loading constraints are presented in Section 4.6. Finally, section 4.7 summarizes our findings and discusses future research.

4.2. Loading Problem of MCVs

Transportation with MCVs needs to be distinguished from distribution with standard vehicles as different loading/unloading processes need to be considered. In this section, we first detail the distribution processes and related costs with MCVs, an then we define the actual loading issues.

As product segments are stored in the DC by temperature zone, an MCV needs to approach different segment-specific shipping gates to pick up pallets from multiple segments (see left-hand side of Figure 4.2 where four shipping gates are visited). Therefore, the loading costs depend on the number of product segments and hence on the number of compartments. Each additional segment leads to additional loading costs as there are set-up costs related to traveling and loading. On the other hand, customer orders from different segments can be combined, which reduces the number of stops at the stores (see right-hand side of Figure 4.2 where each customer receives more than one segment). This feature leads to different unloading costs, related with stopping costs.

Our work focus on the grocery distribution, where the transportation units have identical size and cannot be stacked. For the SCVs, the loading of the orders just needs to follow the opposite sequence of the route so that the orders of the first customer approached are

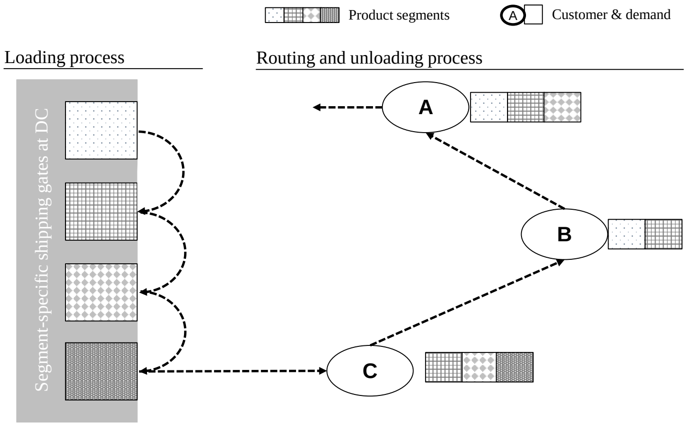

Product segments
A
Customer & demand
Loading process
Routing and unloading process
Segment-specific shipping gates at DC
A
B
C

Figure 4.2: Loading, routing and unloading processes with an MCV

at the rear of the truck. However, when using MCVs, the loading of customer orders from each segment within each compartment needs to be made in a sequence such that they can be easily accessed without any rearrangement. The rearrangement of orders during a tour (e.g., moving orders of store B during a stop at store A) is not desired due to strict legal regulations with regard to maintaining a constant temperature during the transport, short time-windows for unloading, limited space at retail stores, and significantly higher unloading costs. In the remainder, we describe the critical characteristics of (un)loading procedures that have to be considered to achieve a feasible loading of an MCV.

(1) Requirements for the joint loading of orders of one segment. All orders of one segment that will be transported on the same vehicle have to be loaded at the same time (i.e. no other segment can be loaded in between), as otherwise it would be necessary to approach a shipping gate for the same segment several times. This is not practical as it would lead to an unmanageable planning situation, increasing the waiting time for loading at the DC, and the set-up costs for loading. Figure 4.3 illustrates four different loading options for an MCV with four compartments (vehicle (V) 1-4). The loading area is separated into three aisles, and a compartment can be within and across an aisle. Example V1 shows a feasible solution where two segments share the middle aisle, but are separated by a compartment wall. Example V2 shows that two compartments for the same segment can be set up one after another in different aisles of an MCV. This is feasible as the compartments can be loaded after each other at the same shipping gate. Example V3 shows the possibility of splitting one compartment for the same segment crosswise, i.e. into different aisles. Example V4 illustrates that one compartment cannot be split in the same aisle. A loading as given in V4 would require the repeated approach of a shipping gate. When a vehicle leaves the DC the layout of the loading area is fixed and cannot be changed.

Requirements for a feasible (2) customer and (3) segment sequence. The unloading problem of an MCV is essentially driven by the question of whether all orders that need to be unloaded at a customer can be accessed without any obstacles in the form of orders

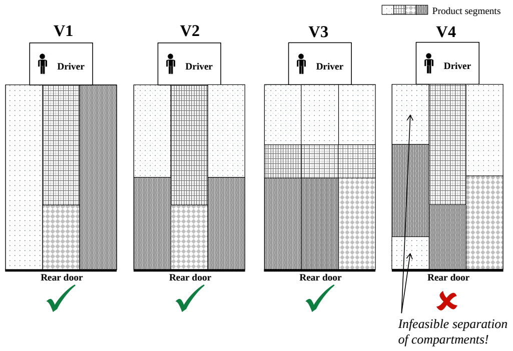


Figure 4.3: Illustrative vehicle loading layout with multiple compartments

from other customers. Figure 4.4 illustrates possible sequencing issues for a tour example. There are three customers (A, B and C, represented by circles) with different order sizes (represented by the number in the rectangle) from a specific segment. Let us assume that the customer sequence with the minimum transportation costs is *B*-*C*-*A*. Three loading layouts are presented for the given customer orders. The first two layouts (V1 and V2) are infeasible because the orders of customer A would block the unloading orders of customer C who needs to be served before A. The only layout that would allow feasible unloading is the third (V3). However, this would require approaching the shipping gate of one segment twice. As a consequence, no feasible loading can be achieved for the tour *B*-*C*-*A* that is solely focused on minimizing the transportation costs.

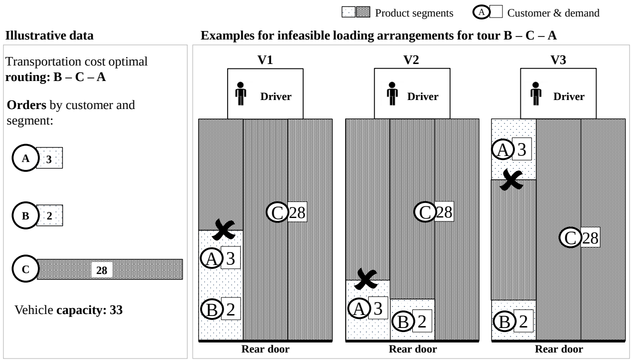

Product segments
A Customer & demand
Illustrative data
Examples for infeasible loading arrangements for tour B – C – A
Transportation cost optimal routing: B – C – A
Orders by customer and segment:
A 3
B 2
C 28
Vehicle capacity: 33
V1
V2
V3
Driver
A 3
B 2
C 28
Rear door

Figure 4.4: Illustrative example for infeasible loading and unloading

Note that backhauling of empty transportation units (pallets or roll-cages) is not considered in this study, as we found that this is not an issue for our partners in practice. In operational practice the driver decides if and how many units he returns from each store. As empty pallets or roll-cages can be stacked and therefore require only a fraction of space of full ones, their return is not a bottleneck in distribution.

4.3. Related Literature

The MCVRP_LC is a variant of the CVRP. Based on the problem characteristics there are two relevant streams within VRP literature to ve covered. On the one hand, our work clearly extends the MCVRP as the use of MCVs is a central aspect in our routing decisions. On the other hand, the problem we are facing concerning the loading of MCVs is related to the VRP with loading constraints.

4.3.1 Literature related to MCVRP

While there is a wide range of publications dealing with the CVRP and its various extensions (see general overviews at Golden et al. (2008) and Toth and Vigo (2014)), there is only a small body of MCVRP literature. We are considering MCVs with flexible compartments, the number and size of which can be adjusted. The MCVRP extends an CVRP by adding constraints for the building and loading of different compartments on each vehicle. Besides the delivery of groceries, there are several areas of application for MCVs, such as fuel distribution or the collection of waste. However, most publications in these areas consider the use of fixed compartments (e.g., Muyldermans and Pang (2010) for waste collection and Avella et al. (2004) for fuel distribution), instead of vehicles where compartments can be adjusted flexibly.

Derigs et al. (2011) are the first to present a comprehensive problem formulation for MCVRP with flexible compartments. They formulate a general model for the application of their MCVRP for food and fuel distribution. They employ a solver suite combining different heuristic approaches to solve the MCVRP. This involves construction heuristics (e.g., Sweep-Algorithm) as well as a combination of search and metaheuristics. In addition, Derigs et al. (2011) identify an LNS with the operator combination Shaw removal and regret insertion as particularly efficient. Henke et al. (2015) consider the use of MCVs for the collection of glass waste. They allow flexible compartments but restrict their sizes to predefined values. They also allow the use of more product categories than compartments. For their MCVRP, they apply a variable neighborhood search. Further, Henke et al. (2017) present a branch-and-cut algorithm for their MCVRP in glass collection. Recently, Koch et al. (2016) propose a genetic algorithm for a similar problem setting. Hübner and Ostermeier (2016) present an MCVRP that considers process costs implied by the use of MCVs in grocery distribution. They show that costs for loading/unloading have to be taken into account to achieve a more realistic evaluation of total costs due to additional operations that are required when using MCVs. The LNS proposed by Hübner and Ostermeier (2016) is used in this work to solve the MCVRP part of the MCVRP_LC.

4.3.2 Literature related to VRP with loading constraints

Different loading constraints have to be faced in many routing problems. VRP with loading constraints are a combination of CVRP and packing problems. A general overview of packing problems and container loading problems can be found in Wäscher et al. (2007) and Bortfeldt and Wäscher (2013). A review on the combination of routing and loading problems was introduced by Iori and Martello (2010) and more recently by Pollaris et al. (2014). Côté et al. (2016) examine the value of integrating the loading decisions instead of solving routing and loading sequentially.

Grocery retailers use standardized transportation units for the transport of goods from the DCs to stores, which cannot be stacked. Two types are available: pallets and roll-cages. As a consequence, the loading of MCVs in grocery distribution considers the assignment of two-dimensional objects of identical size (i.e., pallets or roll cages) to identical containers (i.e., loading area of identical vehicles). We therefore focus on literature concerning two-dimensional loading CVRP (2L-CVRP) that also considers sequence-based loading. An 2L-CVRP is first addressed by Iori et al. (2007). They propose an B&C algorithm for routing and an integrated branch-and-bound framework to check loading feasibility. Similarly, Côté et al. (2014) developed an B&C algorithm that uses a one-dimensional contiguous bin packing problem to identify and eliminate non-feasible loading. We use a similar approach recorring to the MCPP to identify and cut infeasible loadings. Further, both Iori et al. (2007) and Côté et al. (2014) solve the routing and packing sequentially, which can be seen as a starting point for the problem studied in this paper.

Besides these exact methods for the examination of loading constraints, most publications focus on heuristic solution approaches for 2L-CVRP. Attanasio et al. (2007) solve a simplified integer linear program within a cutting plane framework for their variant of the 2L-CVRP, extended by multiple time windows for each day. Gendreau et al. (2008) present a tabu search (TS) for the 2L-CVRP. In their approach, they combine heuristics, lower bounds and a truncated branch-and-bound procedure to solve the loading problem. A guided TS is used by Zachariadis et al. (2009), who later propose a metaheuristic that uses a local search procedure for diversification (Zachariadis et al., 2013). Fuellerer et al. (2009) address an 2L-CVRP where a rotation of items of 90 degrees is allowed. They apply ant colony optimization to solve the respective problem. Finally, Pollaris et al. (2016) develop a special case of 2L-CVRP, an CVRP with sequence-based pallet loading and axle weight constraints. They present a mixed integer linear program for their problem formulation and compare it to a model without axle weight restrictions.

Summary. In our problem we combine an MCVRP and an 2L-CVRP formulation and we therefore leverage both streams in literature. More precisely, Iori et al. (2007) and Côté et al. (2014) inspired our approach to sequentially tackle the loading and routing problem. Further, we use an LNS framework as this showed very good results in various applications of MCVRPs (Derigs et al., 2011; Hübner and Ostermeier, 2016).

4.4. MCVRP with Loading Constraints

The formulation proposed for the MCVRP_LC incorporates the loading constraints into an MCVRP model in the form of sub-tour elimination constraints. This is achieved by defining a set Ω that records tours with infeasible loading. In this section, we first present the overall problem formulation, followed by the MCPP formulation that defines how compartments should be built inside the vehicles, respecting loading requirements. If such is not possible (infeasibility), the tour is added to Ω.

The MCVRP_LC can be defined as follows. For an undirected, weighted graph $G = (N, E)$ a set of vertices $N = \{0, 1, \ldots, n\}$ is given, representing the location of the DC ($\{0\}$) and the locations of n customers. The connection between different locations is represented by the set of edges $E = \{(i, j) : i, j \in N\}$ and each edge has an associated traveling cost t_{ij}. It is assumed that all traveling costs satisfy the triangle inequality and each tour starts and ends at the DC. Let V be the set of vehicles available for transportation at the DC. The number of vehicles available is assumed to be sufficiently large to fulfill all customer demand and consists of identical vehicles with capacity Q. The loading area of each vehicle can be split into a limited number of compartments $c \in C$ with $C = \{1, \ldots, c\}$. The number of compartments used on each vehicle is indicated by $k \in K$, with $K = \{1, ..., c\}$. As in Hübner and Ostermeier (2016) we also consider an MCVRP with loading/unloading costs. The loading cost l_k depends on the number of segment specific-shipping gates approached and is related with the number of compartments used. The unloading cost u is incurred every time a customer is visited.

Let $P = \{1, \ldots, p\}$ be the set of orders to be delivered on a given day and q_m the quantity of each order $m \in P$ ($q_m > 0$). Let $S = \{1, \ldots, s\}$ be the set of segments available for distribution. The subsets D_i and W_h are used to represent all orders from a customer $i \in N$ and all orders from a segment $h \in S$, respectively. Finally, assuming Ω as the set of tours with infeasible loading, let $E_\omega \subseteq E$ be the subset composed by the sequenced edges that are traversed in the infeasible loading tour $\omega \in \Omega$. Moreover, let $P_\omega \subseteq P$ be the subset of orders considered in the infeasible loading tour $\omega \in \Omega$.

For the formulation of the MCVRP_LC we used the following variables:

$$\theta_{mvc} \begin{cases} = 1, & \text{if order } m \text{ is assigned to compartment } c \text{ on vehicle } v \\ = 0, & \text{otherwise} \end{cases}$$

$$b_{ijv} \begin{cases} = 1, & \text{if vehicle } v \text{ transverses the edge } (i, j) \\ = 0, & \text{otherwise} \end{cases}$$

$$\vartheta_{vc} \begin{cases} = 1, & \text{if compartment } c \text{ on vehicle } v \text{ is active} \\ = 0, & \text{otherwise} \end{cases}$$

$$r_{vk} \begin{cases} = 1, & \text{if vehicle } v \text{ has } k \text{ active compartments} \\ = 0, & \text{otherwise} \end{cases}$$

f_v representing the number of customer stops for vehicle v

$\delta_{iv} = d,\ d \in \{1, ..., n\}$ representing position d of customer i on vehicle v

The mathematical model can be formulated as follows:

$$min! \sum_{v \in V} \Big[\sum_{k \in K} l_k \cdot r_{vk} + \sum_{i \in N} \sum_{j \in N} t_{ij} \cdot b_{ijv} + u \cdot f_v \Big] \tag{4.1}$$

subject to:

$$\sum_{j \in N \setminus \{0\}} b_{0jv} \leq 1 \qquad v \in V \tag{4.2}$$

$$\sum_{i\in N} b_{igv} = \sum_{j\in N} b_{gjv} \qquad v\in V,\ g\in N \tag{4.3}$$

$$\delta_{iv} - \delta_{jv} + (n+1)\cdot b_{ijv} \le n \qquad v\in V,\ i\in N,\ j\in N\backslash\{0\} \tag{4.4}$$

$$\delta_{0v} = 1 \qquad v\in V \tag{4.5}$$

$$\sum_{m\in P}\sum_{c\in C} q_m\cdot\theta_{mvc} \le Q \qquad v\in V \tag{4.6}$$

$$\sum_{v\in V}\sum_{c\in C}\theta_{mvc} = 1 \qquad m\in P \tag{4.7}$$

$$\sum_{m\in D_j}\sum_{c\in C}\theta_{mvc} \le p\cdot\sum_{i\in N} b_{ijv} \qquad v\in V,\ j\in N\backslash\{0\} \tag{4.8}$$

$$\sum_{m\in W_h}\theta_{mvc} \le p\cdot(1-\sum_{r\in W_z}\theta_{rvc}) \qquad v\in V,\ c\in C,\ h,z\in S : h\neq z \tag{4.9}$$

$$\sum_{m\in P}\theta_{mvc} \le p\cdot a_{vc} \qquad c\in C,\ v\in V \tag{4.10}$$

$$\sum_{c\in C}\vartheta_{vc} = \sum_{k\in K} k\cdot r_{vk} \qquad v\in V \tag{4.11}$$

$$\sum_{k\in K} r_{vk} \le 1 \qquad v\in V \tag{4.12}$$

$$f_v \ge \sum_{i\in N}\sum_{j\in N\backslash\{0\}} b_{ijv} \qquad v\in V \tag{4.13}$$

$$\sum_{m\in P_\omega}\sum_{c\in C}\theta_{mvc} + \sum_{(i,j)\in E_\omega} b_{ijv} \le |E_\omega| + |P_\omega| - 1 \qquad \omega\in\Omega,\ v\in V \tag{4.14}$$

$$\theta_{mvc}, b_{ijv}, \vartheta_{vc}, r_{vk} \in \{0,1\}; f_v, \delta_{iv} \ge 0 \qquad m\in P, v\in V,\ c\in C,\ i,j\in N,\ k\in K \tag{4.15}$$

The objective function (4.1) minimizes the total cost of all tours and consists of three parts: (i) loading, (ii) transportation and (iii) unloading costs. Constraints (4.2) and (4.3) represent routing constraints, guaranteeing that each vehicle can depart at most once from the depot, and that every vehicle that visits a location also leaves it. Constraints (4.4) and (4.5) are used to eliminate sub-tours. The former determines the position of each location within the tour, while the latter ensures that the depot is the first in the sequence of locations approached. The overall quantity of the orders assigned to a vehicle cannot exceed the vehicle capacity as stated by constraint (4.6). According to constraint (4.7), each order can only be assigned to one compartment and one vehicle. Additionally, an order can only be assigned to a vehicle if the associated customer is visited on the tour (constraint (4.8)). Constraint (4.9) ensures that orders from different segments are not assigned to the same compartment. To model the use of compartments, constraint (4.10) ensures that a compartment is set to active if an order is assigned to it. Constraints (4.11) and (4.12) control the value for the number of compartments used. For this, r_{vk} has to be activated for the correct number of compartments and at the same time it can only be activated once for each vehicle. In addition, constraint (4.13) ensures that the number of stops is chosen accordingly for each vehicle. Constraint (4.14) ensures that the infeasible tours in terms of loading are eliminated. This is accomplish by avoiding tours with the same customer sequence and order mix as the infeasible tours. Lastly, the variables domains are defined by constraint (4.15).

4.4.1 Multi-Compartment Packing Problem

The MCPP can be classified as an *Identical Item Packing Problem* (IIPP) (see Wäscher et al. (2007)), since it involves the assignment of identical items to exactly one large object while maximizing the number of orders loaded. In our context, this requires the loading of all orders assigned to a given tour, while respecting customer and segment sequence. Hence, if not all orders can be loaded, the corresponding tour is declared infeasible.

The MCPP can be defined as follows. Let $A = \{1,...,a\}$ be the set of orders assigned to be loaded on the vehicle in analysis. The loading area and capacity of an MCV can be divided into a certain set of items (see the left-hand side of Figure 4.5). To position each single item into the truck, the loading area is divided into $|X| \cdot |Y|$ possible spots. Here, $x \in X$ represents the aisles available on the loading area and $y \in Y$ represents the different positions within an aisle. Each order $m \in A$ has a fixed quantity q_m that expresses the number of loading items. The orientation of the items is considered to be fixed in order to be able to move them in the required manner (e.g., pallets can be accessed using pallet trucks). Let, now, $N = \{1,...,n\}$ be the set of customers that have to be visited by the vehicle and $S = \{1,...,s\}$ the set of segments to be loaded. Each customer can have more than one order to be delivered, but only one per segment. Once again, all orders from a customer $i \in N$ compose the subset $D_i \subseteq A$ and all orders from a segment $h \in S$ compose the subset $W_h \subseteq A$. To take into account the customer sequence, a set $G_i \subseteq N$ is defined indicating the predecessors of customer $i \in N$, i.e. all customers that need to be visited earlier than customer i in the tour. Lastly, M is used as a sufficiently high number to regulate the setting of our decision variables.

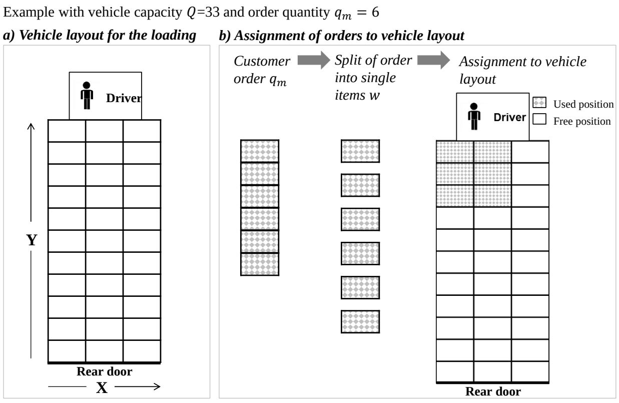

Example with vehicle capacity Q=33 and order quantity q_m = 6
a) Vehicle layout for the loading
b) Assignment of orders to vehicle layout
Driver
Y
Rear door
X
Customer order q_m
Split of order into single items w
Assignment to vehicle layout
Driver
Used position
Free position
Rear door

Figure 4.5: Loading layout for the Multi-Compartment Packing Problem (MCPP) (illustrative example)

We introduce the following decision variables for the formulation of the MCPP:

$$\lambda_{mxy} \begin{cases} = 1, & \text{if an item of order } m \in A \text{ has been assigned to position } (x,y) : x \in X, y \in Y \\ = 0, & \text{otherwise} \end{cases}$$

$$\beta_{hz} \begin{cases} = 1, & \text{if segment } h \in S \text{ is loaded after } z \in S \text{ in the segment sequence} \\ = 0, & \text{otherwise} \end{cases}$$

Since we want to define a feasible loading, all orders assigned to the vehicle have to be loaded. Therefore, the MCPP does not require an objective function, but has to respect the following constraints.

$$\sum_{x\in X}\sum_{y\in Y}\lambda_{mxy} = q_m \qquad m \in A \tag{4.16}$$

$$\sum_{m\in A}\lambda_{mxy} \leq 1 \qquad x \in X,\ y \in Y \tag{4.17}$$

$$\beta_{hz} + \beta_{zh} = 1 \qquad h,z \in S\ :\ h > z \tag{4.18}$$

$$\beta_{hz} + \beta_{zf} \leq 1 + \beta_{hf} \qquad h,z,f \in S\ :\ s \neq z \neq r \tag{4.19}$$

$$y \cdot \lambda_{mxy} \leq y' \cdot \lambda_{rxy'} + M(2 - \lambda_{rxy'} - \beta_{hz}) \qquad h,z \in S\ :s \neq z,\ m \in W_h,\ r \in W_z,\ x \in X,\ y,y' \in Y \tag{4.20}$$

$$y \cdot \lambda_{mxy} \leq y' \cdot \lambda_{rxy'} + M(1 - \lambda_{rxy'}) \qquad i \in N,\ j \in G_i,\ m \in D_j,\ r \in D_i,\ x \in X,\ y,y' \in Y \tag{4.21}$$

$$\lambda_{mxy}, \beta_{hz} \in \{0,1\} \qquad m \in A,\ x \in X,\ y \in Y,\ h,z \in S \tag{4.22}$$

Constraint (4.16) ensures that each order assigned to the vehicle has all items loaded. This is ensured by setting the sum of all assigned items of an order m equal to the order quantity q_m. Only one item can be assigned to each position on the truck, which is guaranteed by constraint (4.17). The requirements for customer and segment sequence described previously are represented by constraints (4.18) to (4.21). First, constraints (4.18) and (4.19) define the segment sequence, i.e. in which sequence the segments shipping gates should be visited. Second, constraint (4.20) ensures that the defined sequence of segments is not violated during the loading and that the joint loading of each segment is guaranteed. This is accomplished by ensuring for each aisle of the vehicle that the orders of segment loaded later are closer to the rear door (see Figure 4.5). Finally, the customer sequence requirement is ensured by constraint (4.21). It ensures that all orders in the same aisle adhere to the customer sequence of the tour, i.e., the correct unloading sequence is respected. The variables domains are defined by constraint (4.22).

4.5. Solution Approaches

The MCVRP_LC is an extension of the CVRP and therefore NP-hard (Toth and Vigo (2014)). In Section 4.5.1 we first present an B&C algorithm that iteratively solves the routing and loading problem by generating cuts for infeasible tours at each integer node. However, this exact method is only capable of solving small instances. Therefore, we further extend an LNS framework to solve practically relevant problem sizes, by incorporating a repair mechanism for infeasible loading tours. In addition, exact and heuristic solution approaches for the MCVRP (without loading constraints) are used to obtain bounds for the assessment of our extension to the MCVRP_LC. Table 4.1 provides an overview of the solution approaches applied.

Table 4.1: Overview of solution approaches that have been applied and developed

Problem	Exact approach	Heuristics	Purpose
MCVRP	B&B	LNS[1]	Used as benchmark
MCVRP_LC	B&C	LNS_LC[2]	Used to benchmark the heuristic quality and evaluate the impact of loading constraints

[1] Based on Hübner and Ostermeier (2016)
[2] Extends the LNS approach with a repair mechanism

4.5.1 Exact approach for the MCVRP_LC

An B&C algorithm is proposed to solve the MCVRP_LC (see Figure 4.6). This exact approach is used to get a benchmark for the heuristics in terms of solution quality. In the first step, an MCVRP is solved without loading constraints by using the branch-and-bound (B&B) framework provided by CPLEX, i.e. constraint (4.14) of the MCVRP_LC is not considered. As soon as a feasible solution for the MCVRP, and therefore an integer node is reached, the check of loading feasibility is carried out in Step (a). The MCPP model proposed in Section 4.4.1 is used to check the feasibility of each tour. The sequence of customers and the orders assigned to each vehicle are passed to the MCPP model, which returns if a feasible loading is possible or not.

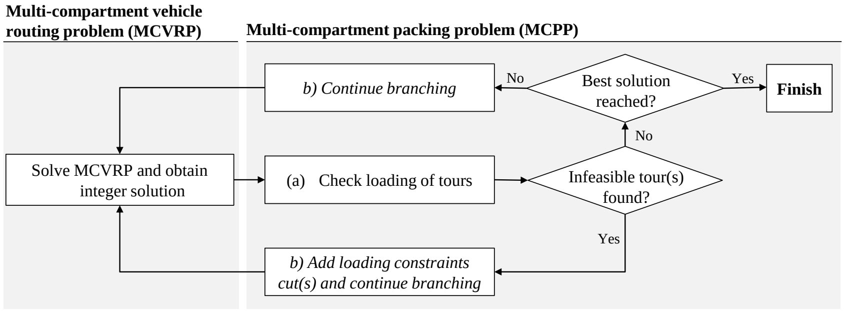
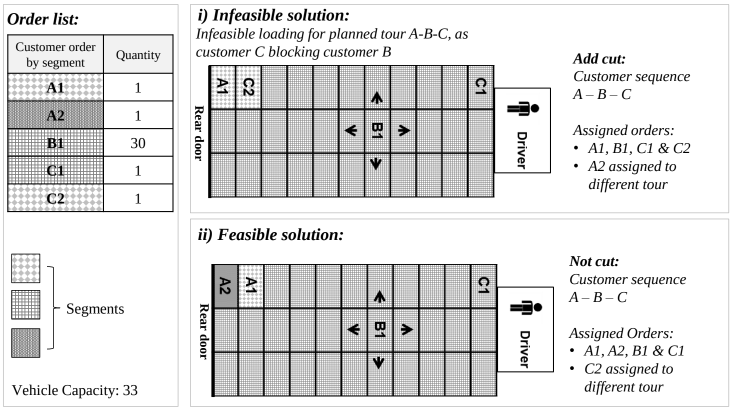

Multi-compartment vehicle routing problem (MCVRP)
Multi-compartment packing problem (MCPP)
b) Continue branching
No
Best solution reached?
Yes
Finish
No
Solve MCVRP and obtain integer solution
(a) Check loading of tours
Infeasible tour(s) found?
Yes
b) Add loading constraints cut(s) and continue branching

Figure 4.6: Scheme of exact approach for MCVRP_LC with branch-and-cut

According to the outcome of Step (a), there are two possible cases for Step (b). If the MCPP returns a feasible solution, no loading constraints are violated and the algorithm continues branching. If no feasible solutions can be found for a tour, the tour is declared infeasible and therefore added to the set Ω. This means that a cut is added to the MCVRP_LC model. The cuts take into account the special characteristics of the problem setting. In contrast to classical VRP with loading constraints, in MCVRPs the orders of one customer can all be put on the same as well as on different tours. Figure 4.7 illustrates this aspect, where the left-hand side denotes the data for the simplified example and the right-hand side the possible loading solutions.

The example shows the solution for a tour with three customers (A, B and C) and the corresponding orders denoted by the customer and the order number for different segments (e.g., A1). The best MCVRP solution found for this data set includes a tour with a customer sequence A-B-C (see part i) of Figure 4.7). However, this is an infeasible tour in terms of



Order list:

Customer order by segment	Quantity
A1	1
A2	1
B1	30
C1	1
C2	1

Figure 4.7: Example for added cuts: customers and orders define cut

loading, due to the fact that order C2 is blocking order B1 and customer B is supplied before customer C. If we add a cut only considering the customer sequence, all possible order combinations for the sequence A-B-C would also be cut. Nevertheless, the same customer sequence can result in a feasible tour if other orders from the same customers are assigned to the vehicle as displayed in the part ii) of Figure 4.7. Just by exchanging the segments assigned, a feasible loading is possible for the tour. The order mix is therefore decisive to make sure that a tour that is identified as having infeasible loading is not further regarded in the solution process. This means that not only the customer sequence needs to be considered, but also which orders of each customer are assigned. After the cuts are performed, the branching procedure continues until the best solution is reached (and proved optimal).

4.5.2 Heuristic approach for the MCVRP_LC

The complexity of the MCVRP is significantly increased compared to classical CVRP variants due to the consideration of multiple compartments. Therefore, in order to solve large instances relevant for practice, heuristic solution approaches have been proposed for the MCVRP. To tackle our problem, the LNS framework is extended to incorporate loading constraints. This involves the inclusion of a repair mechanism that checks the tours loading feasibility and repairs it, if no feasible loading is possible. The resulting solution approach is labeled as LNS_LC, and the overall structure is given in Figure 4.8.

The LNS_LC operates in two phases. First, the LNS is applied solely for the MCVRP, which is denoted as standard LNS (sLNS), and provides a solution for the MCVRP. Afterwards, the final solution is checked for infeasibilities. If the sLNS solution has tours with infeasible loading, the search continues solving the MCVRP_LC. At this stage, a repair mechanism is incorporated into the previous LNS framework, now denoted as repair LNS (rLNS). Both LNS variants are similar, however the rLNS calls the repair mechanism each time a new best candidate solution is found. If an infeasible loading is detected in one tour, it is repaired and the resulted solution checked again by the acceptance criteria. The rLNS only saves as best solutions, solutions respecting the loading constraints. The main features

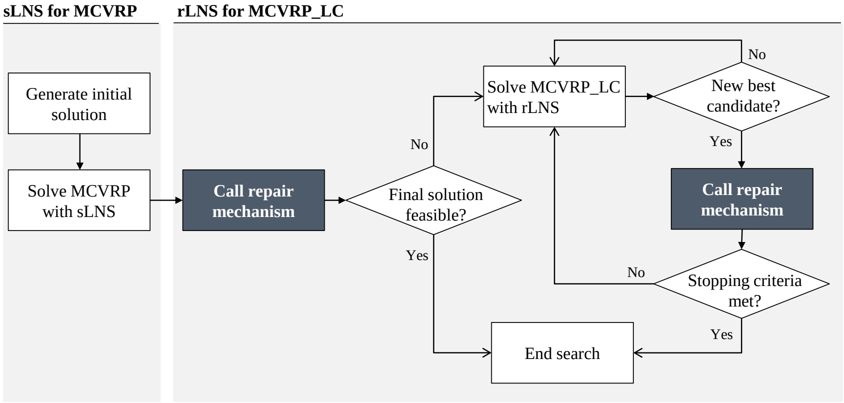


Figure 4.8: Scheme of LNS_LC approach

of the sLNS framework used are described in Section 4.5.2.1, followed by the description of the repair mechanism used in the rLNS (Section 4.5.2.2).

4.5.2.1 Standard LNS (sLNS) for the MCVRP

As previously mentioned, we used the LNS framework proposed by Hübner and Ostermeier (2016). Algorithm 1 outlines the sLNS framework.

Algorithm 1 standard Large Neighborhood Search (sLNS)

```
generate a solution $S$
set regret parameter $k$
set allowed deviation for Record-To-Record Travel $RRT$
set $S_{best} := S$
repeat
    $S' := S$
    remove $r$ orders from $S'$ using Shaw Removal
    reinsert removed orders into $S'$ using Regret-k Insertion
    increase number of unsuccessful runs
    if $S'$ better than $S_{best}$ then
        $S_{best} = S'$
        $S = S'$
        reset number of unsuccessful runs to zero
    else if $S'$ complies with acceptance criteria $RRT$ then
        $S := S'$
    end if
    if number of unsuccessful runs = reset border then
        remove $r$ orders from $S$ using Shaw Removal
        reinsert removed orders into $S$ using Regret-k Insertion
    end if
until number of unsuccessful runs < limit
return $S_{best}$
```

The sLNS starts by generating a solution using the parallel Savings Algorithm presented by Clarke and Wright (1964). It operates by building single customer tours for each given order and then combines tours according to a calculated saving in traveling distance.

From the initial solution we move on by applying the Shaw removal (Shaw, 1997) as destroy operator to remove orders of the incumbent solution. In the following step, the orders removed are reinserted in tours using regret-k insertion as repair operator. It was shown in previous works on MCVRPs that these operators are able to efficiently generate good solution (Derigs et al., 2011). If the new solution fulfills the acceptance criterion defined by Record-To-Record Travel (RRT), it replaces the incumbent solution. The procedure is repeated until a termination criterion is met. In the subsequent paragraphs a description of the operators used is presented.

Shaw Removal. The Shaw removal was introduced by Shaw (1997) and resembles a remove operator based on a similarity measure for pairs of customers. The removal approach is based on a defined similarity measure R_{ml} between any two orders m and l (either from the same customer or different customers) with $m, l \in P$ and a randomized selection. In total, a defined number of r orders has to be removed from all orders P. In the following, we divide the set of all orders P into removed orders (P^-) and assigned orders (P^+), such that $P^+, P^- \subseteq P$, $P^+ \cup P^- = P$ and $P^+ \cap P^- = \emptyset$. The first order m removed is chosen randomly from all orders P. Further orders are gradually removed according a defined procedure based on the calculated similarity measure R_{ml}. After one order m has been selected randomly in Step 1, the similarity of orders is calculated in Step 2. The similarity index for all orders is calculated using the measure R_{ml} that is defined in Equation (4.23). It expresses the similarity of two orders m and l, with $m \in P^-$ and $l \in P^+$. The smaller the calculated value of R_{ml} becomes, the more similar the orders compared are. Derigs et al. (2011) propose a modified similarity measure to include MCVRP particularities. The measure is given below and combines traveling costs, order quantity and order segment.

$$R_{ml} := \phi \cdot \frac{cost_{ml}}{cost_{max}} + \psi \cdot \frac{|quantity(m) - quantity(l)|}{quantity_{max}} + \varphi \cdot segment_{ml} \tag{4.23}$$

The parameters $cost_{max}$ and $quantity_{max}$ indicate the maximum traveling cost and order quantity across orders that are considered. Further, $segment_{ml}$ is set to 1 if two orders are from different segments and cannot be allocated to the same compartment, and 0 otherwise. For the removal process, this similarity measure is combined with a randomization step in order not to choose the most similar order, but a random order amongst the similar orders. In Step 2 the orders are ranked according to similarity in descending order. Then the next order for removal needs to be selected in Step 3. This selection process is based on the random number $z \in [0, 1)$ and a parameter α. Thus, for the selection of a new order, it is not the order with the highest similarity that is chosen but one that can be found z^α percent down the similarity ranking. The parameter α can be seen as a parameter for diversification. If α is chosen to be 1, the similarity is not taken into account and the choice of orders is completely random. The smaller the value of α is set, the more decisive the similarity calculated. After r orders are removed, the algorithm continues with the reinsertion.

Regret-k Insertion. As repair operator, the regret-k insertion is used following the approach by Ropke and Pisinger (2006). It is a more foresighted heuristic compared to a greedy approach as it considers information on postponed insertion of orders. More precisely, it takes into account the k best possibilities for insertion instead of only the best one. Based on the cost difference between these k options and the best option, a regret value is calculated in the form of Equation (4.24). Let $\Delta_{v_u}^m$ denote the cost of adding a selected order $m \in P$ at its best position on the u-th best tour. The cost calculation is based on the objective function (4.1).

$$regret_k^m := \sum_{u=2}^{k} (\Delta_{v_u}^m - \Delta_{v_1}^m) \tag{4.24}$$

Then the order m with the highest regret value is chosen to be inserted into the tour that is best suited to it. This means the insertion does not just consider the actual state but also uses the regret criterion to evaluate possible future costs. After the order s with the highest regret is inserted into P^+, the regret value is recalculated for each order m on the removal list P^-, as the insertion options might have changed. The insertion procedure is thus iterated until all orders removed, $m \in P^-$, are allocated to a new position in the tour planning, and hence $P^- = \oslash$ and $P^+ = P$.

Record-To-Record Travel. The search operators of the sLNS are embedded into a RRT framework as presented in Dueck (1993). A solution obtained is only accepted during the search if it improves the search, or if it lies within a defined deviation from the solution that is best so far. Furthermore, additional parameters are integrated for regulation of the application runtime. If the number of unsuccessful iterations reaches a reset border, the next iteration of sLNS is run with a high value (e.g., one half of the order list) for the number of orders removed (r) in order to create a completely different neighborhood, thus aiming to avoid local minima. A limit is set for the maximum number of succeeding fruitless iterations to terminate the algorithm.

4.5.2.2 LNS with repair mechanism (rLNS) for the MCVRP_LC

The incorporation of a repair mechanism into the LNS framework is one central aspect of our extension to consider loading constraints for MCVRP. The LNS with repair mechanism (rLNS) framework starts by calling the repair mechanism to check the loading feasibility of the sLNS solution obtained. If the sLNS solution provides feasible tours, the search ends. Otherwise, the repair mechanism changes the tours in order to make them feasible and the rLNS continues solving the MCVRP_LC (see Figure 4.8).

The rLNS framework is similar with the sLNS (see algorithm 1). However, each time a new best candidate solution is found (line 10 of algorithm 1), instead of saving it as best solution, it calls the repair mechanism to check its feasibility. If the solution is feasible it is saved as best and the search continues. Otherwise, the solution is repaired and checked again by the acceptance criteria. Note that by repairing the solution we are changing some decisions that could lead to an increase of the solution cost. An overview of the repair mechanism is presented in Figure 4.9.

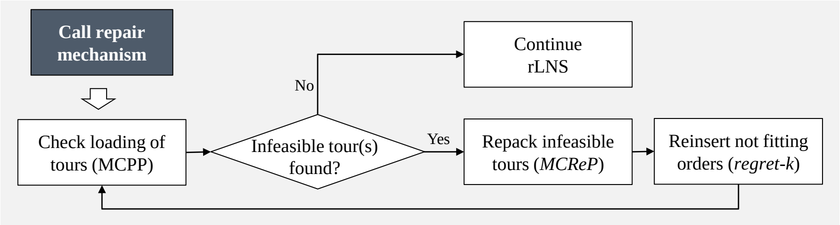

Call repair mechanism
Continue rLNS
No
Yes
Check loading of tours (MCPP)
Infeasible tour(s) found?
Repack infeasible tours (MCReP)
Reinsert not fitting orders (regret-k)

Figure 4.9: Procedure of repair mechanism within rLNS

First, the repair mechanism applies the MCPP model presented in Section 4.4.1 to each tour of the solution to assess its loading feasibility. If an infeasible tour is found, a repacking phase is applied to remove the orders that do not completely fit on the truck without violating the loading constraints. This procedure is carried out for all tours in the current solution until every tour has been checked for feasibility. By the end of the repacking phase, every order that caused non-feasible loading for a corresponding tour will have been added to the list of orders removed. Based on this list, a regret-k insertion step is applied to find the best option for the removed orders. A tour with removed orders is set tabu for the corresponding orders in the reinsertion phase. This is done to prevent the circling of the algorithm. As the reinsertion causes changes to the structure of affected tours, the feasibility has to be assessed again. Therefore, the MCPP is applied to all modified tours within the reinsertion process and if a tour is found to be infeasible, the repacking is applied again as described above. This process is repeated, until a feasible loading has been found for all tours within the solution. Note that a new tour is created if it is the lowest total cost option or if all available tours are set tabu for an order.

The repacking phase uses a model denoted as *Multi Compartment Repacking Problem* (MCReP). The MCReP is based on two modifications of the MCPP. Instead of aiming to find a feasible loading for a tour, it tries to maximize the number of items loaded (from the orders assigned to the tour) while satisfying the loading constraints. Therefore, the MCPP model is transformed in the MCReP model by adding an objective function and changing constraint (4.16). The MCReP can be read as follows:

$$\max! \sum_{m\in P}\sum_{x\in X}\sum_{y\in Y}\lambda_{mxy} \tag{4.25}$$

$$\text{subject to: } (4.17)-(4.21) \tag{4.26}$$

$$\sum_{x\in X}\sum_{y\in Y}\lambda_{mxy} \leq q_m \qquad m\in P \tag{4.27}$$

$$\lambda_{mxy},\beta_{hz} \in \{0,1\} \qquad m\in P,\ x\in X,\ y\in Y,\ h,z\in S \tag{4.28}$$

The new constraint (4.27) allows the assignment of a number of items lower than the quantity of each order. Together with the objective function (4.25) of the MCPP, this results in the assignment of the maximum number of items without violating the loading constraints. We therefore obtain a solution with feasible loading for the considered tour, but also a solution that does not assign all orders completely as the loading of all orders is not feasible. The order(s) that could not be assigned to a tour due to loading constraints are removed from the corresponding tour and reinserted into another tour within the LNS.

4.6. Numerical Experiments

In this section we present numerical experiments to evaluate our solution approaches and grasp the relevance of including loading constraints in the MCVRP. As a basis for our experiments we present the structure of our data generation in Section 4.6.1. An overview of all tests in Sections 4.6.2 to 4.6.4 is given in Table 4.2.

Table 4.2: Overview of numerical tests

Section	Model	Solution approaches	Data applied	Test purpose
4.6.2.1	MCVRP and MCVRP_LC	MIP Solver, B&C	Small randomly generated data	Impact of loading constraints for small instances
4.6.2.2	MCVRP_LC	B&C, LNS_LC, sLNS_ExPost	Small randomly generated data	Effectivity of heuristic approach
4.6.3	MCVRP_LC	LNS_LC, sLNS_ExPost	Large randomly generated data	Efficiency of heuristic and impact of loading constraints for larger instances
4.6.4	MCVRP_LC	LNS_LC,	Case study sLNS_ExPost	Influence of loading constraints in practice

In Section 4.6.2 we use small size instances and compare the results of the MCVRP_LC with the results of the MCVRP (without loading constraints) obtained by the corresponding exact approaches. Furthermore, we investigate the differences of the heuristic to the B&C. This aims to identify the efficiency of the heuristic in terms of solution quality and runtime. Additionally, the results of the LNS with the repair mechanism (LNS_LC) are compared to the sLNS solution with an ex-post evaluation of loading feasibilities. This is done by applying the repair mechanism only to the final solution of the sLNS framework. We denote this approach as sLNS_ExPost. In Section 4.6.3 we consider tests with larger instances. Here we test the LNS_LC as well as the sLNS_ExPost to obtain further insights regarding runtime and the influence of different data structures on loading constraints. In Section 4.6.4 we apply the methodology to a real life case example of a major European grocery retailer.

4.6.1 Overview of test instances

In order to perform the numerical experiments to analyze the influence of loading constraints in the MCVRP, we use randomly generated problem instances. We leverage our data generation on the data structures we obtained from our case study (see Section 4.6.4). The resulting instances can be found on http://www1.ku.de/wwf/pw/forschung/. For both small and large size problems we define the number of customers, the number of orders per customers, i.e. the number of different segments each customer orders, and the order quantity.

Small instances. As the problem becomes very hard to solve exactly with an increasing number of customers and/or orders, we generate small instances. Please note that for MCVRP the main drivers for complexity are the number of orders and related segments, not the number of customers as in classical VRP formulations. The first type of instance comprises ten customers, and each customer orders one out of four available segments. In the second type of instances there are four customers and four segments, but each customer orders all segments. This results in a total of 16 orders. For these small problems the distance matrix is randomly generated with distances between any two customers set between 7 and 80 kilometers. The distances between customers and the DC are set between 90 and 120 kilometers to create a clustered group of customers.

Large instances. The larger problems are created for 25, 50, 75 and 100 customers with two or four segments. As not all customers have demand for all segments, it is also possible that not all available segments are ordered. Therefore, different sets of instances will be generated changing the number of segments ordered. The distance design in Solomon (1983) for VRP with 100 customers is used for all tests and has been adjusted to our problem instances. We use the instances related to clustered customers (C-type) and uniformly distributed customers (R-type). In food distribution, the mix of orders between customers and segments is very heterogeneous as the analysis of our case study has shown (see Section 4.6.4). Using this information, the quantity of each order will depend on the segment in question and will vary between a given range. The ranges used are presented in Table 4.3. For instance, segment 4 could comprise the ambient assortment with larger order sizes and segment 1 could be seen as deep frozen where orders are usually small. The segments associated with each customer are chosen randomly depending on the number of orders a customer submits. For instance, in the case where each customer orders three out of four segments, three segments are selected randomly to generate the orders. The quantity of orders is defined according to the segments in question.

Table 4.3: Range of order quantity per segment (in TU)

Segment	Minimum Quantity	Maximum Quantity
1	1	5
2	1	10
3	5	20
4	10	25

Homogeneous MCVs with a total capacity of 33 transportation units (TU) were used for all tests. The loading/unloading costs used in the experiments have been derived from our case study. The loading costs depend on the number of segments to be loaded at the distribution center and are presented in Table 4.4. Unloading costs accrue with every customer stop and are set to 2.20 currency units (CU). The transportation costs are based on the travel distances between any two locations.

Table 4.4: Applied costs for loading MCV

# Segments to be loaded	1	2	3	4
Loading (CU/shipping gate)	2.70	5.57	8.27	10.97

The computational results were obtained on a 3.60 GHz PC with 16 GB memory. The implementation of the B&C algorithm has been done in C++ and the LNS algorithm in Java. The setting of the heuristic specific parameters is informed by previous research on MCVRPs and the application of an LNS as they performed very well on different instances related to our case (Hübner and Ostermeier, 2016; Derigs et al., 2011). Accordingly, for the Shaw Removal the weights for the calculation of the similarity measure R_{ml} were set to $\phi = \varphi = 0.4$ and $\psi = 0.2$ and $\alpha = 4$. This choice of weights is based on the higher influence of distance costs and product segments compared to order size. The number of items removed r are chosen randomly using a uniform distribution dependent on the respective problem size as given in Table 4.5. The regret parameter has been set at $k = 2$. Furthermore, the termination limit is 1,000 and the limit for a solution reset is 500 for all LNS applications. The maximum deviation D allowed for the RRT equals 0.4% for all tests.

Table 4.5: Number of removed orders (r)

# Customer	Minimum	Maximum
10	2	5
25	2	10
50	5	15
75	5	20
100	5	30

4.6.2 Comparison between exact and heuristic approaches for small instances

We present two comparisons concerning small problems. First, in Section 4.6.2.1 we compare the exact solutions of an MCVRP, i.e. without loading constraints, with the exact solutions for the MCVRP_LC provided by our B&C algorithm. With this experiment we aim to understand the cost implications of adding loading constraints. Second, in order to validate the efficiency of the heuristics, the results obtained with it are compared to the results obtained with the B&C in Section 4.6.2.2.

These comparisons comprise the results of ten different instances for each scenario tested. The B&C approach has been terminated after two hours as it becomes very hard to prove optimality. However, the search ends with an average solution gap of 0.01%. For the heuristic approaches 50 repetitions are applied to create a sample for results comparison. As the solution approaches provide stable results with an average variation coefficient (standard deviation/mean) below 1%, the comparisons are made based on the average result achieved for each instance.

4.6.2.1 Comparison of exact approaches for MCVRP_LC vs. MCVRP

This experiment uses the B&B and B&C algorithms to solve the MCVRP and MCVRP_LC, respectively. Instances from two scenarios are tested. Scenario 1 correspondents to instances with 4 customers and 4 out of 4 segments ordered, resulting in 16 orders to distribute. Scenario 2 relates to 10 customers and 1 out of 4 segments orders, with a total of 10 orders. The cost deviation between the MCVRP_LC solution and the MCVRP solution achieved for each scenario are presented in Figure 4.10.

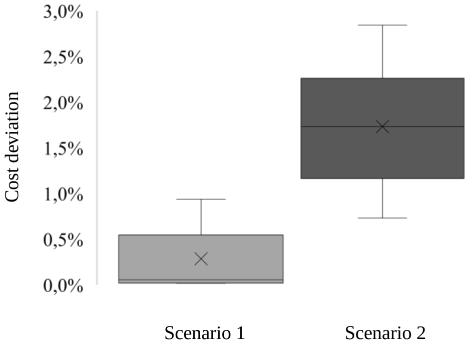


Figure 4.10: Comparison of the results for MCVRP_LC vs. MCVRP with exact approaches,*10 instances per scenario*

The results show that the cost deviation of considering loading constraints is around 0.0%-0.5% for Scenario 1 and between 1.2% and 2.3% for Scenario 2. This means that even for small instances, the loading constraints have an impact on the routing decisions, particularly for Scenario 2. Here, the impact on costs is higher than in Scenario 1, as ten different customers are considered with only one order each. Any move of an order is equal to the move of a customer to a different tour and has a high influence on the routing as only ten customers are available. However, tours with feasible loading can be reached with a small increase in the costs and thus would be preferred to a solution that would require rearrangement of the orders. In most cases the exchange of two orders or the reassignment of a single order is sufficient to obtain feasible tours.

4.6.2.2 Comparison of exact approaches vs. heuristics for MCVRP_LC

This second numerical experiment aims to analyze the efficiency of the heuristic approaches. It compares the solutions reached by the LNS_LC and sLNS_ExPost with the exact solutions achieved with the B&C for the MCVRP_LC. Figure 4.11 summarizes the cost deviations of both heuristic approaches to the B&C for the same two scenarios (with 4 and 10 customers) used previously.

The results indicate that the LNS_LC reaches most of the times the optimal solutions of the MCVRP_LC for both scenarios. The ex post approach also achieves solutions with an average deviation close to 0% for Scenario 1, but for Scenario 2 the deviations increase. For Scenario 2, where only 1 out of 4 segments are ordered, the sLNS_ExPost provides solutions with an average deviation of 8%.

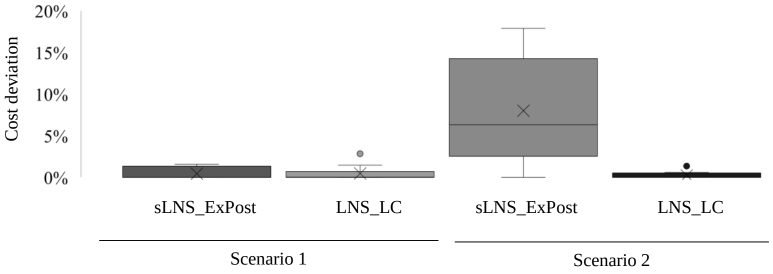


Figure 4.11: Comparison of heuristics vs. exact approach for MCVRP_LC, *10 instances per scenario*

4.6.3 Analysis of loading constraints for larger instances

Subsequent to our tests with small instances, we analyze the impact of loading constraints for larger instances. Firstly, we analyze the runtime development of the heuristic for different problem sizes. Secondly, we present tests for different problem settings to examine the influence of loading constraints. For the tests with larger instances, the results of 20 instances for each of the given scenarios are compared. As in the exact approach analysis, 50 repetitions of the heuristic approaches are applied to each instance to create a sample for results comparison. As the solution approaches also provide stable results for the larger instances with an average variation coefficient below 1%, the comparisons are made based on the average result achieved for each instance.

4.6.3.1 Runtime development

The problem complexity and requirements for the heuristic approach increase quickly for larger problems. We therefore decided to focus on the runtime development for the LNS_LC and sLNS_ExPost for 25, 50, 75 and 100 customers. A C-type distance matrix is considered for the performed tests. In all tests, four segments are available. However, as in the tests with the exact approach, we consider two different scenarios for all order structures, where either all four segments are ordered by each customer, or only one segment per customer. The average, maximum and standard deviation of the computational time (runtime) required to solve each instance of each scenario are summarized in Table 4.6. Note that as each scenario is tested for 20 instances, which are run for 50 repetitions, the results presented are calculated considering the 1000 ($20 \cdot 50$) individual runs.

Our tests show that the number of segments ordered seem to be an important driver for the computational time required to solve an MCVRP_LC instance. We see that for the same number of customers, changing the segments ordered from the 4 out of 4 case to the 1 out of 4 case leads to an increase of the LNS_LC runtime. Both the average and maximum runtime are increased, as well as the standard deviation. An increase of the number of customers does not seem to have an influence on the runtime development. This can be attributed to the number of infeasible solutions reached during the search. The increase in the number of customers could change the routing decisions, but it is the mix of orders that indicates the number of compartments to be built in the MCVs and that could influence

Table 4.6: Runtime development for increasing problem sizes (minutes), *20 instances*

#Customers	#Segments ordered	sLNS_ExPost Runtime Average	Maximum	σ	LNS_LC Runtime Average	Maximum	σ
25	1 out of 4	0.10	0.94	0.10	4.54	27.30	2.88
	4 out of 4	0.06	0.64	0.07	0.41	16.80	0.78
50	1 out of 4	0.12	1.12	0.12	6.16	61.62	7.61
	4 out of 4	0.21	2.18	0.17	1.21	29.65	1.77
75	1 out of 4	0.17	1.03	0.12	16.01	158.50	22.90
	4 out of 4	0.34	1.56	0.18	1.27	37.40	2.25
100	1 out of 4	0.54	0.80	0.09	7.12	66.90	8.60
	4 out of 4	0.17	1.36	0.17	1.43	11.80	1.44

the loading feasibility. This aspect will be analyzed in more detail in Section 4.6.3.2. The bottleneck for the LNS_LC is the repair mechanism and therefore the call of the MCPP and MCReP models. As more infeasible solutions are found during the search, the more repacking phases are required and the runtime of the LNS_LC increases. This is noticeable in the tests with 75 customers and 1 out of 4 segments ordered, which have higher runtimes than the remaining scenarios. As the runtime is highly driven by infeasibility tests, even though fewer customers are considered, three out of the 20 instances tested had a high number of infeasible tour combinations. The analysis of the sLNS_ExPost results shows that it is able to solve all instance scenarios in less than 3 minutes, as it only calls the repair mechanism once.

4.6.3.2 Scenario analysis for solution quality

In this subsection we present the analysis of different larger data settings. This is done to examine the influence of loading constraints with different data settings. This also enables us to gain further insight into the runtime development of the suggested approach. The analysis is based on the random generated instances with 25 and 100 customers. We would like to note that even in the case of 25 customers we consider up to 100 orders. In contrast to other VRP, the order number is an essential parameter for defining the problem size. Three different scenarios are considered for the number of segments ordered: the 1 out of 4 and 4 out of 4 scenarios, previously used, and an additional scenario with only two segments available, and both ordered by all customers (2 out of 2). We emphasize the case with four segments as usually retailers are confronted with this situation in practice. In each scenario we further examine changes in customer distribution (matrix type) and loading costs (costs). The matrix types comprise clustered (C) and uniformly distributed (R) customers. Loading costs are used as given in Table 4.4 (NC) as well as costs reduced by 50% (LC) for additional scenarios. This change of loading costs can lead to changes in the compartment setting of MCVs (see Hübner and Ostermeier (2016)), i.e. the lower the loading costs the more compartments could be activated. Besides the number of customers, all settings and scenarios were equal for the cases with both 25 and 100 customers.

In this numerical experiment, we run the sLNS for the MCVRP (without loading constraints) and if the final solution reached has infeasible tours, we run the sLNS_ExPost and the LNS_LC. For each scenario we indicate the number of instances (out of the 20) that revealed infeasible tours for the best final solution achieved by the sLNS for the MCVRP (see column 4 in Tables 4.7 and 4.8). Further, the cost of the final solution provided by the LNS_LC and sLNS_ExPost is compared. The average and minimum deviations reached for the group of instances of each scenario are displayed in Table 4.7 for 25 customers and in Table 4.8 for 100 customers. Note that the LNS_LC never provides a worse solution than the sLNS_ExPost, as it continues the search from the sLNS_ExPost solution (see Figure 4.8). The average and maximum runtime required to solve each instance with the LNS_LC are also provided in Tables 4.7 and 4.8. The runtime of the sLNS_ExPost is always below 1 minute for all scenarios with 25 customers, and below 4 minutes for the scenarios with 100 customers.

Table 4.7: Test overview for instances with 25 customers, *20 instances*

# Segments	Matrix	Loading	**Infeasible**	**Cost deviation**		**LNS_LC Runtime** (minutes)	
ordered	type[1]	costs[2]	**instances**[3]	Average	Minimum	Average	Maximum
Scenario 1	C	NC	25%	-1.32%	-2.81%	1.17	16.80
	C	LC	25%	-1.04%	-2.65%	0.92	10.46
(4 out of 4)	R	NC	10%	-1.11%	-1.65%	0.62	2.88
Scenario 2	C	NC	35%	-3.62%	-4.30%	3.62	27.30
	C	LC	15%	-4.79%	-8.44%	1.75	6.99
(1 out of 4)	R	NC	20%	-2.23%	-2.66%	2.44	16.64
Scenario 3	C	NC	0%	-	-	-	-
	C	LC	0%	-	-	-	-
(2 out of 2)	R	NC	0%	-	-	-	-

[1] According to Solomon (1983); C: clustered customers; R: uniformly distributed
[2] Loading cost: NC - normal costs; LC - lower costs, reduced by 50% compared to NC
[3] Share of sLNS solutions that contain at least one infeasible tour (i.e., violate loading constraints)

Table 4.8: Test overview for instances with 100 customers, *20 instances*

# Segments	Matrix	Loading	**Infeasible**	**Cost deviation**		**LNS_LC Runtime** (minutes)	
ordered	type[1]	costs[2]	**instances**[3]	Average	Minimum	Average	Maximum
Scenario 1	C	NC	50%	-0.28%	-3.38%	1.27	11.80
	C	LC	35%	-0.27%	-2.94%	1.88	27.77
(4 out of 4)	R	NC	55%	-0.36%	-3.29%	2.60	37.33
Scenario 2	C	NC	100%	-1.98%	-9.71%	7.12	66.90
	C	LC	80%	-1.74%	-10.57%	8.70	101.80
(1 out of 4)	R	NC	95%	-1.44%	-7.51%	11.15	85.15
Scenario 3	C	NC	25%	-0.22%	-3.23%	0.57	9.81
	C	LC	30%	-0.22%	-3.73%	0.47	9.88
(2 out of 2)	R	NC	30%	-0.22%	-2.80%	0.76	16.20

[1] According to Solomon (1983); C: clustered customers; R: uniformly distributed
[2] Loading cost: NC - normal costs; LC - lower costs, reduced by 50% compared to NC
[3] Share of sLNS solutions that contain at least one infeasible tour (i.e., violate loading constraints)

Scenario 1. In the first scenario up to 25% of instances reveal an infeasible solution in the tests with 25 customers (see Table 4.7). For 100 customers (see Table 4.8), this value amounts to over 50%. Our solution approach shows that the LNS_LC can improve the ex-post application of loading constraints from 1% up to 3%, for the case of 25 customers. When we increase the number of customers, the average cost deviation between the two approaches is almost null. However, the best solutions found by the LNS_LC can reach an improvement of up to 3.4%. The runtime of the LNS_LC for the 25 and 100 customer case shows a moderate increase considering that the number of orders increases from 100 (25 customer case) to 400 (100 customer case).

Scenario 2. The order structure with 1 out of 4 segments ordered shows an higher impact on the loading feasibility of tours. For 25 customers, the results indicate up to 35% infeasible best solutions, and 100% for 100 customers. This increase in infeasible solutions impacts both runtime and solution deviation from the sLNS_ExPost. As more feasibility checks are required, the runtime increases, with some instances requiring more than one hour to be solved. On the other hand, this increase in runtime is accompanied by a solution improvement. For the 25 customer case the average cost deviation increases to 3% (maximum 8%). The average cost deviation for the 100 customer case is still small (below 2%), but reaches improvements of up to 10.57%.

Scenario 3. It is not surprising that in our last scenario with two segments the influence of loading issues is smaller compared to the other scenarios. The reduction of available segments also reduces the planning complexity for the compartment setup on a vehicle and therefore also for loading at the DC. In the tests with 25 customers these results in feasible solutions for all tested instances. However, when analyzing the 100 customers case, 30% of the instances were found to be infeasible, with cost deviations reaching 3% improvement by using the LNS_LC. This shows that we cannot neglect loading constraints even when only two segments are distributed.

Summary. In our tests we compare the performance of the LNS_LC and sLNS_ExPost for different larger data settings. This enabled the analysis of the solution quality, runtime efficiency and influence of loading constraints. As shown in different scenarios, the LNS_LC performs well in terms of solution quality. The solution approach is able to improve the solutions of the sLNS_ExPost, reaching maximum improvements between 3% and 10%. However, the improvements achieved by the LNS_LC comes with a greater computational effort. The average runtime of the LNS_LC can increase on average to 11 minutes (with maximums of more than one hour), whereas the sLNS_ExPost has a maximum runtime of 4 minutes. The main driver for these differences in solution quality and runtime is the frequency of infeasible solutions within the search. With an increasing number of infeasible solutions, the LNS_LC provides a higher solution improvement but also consumes more computational time. The scenarios considered show that the number of available segments together with the order structure of customers is the most influential driver for loading issues. Other factors such as the geographic positioning of customers and costs do not have an impact on runtime and solution quality. With a larger number of segments, the number of infeasible tours increases. If only two segments are available, fewer solutions will contain infeasible tours. However, the occurrence of non-feasible tours increases for four segments, specially if customers only order 1 out of 4 segments, with up to 100% of the problems having an infeasible best solution.

4.6.4 Case study

To conclude the numerical experiments, we applied our solution approach to a case study with a major European grocery retailer. For our analysis we use data from one specific distribution center and area. The data set received comprises one example week with 4 different product segments available for delivery. As before, the distribution is carried out by identical vehicles with a capacity of 33 TU. The longest distance from the DC to a customer is around 300 km. The considered DC supplies around 100 customers spread around an area of 54.000 km^2. The example week comprises seven delivery days and more than 2,000 customer orders. The complete information regarding our case is presented in Table 4.9.

Table 4.9: Case data for one example week

Day	# of orders	Ø order size	Std. dev. order size	Ø number of segments per customer
1	261	6	6	2.7
2	330	10	8	3.4
3	353	10	8	3.7
4	335	10	8	3.5
5	357	11	9	3.7
6	338	11	9	3.5
7	202	11	5	2.1

We solved the MCVRP_LC using the LNS_LC and sLNS_ExPost for each delivery day and thus seven different instances. The summary of our findings is displayed in Table 4.10.

Table 4.10: Summary of case study results

Day	**Cost deviation**		**LNS_LC Runtime (minutes)**	
	Average	Minimum	Average	Maximum
1	-0.77%	-4.46%	5.29	50.50
2	-0.15%	-0.94%	1.06	3.51
3	-0.31%	-2.78%	3.71	18.51
4	-0.86%	-4.09%	2.31	11.34
5	-0.33%	-1.70%	3.85	18.12
6	-0.52%	-1.85%	6.58	20.53
7	-	-	-	-

In all but the last day infeasible tours have been found within the search procedure of the sLNS for the MCVRP. This is due to the small number of segments per customer (on average only 2 out of 4 segments, see Table 4.9). On day 7 most of the customers only receive two of the segments. Hence, it is similar with a 2 out of 2 structure, as all customers receive the same two segments. Consequently, the results indicate that the consideration of loading constraints in the MCVRP has practical relevance.

The improvement ratio of the LNS_LC is similar with the results for scenario 1 in Table 4.8 (100 customers case). The average cost deviation between the solution approaches is below 1%, with the highest improvement reaching 4%. The average runtime shows a small increase but the maximum runtime is reduced for most of the days. The sLNS_ExPost provided solutions with a computational time lower than 1 minute. Since the LNS_LC logic uses the solution of the sLNS_ExPost as a starting point for the rLNS search (see Figure 4.8), retailers can use the sLNS_ExPost to generate tours with feasible loading and decide if they want to continue with the rLNS search, knowing that option requires a higher computational effort. Further, the sLNS solution cost for the MCVRP can be compared to the one of the sLNS_ExPost, i.e. after repairing the infeasibilities, and depending on the cost increase decide if the computational effort of using the LNS_LC is worthy. Most of the infeasible tours found could be repaired by exchanging one or two orders. However, it is important to note that even though the number of orders to be moved can be reduced, the impact of rearranging instead of changing the tour can be high. For instance, if the unloading of a truck is made in a street and not within a shipping gate with controlled temperature, performing the unloading and reloading of the blocking orders could lead to the degradation of the products.

4.7. Conclusion

The focus of this work was to define and examine the problem of loading constraints that occurs in grocery distribution with MCVs, developing a solution approach able to solve the arising MCVRP variant. We therefore presented a detailed problem description to highlight loading issues with flexible compartment sizes. We showed that even in the simplest cases problems can arise for the unloading of goods at the customer if the corresponding loading of a vehicle is not taken into account during the loading process. The loading of MCVs has never been studied in the literature, therefore we develop a packing problem called Multi-Compartment Packing Problem that defines how the vehicle should be loaded in order to respect the loading constraints, if such is possible.

We introduce an extension for MCVRP with flexible compartments, the MCVRP with Loading Constraints, for which we propose an B&C algorithm to exactly solve the problem. The solution approach makes use of the packing problem developed to solve the loading problem for MCVs and add cuts for infeasible tours. Moreover, we adapt the LNS framework propose by Hübner and Ostermeier (2016) to consider loading constraints. The packing problem introduced is used to check the loading feasibility of tours and a modified packing model is also introduced to repair the infeasible tours and thus guide the search for feasible solutions.

The numerical experiments indicate that (i) loading constraints matter even for small instances, (ii) feasible loading can often be achieved by only minor changes to the routing solution and therefore with limited additional costs, and that (iii) the number of infeasible solutions increases as the problem sizes increase, specially when heterogeneous mix of segments are ordered. The numerical experiments also show that there is a small deviation between the average results achieved with the LNS_LC and an ex post check of loading constraints and that the first requires a higher computational effort. Therefore, for complicated and slower problems the sLNS_ExPost approach can be used to generate good feasible solutions in a short running time. Nevertheless, if the obtained solution yields a high cost increase, the LNS_LC approach can be used to improve it.

We consider a variant of an MCVRP with flexible compartments, which is still a very novel research area. Literature is very limited as most publications deal with fixed compartment sizes. As a consequence, there are various possibilities for extending variants of the MCVRP with loading constraints. For a start, literature on MCVRP focuses on problems with homogeneous fleets. An extension of our approach to address a heterogeneous fleet and determining an optimal fleet mix would be a logical next step. Furthermore, the packing problem proposed could be adapted for other types of MCVs with less flexibility. Following classical VRP formulations, an MCVRP with loading constraints could also be extended to account for backhauls or pick-up and delivery problems. Additionally, there is a lack of literature concerning MCVRPs across multiple periods. The delivery with MCVs might impact delivery frequencies and therefore also delivery patterns (Holzapfel et al., 2016). Usually delivery patterns are defined by retailers for their stores. Using multiple compartments opens up new possibilities for defining the corresponding delivery patterns. Lastly, the scope of the research could be extended to include the impact on store operations or inventory costs (Gaur and Fisher, 2004; Hübner and Schaal, 2017).

Acknowledgements The research of the first author is founded by the German Ministry of Education and Research and the Hanns Seidel Foundation. The initialization of this international research cooperation was financially supported by the Bavarian Research Alliance. Further, this work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The second author was also supported by Grant SFRH/BD/102013/2014 from FCT.

Bibliography

Attanasio, A., Fuduli, A., Ghiani, G., and Triki, C. (2007). Integrated shipment dispatching and packing problems: a case study. *Journal of Mathematical Modelling and Algorithms*, 6(1):77–85.

Avella, P., Boccia, M., and Sforza, A. (2004). Solving a fuel delivery problem by heuristic and exact approaches. *European Journal of Operational Research*, 152(1):170–179.

Bortfeldt, A. and Wäscher, G. (2013). Constraints in container loading a state-of-the-art review. *European Journal of Operational Research*, 229(1):1–20.

Clarke, G. and Wright, J. W. (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research*, 12(4):568–581.

Côté, J.-F., Gendreau, M., and Potvin, J.-Y. (2014). An exact algorithm for the two-dimensional orthogonal packing problem with unloading constraints. *Operations Research*, 62(5):1126–1141.

Côté, J.-F., Guastaroba, G., and Speranza, M. G. (2016). The value of integrating loading and routing. *European Journal of Operational Research*.

Derigs, U., Gottlieb, J., Kalkoff, J., Piesche, M., Rothlauf, F., and Vogel, U. (2011). Vehicle routing with compartments: Applications, modelling and heuristics. *OR Spectrum*, 33:885–914.

Dueck, G. (1993). New optimization heuristics. *Journal of Computational Physics*, 104(1):86–92.

Fuellerer, G., Doerner, K. F., Hartl, R. F., and Iori, M. (2009). Ant colony optimization for the two-dimensional loading vehicle routing problem. *Computers & Operations Research*, 36(3):655–673.

Gaur, V. and Fisher, M. (2004). A periodic inventory routing problem at a supermarket chain. *Operations Research*, 52(6):813–822.

Gendreau, M., Iori, M., Laporte, G., and Martello, S. (2008). A tabu search heuristic for the vehicle routing problem with two-dimensional loading constraints. *Networks*, 51(1):4–18.

Golden, B. L., Raghavan, S., and Wasil, E. A. (2008). *The Vehicle Routing Problem: Latest Advances and New Challenges*. Operations Research/Computer Science Interfaces Series. Springer.

Henke, T., Speranza, M. G., and Wäscher, G. (2015). The multi-compartment vehicle routing problem with flexible compartment sizes. *European Journal of Operational Research*, 246(3):730–743.

Henke, T., Speranza, M. G., and Wäscher, G. (2017). A branch-and-cut algorithm for the multi-compartment vehicle routing problem with flexible compartment sizes. *Working Paper No. 04/2017, Otto-von-Guericke-University Magdeburg*.

Holzapfel, A., Hübner, A., Kuhn, H., and Sternbeck, M. (2016). Delivery pattern and transportation planning in grocery retailing. *European Journal of Operational Research*, 252:54–68.

Hübner, A. and Ostermeier, M. (2016). A multi-compartment vehicle routing problem with loading and unloading costs. *Working Paper*.

Hübner, A. and Schaal, K. (2017). Effect of replenishment and backroom on retail shelf-space planning. *Business Research*, page forthcoming.

Iori, M. and Martello, S. (2010). Routing problems with loading constraints. *TOP*, 18(1):4–27.

Iori, M., Salazar-González, J.-J., and Vigo, D. (2007). An exact approach for the vehicle routing problem with two-dimensional loading constraints. *Transportation Science*, 41(2):253–264.

Klingler, R., Hübner, A., and Kempcke, T. (2016). *End-to-end supply chain management in grocery retailing*. European Retail Institute.

Koch, H., Henke, T., and Wäscher, G. (2016). A genetic algorithm for the multi-compartment vehicle routing problem with flexible compartment sizes. *Working Paper No. 04/2016, Otto-von-Guericke-University Magdeburg*.

Muyldermans, L. and Pang, G. (2010). On the benefits of co-collection: Experiments with a multi-compartment vehicle routing algorithm. *European Journal of Operational Research*, 206(1):93–103.

Pollaris, H., Braekers, K., Caris, A., Janssens, G. K., and Limbourg, S. (2014). Vehicle routing problems with loading constraints: state-of-the-art and future directions. *OR Spectrum*, 37(2):297–330.

Pollaris, H., Braekers, K., Caris, A., Janssens, G. K., and Limbourg, S. (2016). Capacitated vehicle routing problem with sequence-based pallet loading and axle weight constraints. *EURO Journal on Transportation and Logistics*, 5(2):231–255.

Ropke, S. and Pisinger, D. (2006). An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. *Transportation Science*, 40(4):455–472.

Shaw, P. (1997). A new local search algorithm providing high quality solutions to vehicle routing problems. *APES Group, Dept of Computer Science, University of Strathclyde, Glasgow, Scotland, UK*.

Solomon, M. (1983). Vehicle routing and scheduling with time window constraints: Models and algorithms. *Technical report, College of Business Admin., Northeastern University, USA*.

Toth, P. and Vigo, D. (2014). *Vehicle Routing: Problems, Methods, and Applications, Second Edition*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics.

Wäscher, G., Haußner, H., and Schumann, H. (2007). An improved typology of cutting and packing problems. *European Journal of Operational Research*, 183(3):1109–1130.

Zachariadis, E. E., Tarantilis, C. D., and Kiranoudis, C. T. (2009). A guided tabu search for the vehicle routing problem with two-dimensional loading constraints. *European Journal of Operational Research*, 195(3):729 – 743.

Zachariadis, E. E., Tarantilis, C. D., and Kiranoudis, C. T. (2013). Integrated distribution and loading planning via a compact metaheuristic algorithm. *European Journal of Operational Research*, 228(1):56 – 71.

Chapter 5

Grocery Retail Distribution with Consistent Product Deliveries

Product-Oriented Time-Window Assignment for Multi-Compartment Vehicle Routing Problem

Sara Martins[*] · **Manuel Ostermeier**[†] · **Pedro Amorim**[*] · **Alexander Hübner**[†] · **Bernardo Almada-Lobo**[*]

To be submitted to *European Journal of Operational Research*, 2017

Abstract This work extends the research on multi-compartment vehicle routing problems (MCVRPs) by tackling a multi-period setting with a product-oriented time-window assignment. In this problem, a fleet of multi-compartment vehicles (MCVs) is used for distribution and a unique time-window for the delivery of each product to each customer should be defined and used consistently throughout the planning horizon. Besides fuel and waste distribution, one core application of MCVs is the distribution of groceries, as retailers may jointly transport products with different temperature requirements, thus reducing the number of visits to a store. Grocery stores usually define preferable time-windows, which depend on the temperature of products (for example, fresh products such as fruits and vegetables in the morning), to indicate when deliveries should occur to better plan their in-store operations. Hence, distribution planning should take these preferences into consideration to obtain consistent delivery times. An adaptive large neighborhood search is proposed to solve the product-oriented time-window assignment for MCVRP. Daily operators focusing on the improvement of routing aspects of the problem on each day, and weekly operators designed to align the time-window assignment consistently throughout the planning horizon are developed. The approach is tested on benchmark instances from the literature as well as on randomly generated instances to demonstrate its effectiveness. Numerical experiments demonstrated that planning a consistent distribution leads to better overall solutions than an ex-post time-window assignment of daily plans, facilitating more on-time deliveries.

Keywords Grocery retailing · Multi-compartment · Vehicle routing · Consistency · Adaptative Large Neighborhood Search

[*] INESC TEC and Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

[*] Catholic University Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany

5.1. Introduction

This paper introduces a multi-compartment vehicle routing problem (MCVRP) for the assignment of product-oriented time-windows and extends the capacitated vehicle routing problem (Toth and Vigo, 2014; Golden et al., 2008). The problem occurs in grocery distribution of different products from a distribution center (DC) to stores.

An efficient supply chain is essential for the success of any retailer and the distribution planning plays a central role. High product variety with particular temperature requirements characterize grocery distribution. For instance, dairy products require continuous cooling while other products are not sensitive to temperature.

In the past, not all products could be transported within the same vehicle as only one temperature could be set up at the same time. However, in the last years, vehicles with new technologies have been deployed, such as multi-compartment vehicles (MCVs) that are able to split their loading area flexibly into different compartments. These compartments can be individually adjusted to a given temperature. Hence, MCVs allows for the joint distribution of different products on compartments set up with distinct temperatures, while triggering no capacity loss.

With MCVs, retailers need to decide which products are (or not) supplied jointly for each store. Such decision also implies a higher variety of possible delivery times for the stores. In practice, grocery stores usually define preferable time-windows to indicate when deliveries should occur. Stores do need to align their operations to the delivery schedule to make sure their resources are available for unloading, replenishment and stocking operations, which impacts staff scheduling and backroom capacity planning. This is especially significant as these resources are scarce and shared by different activities. Additionally, stores need to manage the available inventory. Besides capacity restrictions in the backroom, the shelf inventory is crucial for sales. The defined time-windows help managing the on-hand inventories at stores as the time until the next supply is known. The scheduling of deliveries therefore needs to be defined according to the stores' time-window requirements, independent from joint or separate deliveries across products. Only in this way a smooth operation regarding the supply of stores can be guaranteed. Delivery times have to be pre-defined and fixed for a given period to efficiently coordinate store resources during the day. This leads to consistent deliveries throughout the planning horizon.

While the current VRP literature usually assigns time-windows to customers (Spliet and Gabor, 2014; Spliet and Desaulniers, 2015), when dealing with multiple products with distinct characteristics, such as grocery products, a more specific product oriented assignment is required. It is not sufficient to consider multiple delivery time-windows per customer (Belhaiza et al., 2014). It is rather necessary to define in which of the available time-windows each product will be delivered (assigned to), ensuring a consistent delivery for a given planning horizon.

In combination with the transportation in MCVs, this rises the question of when each product should be supplied and if the same time-window should be assigned to different products to enable a joint delivery. Our work addresses this special variant of VRP with MCV and time-window assignment. It extends the research on MCVRP by tackling a multi-period setting with a product-oriented time-window assignment that should be consistently used throughout the planning horizon. The main contributions of this paper are as follows. Firstly, a mixed integer programming model for the product-oriented time-window assignment for MCVRP (PTWA-MCVRP) is proposed. Secondly, an adaptive large neighborhood search (ALNS) is designed to solve the problem with daily operators, focusing on the improvement of routing aspects of the problem on each day, and weekly operators, called to align the time-windows consistency throughout the periods. Its effectiveness is tested on benchmark instances of related problems. Thirdly, the effects of product-oriented time-windows and consistent deliveries in an MCVRP environment are analyzed, as well as the decision relevant costs of the problem.

The remainder of this paper is organized as follows. A detailed description of the problem is provided in Section 5.2. Section 5.3 overviews the related literature. In Section 5.4 the mathematical model formulation is given. The ALNS algorithm developed to tackle the problem is described in Section 5.5. Extensive numerical experiments are carried out in Section 5.6, where tests on benchmark instances demonstrate the effectiveness of the solution approach. Additionally, tests on randomly generated data are performed to identify the impact of consistent time-windows assignment on route planning with MCVs. Finally, our findings are summarized in Section 5.7.

5.2. Distribution Process and Requirements

The different aspects informing our problem specification need to be understood thoroughly and are therefore detailed in this section. We will first state the overall planning situation for retailers before we highlight the implications of using MCVs for transportation. Following, we move on to the need to consider multiple periods and a consistent delivery across segments. For the remainder of this paper, we refer to products with similar characteristics and temperature requirements as segments. Additionally, the terms customers and stores are used as equivalents in our context.

Distribution of Groceries. The overall objective of our research is a highly efficient supply chain that runs simultaneously different temperatures (e.g. frozen, chilled and ambient). In grocery distribution, the majority of products are sent from suppliers to retailer distribution centers (DCs), which are responsible for supplying stores according to their needs (Hübner et al., 2013; Martins et al., 2017). The distinct products can be allocated to the same DC, however separate warehouse zones need to be secured according to the products temperature requirements to prevent spoilage. The same reasoning applies to the transportation process, during which the preferred temperature for each product needs to be maintained to guarantee a high product quality and to adhere to legal regulations.

Distribution with MCVs. The integration of MCVs into the distribution enables the simultaneous supply of multiple segments but also poses some new challenges for the planning. More precisely, the joint distribution of segments influences not only the transportation process, but also the upstream and downstream supply chain operations (Hübner and Ostermeier, 2018). On the one hand, different gates have to be approached by an MCV to collect different segments as they are allocated to distinct zones of an DC. This leads to an increase in loading costs that depend on the number of segments assigned to a tour and therefore on the number of compartments needed on each vehicle. This means that an increase in the number of segments on a vehicle leads to additional costs for the corresponding tour regarding the loading. On the other hand, separate deliveries with a single-compartment vehicle for each segment may be avoided, reducing the number of visits to a store and transportation kilometers. Trading-off loading and transportation is one of the central decisions while using MCVs. The underlying problem results in the MCVRP which is a special variant of the capacitated VRP.

The MCVRP considers the traditional routing decisions (assignment of customers to vehicles and the route sequence), but additionally determines the number and size of the compartments used in each vehicle together with the orders assignment to the corresponding compartments. These decisions influence the loading, transportation and unloading costs and therefore the overall decisions on routing (Hübner and Ostermeier, 2018). The objective of the MCVRP is consequently to minimize the total operation costs by defining a routing plan, in which the stores may receive all segments together in one delivery, receive one delivery per segment or a mix of both.

Summing up, the MCVRP decides on what segments are delivered together or separately to a customer and thus how often deliveries take place for each customer on a single day. In our extended problem formulation, this decision needs to be synchronized with the decision on when those deliveries should be performed on each day. This requires the assignment of product-oriented time-windows and a consistent delivery plan.

The Assignment of Product-Oriented Time-Windows. The aim of the PTWA-MCVRP is to determine an individual time-window for each segment that a store has ordered and to use this time-window consistently for the complete planning horizon. This means that a store receives each segment within the same time-window every day. Each segment is thereby considered independently of all other segments for the time-window assignment. However, the same time-window can be selected for different segments as stores usually order more than one segment. This also means that the joint delivery of multiple segments is possible if they are demanded on the same day. Furthermore, individual segments impose distinct replenishment constraints to the stores. For instance, fresh products, such as fish, fruit and vegetables, are normally required to be delivered before the store opening to guarantee a good merchandising performance. Yet, the requirement for consistent delivery times applies to all segments and emerges from the stores need to align their in-store operations. Hereafter, we refer to a given segment of a store for which a product-oriented time-window has to be determined as store-segment pair.

In our application, the assignment of product-oriented time-windows underlies some further requirements and rules. As such, time-window restrictions are not strict and both early and late deliveries to the stores are allowed but undesirable, having a negative impact on store operations. Usually, the delivery products are dispatched to the sales area for replenishment and the excess is stored in the backroom area for later replenishment. Therefore, the backrooms are designed to store just a portion of the deliveries, securing future replenishment, and an early delivery can trigger storage problems at this stage if the deliveries are not processed immediately upon arrival, specially for the refrigerated products (Pires et al., 2017). A late delivery, however, leads to idle times of the personal assigned to the receiving activity and might delay the replenishment of shelves with the risk of causing out-of-stocks. We note that in-store operations yield the highest share of operational costs within the internal supply chain of a retailer, accounting for up to 50% (see e.g., van Zelst et al. (2009) and Kuhn and Sternbeck (2013)). We therefore propose penalty costs for early and late deliveries, resembling the setting of VRPs with soft time-windows (Koskosidis et al., 1992).

An example for the assignment of product-oriented time-windows for different store-segment pairs is depicted in Figure 5.1. It illustrates two examples of routing schedules, considering time-window assignments with and without consistent deliveries. For our application we assume that the daily demand for each segment is known across the planning horizon, i.e., the delivery patterns are an input for the decision problem. Accordingly, the days in which each segment should be delivered for a given store (delivery pattern) are provided on the left part of the figure. For example, segment A (blue square) is delivered every day, while segment D (dotted square) is only delivered on Monday, Wednesday and Friday. Two possible solutions for the time-window assignment are presented. The left schedule represents a solution where consistency is not taken into account, while the right one contains a consistent schedule. In the first case, segments are delivered at different times of the day over the week. In the second case, a delivery within the same time bounds is guaranteed over the whole planning horizon. It shows that segments A & B are supplied between 6 and 7 am every day of the planning horizon. Likewise, deliveries for segments C & D always happen between 8 and 9 am. Clearly, this schedule enables stores to plan their resources according to the consistent plan, allowing for an efficient processing of deliveries. Ultimately, the product-oriented time-window is motivated to reduce the planning complexity for stores and with this increase the stores satisfaction with its supplier.

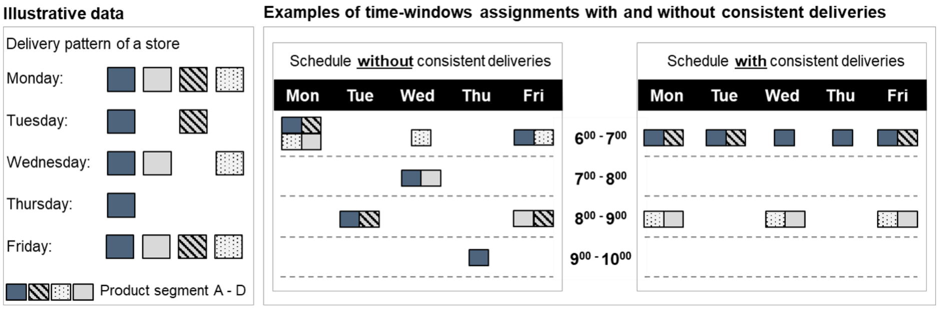

Illustrative data
Delivery pattern of a store
Monday:
Tuesday:
Wednesday:
Thursday:
Friday:
Product segment A - D
Examples of time-windows assignments with and without consistent deliveries
Schedule without consistent deliveries
Schedule with consistent deliveries
Mon
Tue
Wed
Thu
Fri
6⁰⁰ - 7⁰⁰
7⁰⁰ - 8⁰⁰
8⁰⁰ - 9⁰⁰
9⁰⁰ - 10⁰⁰

Figure 5.1: Illustrative example of routing schedules with and without consistent deliveries

Summing up, in a nutshell, the PTWA-MCVRP combines a multi-period VRP for consistent deliveries with an MCVRP. As a consequence, it involves the following partial decisions that are made simultaneously and define the uniqueness of the problem formulation:

- assignment of product-oriented time-windows to each store-segment pair (this decision defines the arrival time of each segment at a store on each delivery day);

- assignment of orders to tours/vehicles (this decision determines which segments are delivered simultaneously);
- number and size of each compartment (this decision defines the number of compartments on each vehicle, and how the capacity is divided between compartments).
- sequence of store visits, as in every VRP (this decision specify the order by which each vehicle should perform the delivery of the orders assigned).

Please note that the first decision addresses time-window assignment to achieve consistent deliveries while the later three are related to MCVRPs.

5.3. Related Literature

The PTWA-MCVRP is a variant of the CVRP (Toth and Vigo, 2014). As previously described, the problem deals with two main groups of decisions: (i) the routing decisions and (ii) the consistent time-window assignment decisions. These decisions relate to three streams of VRP variants. Since the routing decisions are defined considering MCVs, the PTWA-MCVRP clearly extends literature on MCVRPs. The time-window assignment decisions derive from an extension of the time-window assignment vehicle routing problem (TWAVRP), in which one time-window has to be selected for each customer with stochastic demand in a single-period problem, and the consistent vehicle routing problem (ConVRP), where a consistent arrival time has to be achieved in a multi-period setting. The related literature is presented below.

5.3.1 Multi-compartment Vehicle Routing Problems

One of the first references to the MCVRP as a variant of the VRP was given in 1979 (Reed et al., 2014). While at that time the MCVRP did not attract much attention from researchers, the raising deployment of MCVs in the last years is increasing the research on this problem.

The MCVRP literature focuses mainly on its applications to the fuel distribution (Avella et al., 2004; Cornillier et al., 2008; Coelho and Laporte, 2015), waste collection (Muyldermans and Pang, 2010; Reed et al., 2014; Henke et al., 2015) and food distribution (Chajakis and Guignard, 2003; Derigs et al., 2011; Hübner and Ostermeier, 2018). Most of the works on the MCVRP assume that the customers can only be served by one vehicle (Chajakis and Guignard, 2003; Reed et al., 2014; Abdulkader et al., 2015) and/or that the number of compartments and their size are fixed (El Fallahi et al., 2008; Muyldermans and Pang, 2010), which are too restrictive.

Derigs et al. (2011) are the first to consider flexible compartment sizes with multiple deliveries to customers. This new feature creates a more general MCVRP by adding the decisions on the number and size of compartments to the problem. Henke et al. (2015) tackle the MCVRP with discrete flexible compartments, instead of continuous as in Derigs et al. (2011). The authors allow the number of compartments to be smaller than the number of products to be collected and apply a variable neighborhood search to the problem.

Hübner and Ostermeier (2018) study the distribution of groceries with flexible MCVs, incorporating the operational costs of the loading and unloading processes. The authors propose a large neighborhood search to solve the problem. For our problem, the model formulation can be simplified compared to previous works. Since we group the products with the same transportation characteristics in segments and consider completely flexible compartments, the specific allocation of orders to compartments is not required as the order segment represent itself a compartment. The ALNS that we propose to solve the problem uses daily operators based on Derigs et al. (2011) and Hübner and Ostermeier (2018) approaches and incorporate the operational features of the last.

The MCVRP was also extended by some authors to incorporate time-window restrictions, considering fixed compartment sizes. Kaabi and Jabeur (2015) describe an MCVRP where customer orders are composed of different products with associated profits, which are collected once the customer is visited within their time-window. Not all customers may be visited. The authors propose an hybrid approach combining a genetic algorithm and an iterated local search to solve the problem. Kabcome and Mouktonglang (2015) also consider time-windows but as soft constraints. They allow the customers to be visited outside their time-windows at a given cost, distinct for earliness and lateness. An upper bound for the violation of time-windows is set as a hard constraint. The authors use a commercial software to solve exactly small instances. Although both works consider time-window restrictions, they are incorporated as inputs rather than decision variables. Therefore, a model and solution approach that integrates the MCVRP with time-window assignment decisions has not yet been analyzed in the literature but has a practical relevance as described in Section 5.2.

5.3.2 Time-Window Assignment & Consistent Vehicle Routing Problems

The TWAVRP was first introduced by Spliet and Gabor (2014). The problem is described as the assignment of time windows to each customer before demand is known. Afterwards, when the demand is revealed, a vehicle routing schedule is made to satisfy the assigned time windows. The demand realization is described based on a given scenario, from a finite set of scenarios, each with a known probability of occurrence. The work on TWAVRP by Spliet and Gabor (2014) assigns time-windows of pre-specified width from a set of exogenous time-windows. The authors develop a branch-price-and-cut algorithm to find the optimal expected traveling time, according to the assignment made. Spliet and Desaulniers (2015) extended that work by considering the assignment of time-windows from a discrete set of a priori constructed windows. They propose a branch-price-and-cut algorithm and five column generation heuristics. Spliet et al. (2017) introduced time-dependent travel times to the TWAVRP, focusing on predictable variations. The authors develop a branch-price-and-cut algorithm, using an exact labeling and a tabu search heuristic to solve the pricing problem. Jabali et al. (2015) consider a similar problem, in which travel times are stochastic but the demand is deterministic. The goal is to define a single route plan that should be fixed for all days of the year together with the time-windows assignment. The authors propose a tabu search algorithm to minimize the total traveling costs and expected earliness and tardiness penalty costs, assuming soft time-windows. Our work is an extension of the

TWAVRP for a multi-period setting, in which a consistent time-window assignment has to be defined for all periods of the planing horizon with deterministic demand. Additionally, our problem tackles a product-oriented time-window assignment rather than just customer oriented.

Feillet et al. (2014) present a bi-objective time-consistent VRP, which aims at minimizing besides the total travel time, the maximum number of time classes in which a customer is visited. The authors solve the problem with a large neighborhood search. Subramanyam and Gounaris (2016) develop an exact solution for the consistent traveling salesman problem (CTSP), which aims to identify the minimum-cost set of routes that a single vehicle should follow for a given period, ensuring that the customers are visited roughly at the same time of the day. Subramanyam and Gounaris (2017) extend the previous work from an TSP to an VRP variant and show that each scenario of the TWAVRP stochastic model can be reduced to an VRP variant known as consistent vehicle routing problem (ConVRP). The authors adapted an exact algorithm for the ConVRP and solved the TWAVRP benchmark instances.

The ConVRP is a multi-period problem proposed by Groër et al. (2009) that aims to design consistent routes over a given planning horizon. Kovacs et al. (2014a) describe three different types of a consistent routing: (i) arrival-time consistency, which ensures visits to customers at roughly the same time of the day, (ii) person-oriented consistency, which means that customers are visited by the same driver, and (iii) delivery consistency, where customers receive roughly the same quantity of goods. The arrival-time consistency requirement of the ConVRP is similar to an TWAVRP, as both problems define an interval of time within which customers visits should occur.

Most of the literature on ConVRP focuses on arrival-time and driver consistency. The first approaches to solve the problem are based on template concepts, wherein template routes are built considering the frequent customers and afterwards a daily plan is derived to include the remaining customers to be visited on each day (Groër et al., 2009; Tarantilis et al., 2012; Kovacs et al., 2014b). Groër et al. (2009) propose a multi-start solution construction combined with a Record-to-Record travel local search metaheuristic algorithm, and Tarantilis et al. (2012) a tabu search algorithm. Kovacs et al. (2014b) present an ALNS to solve the problem and analyze the variant of allowing later departures from depot. Sungur et al. (2010) also use the template concept for the courier delivery problem with uncertainty, where the problem is solved by a master and daily scheduler heuristic (MADS).

Kovacs et al. (2015a) generalize the ConVRP, allowing the customer to be visited by a limited number of drivers and penalizing the arrival-time variation in the objective function. The authors propose a large neighborhood search algorithm to solve the problem. They found that both driver and arrival-time consistency have a small impact on the fleet size. Kovacs et al. (2015b) and Lian et al. (2016) study the multi-objective ConVRP, where driver and arrival-time consistency are considered as objectives, besides the traveling cost. Both works analyze the trade-off between traveling cost and service consistency, and develop a multi-directional local search (MDLS), combined with a large neighborhood search to approximate a Pareto frontier. Kovacs et al. (2015b) also propose two exact approaches based on the ε-constraint framework and state that, on average, it is possible to achieve

70% better arrival-time consistency by increasing the traveling cost by not more than 4%.

In this work we use the concept of consistency regarding the time-window assignment. A maximum width between arrivals can be defined as for the ConVRP, but not all arrivals times are preferable. This means that we need to define a set of time-windows from where the assignment is made in order to incorporate the stores preferable times for deliveries. The instances proposed for the ConVRP by Groër et al. (2009) and Kovacs et al. (2014b) will be extended and used in this work to evaluate the effectiveness of the solution approach proposed.

5.4. Problem definition

The PTWA-MCVRP is defined on a complete undirected, weighted graph $G = (N, E)$, where $N = \{0, 1, \dots, n\}$ is the set of nodes and $E = \{(i, j) : i, j \in N\}$ is the set of edges. Node 0 represents the DC location. Each edge $(i, j) \in E$ is associated with traveling cost tc_{ij} and traveling time tt_{ij}. It is assumed that all traveling costs satisfy the triangle inequality and each tour starts and ends at the DC. Let V be the set of vehicles available for transportation at the DC. The number of vehicles available is assumed to be sufficiently large to fulfill the demand of all stores and consists of identical vehicles with total capacity Q. Each vehicle departs from the depot at time zero and must return before time T, defined as the tour maximum duration. Waiting time between deliveries is not allowed. As we consider MCVs, the loading area of each vehicle can be split into a limited number of compartments $c \in C$, with $C = \{1, ..., c\}$. The number of compartments a vehicle may have active is indicated by $k \in K$, with $K = \{1, ..., c\}$. Due to the explained characteristics of an MCVRP, besides the traveling costs also loading and unloading costs need to be considered. In line with this, l_k represents the loading cost of a vehicle dependent on the number of k compartments used and u indicates the unloading cost of each stop (Hübner and Ostermeier, 2018).

The DC is responsible for the distribution of products from S segments for a given planning horizon consisting of D days. It is assumed that the delivery pattern of each store for each segment (days in which a delivery should be made (Holzapfel et al., 2016)) are known, as well as the quantities to be delivered. Thus, q_{is}^d defines the quantity of an order for segment s to be delivered to store $i \in N \backslash \{0\}$ on day d, and st_{is}^d the correspondent variable service time that depend on the delivery quantity. Additionally, each time a vehicle stops at a store a fixed service time sf is incurred, which is independent on the delivery quantity. Note that if day d does not belong to the delivery pattern of store i for segment s, q_{is}^d is set to 0.

Time-windows are defined by the set $TW = \{1, ..., tw\}$ and the intervals $[e_t, h_t]$, for every $t \in TW$, indicating the earliest (e_t) and latest (h_t) delivery time of time-window t. The assignment of time-windows will be made for each pair store-segment denoted as $((i, s) : i \in N \backslash \{0\}, s \in S)$. Since, not all time-windows can be used to each pair store-segment, the subset $TW_{is} \in TW$ indicates which time-windows can be used for each pair (i, s). The negative impact of having earlier and later deliveries than the time-window assigned to a store is accounted by λ and β, which represent the unitary cost associated with one unit of time the delivery is too early or late, respectively (Ioannou et al., 2003).

The objective of the PTWA-MCVRP is to minimize the total routing costs, considering traveling, loading and unloading costs, as well as the penalties for earlier or later deliveries than planned, while satisfying the stores demand and performing consistent deliveries. A solution for the problem determines not only (i) the sequence of store visits, but also (ii) the number of compartments used in each vehicle, (iii) the assignment of orders (from different stores and segments) to each compartment, and (iv) the size of each compartment. Additionally, (v) the assignment of a time-window to be used consistently throughout the planning horizon for each pair store-segment is performed. The first group of decisions is related with MCVRP. The last decision is associated with the new feature of the problem, regarding the consistent assignment of time-windows along with penalty costs. The model developed uses the following decision variables to map the discussed decisions.

- $b^d_{ijv} = 1$ if vehicle v travels from location i to j on day d and $b^d_{ijv} = 0$ otherwise
- $\theta^d_{isv} = 1$ if segment s is delivered by vehicle v to store i on day d and $\theta^d_{isv} = 0$ otherwise
- $y_{ist} = 1$ if store-segment pair (i,s) is assigned to time-window t and $y_{ist} = 0$ otherwise
- $\vartheta^d_{vs} = 1$ if vehicle v transports segment s on day d and $\vartheta^d_{vs} = 0$ otherwise
- $r^d_{vk} = 1$ if vehicle v has k active compartments on day d and $r^d_{vk} = 0$ otherwise

The continuous variables w^d_{iv} denote the arrival time at store i by vehicle v on day d and p^d_{is} the penalty cost incurred on day d by the pair (i,s) in case of early or late deliveries. Additionally, the discrete auxiliary variables f^d_v represent the number of store stops performed by vehicle v on day d.

Finally, the formulation of the PTWA-MCVRP is given below:

$$Minimize \sum_{d\in D}\sum_{v\in V}\Big[\sum_{k\in K} l_k \cdot r^d_{vk} + \sum_{i\in N}\sum_{j\in N} tc_{ij}\cdot b^d_{ijv} + u\cdot f^d_v\Big] + \sum_{d\in D}\sum_{i\in N\backslash\{0\}}\sum_{s\in S} p^d_{is} \tag{5.1}$$

subject to

$$\sum_{j\in N\backslash\{0\}} b^d_{0jv} \le 1 \qquad v\in V,\ d\in D \tag{5.2}$$

$$\sum_{i\in N} b^d_{igv} = \sum_{j\in N} b^d_{gjv} \qquad v\in V,\ g\in N,\ d\in D \tag{5.3}$$

$$\sum_{v\in V} \theta^d_{isv} \ge M\cdot q^d_{is} \qquad i\in N\backslash\{0\},\ s\in S, d\in D \tag{5.4}$$

$$\sum_{s\in S} \theta^d_{jsv} \le |S| \sum_{i\in N} b^d_{ijv} \qquad v\in V,\ j\in N\backslash\{0\},\ d\in D \tag{5.5}$$

$$\sum_{i\in N\backslash\{0\}}\sum_{s\in S} q^d_{is}\cdot\theta^d_{isv} \le Q \qquad v\in V,\ d\in D \tag{5.6}$$

$$\sum_{i\in N\backslash\{0\}} \theta^d_{isv} \le M\cdot\vartheta^d_{vs} \qquad v\in V, s\in S, d\in D \tag{5.7}$$

$$\sum_{s\in S} \vartheta^d_{vs} = \sum_{k\in K} k\cdot r^d_{vk} \qquad v\in V, d\in D, k\in K \tag{5.8}$$

$$\sum_{k \in K} r_{vk}^d = 1 \qquad v \in V, d \in D \tag{5.9}$$

$$f_v^d \geq \sum_{i \in N} \sum_{j \in N \backslash \{0\}} b_{ijv}^d \qquad v \in V, d \in D \tag{5.10}$$

$$\sum_{t \in TW_{is}} y_{ist} = 1 \qquad i \in N \backslash \{0\},\ s \in S \tag{5.11}$$

$$w_{0v}^d \leq 0 \qquad v \in V, d \in D \tag{5.12}$$

$$w_{iv}^d \leq M \cdot \sum_{s \in S} \theta_{isv}^d \qquad v \in V,\ i \in N,\ d \in D \tag{5.13}$$

$$w_{iv}^d + tt_{ij} + sf + \sum_{s \in S} \theta_{isv}^d \cdot st_{is}^d - M \cdot (1 - b_{ijv}^d) \leq w_{jv}^d \qquad v \in V,\ i \in N,\ j \in N \backslash \{0\},\ d \in D \tag{5.14}$$

$$w_{iv}^d + tt_{ij} + sf + \sum_{s \in S} \theta_{isv}^d \cdot st_{is}^d + M \cdot (1 - b_{ijv}^d) \geq w_{jv}^d \qquad v \in V,\ i \in N,\ j \in N \backslash \{0\},\ d \in D \tag{5.15}$$

$$w_{iv}^d + tt_{i0} + sf + \sum_{s \in S} \theta_{isv}^d \cdot st_{is}^d - w_{0v}^d - M \cdot (1 - b_{i0v}^d) \leq T \qquad v \in V,\ i \in N \backslash \{0\},\ d \in D \tag{5.16}$$

$$p_{is}^d \geq ((\sum_{t \in TW_{is}} y_{ist} \cdot e_t) - w_{iv}^d) \cdot \lambda - M \cdot (1 - \theta_{isv}^d) \qquad v \in V,\ i \in N,\ s \in S,\ d \in D \tag{5.17}$$

$$p_{is}^d \geq (w_{iv}^d - \sum_{t \in TW_{is}} y_{ist} \cdot h_t) \cdot \beta - M \cdot (1 - \theta_{isv}^d) \qquad v \in V,\ i \in N,\ s \in S,\ d \in D \tag{5.18}$$

Objective function (5.1) minimizes the total routing costs, including loading, traveling and unloading costs, plus the penalty costs of performing deliveries outside the bounds of the time-windows assigned. The constraints of the problems can be aggregated in two groups. Constraints (5.2)-(5.10) compose the first group, which is related to the routing decisions of the MCVRP. Inequalities (5.2) and (5.3) ensure that each route starts at the depot, and that a store has only one predecessor and one successor in the route. Constraints (5.4) guarantees that a store receives all segments that it requires on each day. Constraints (5.5) ensure that store deliveries are only performed by a vehicle that actually visits the store. The vehicles' capacity is controlled by constraints (5.6). Inequalities (5.7)-(5.9) define which segments are loaded on the vehicles and, consequently, how many compartments will be used. Constraints (5.10) determine the number of store stops each vehicle performs on a given day. The remaining constraints compose the second group and regard the time-windows assignment. Constraints (5.11) ensure that only one time-window can be assigned to each pair store-segment. These constraints guarantee consistent deliveries and are hereafter denoted as *time-window consistency constraint*. The departures from the depot at time zero is ensured by inequalities (5.12). The arrival times to stores are set by constraints (5.13)-(5.15), guaranteeing that waiting time between deliveries is not allowed. Constraints (5.16) ensure that the tours do not exceed the maximum duration established. The penalty costs incurred by performing earlier or later deliveries than the bounds of the time-window assigned are determined by constraints (5.17) and (5.18).

5.5. Solution Approach

The main difficulty in solving the PTWA-MCVRP arises from the interrelation between the individual days of the planning horizon. In the considered problem, this interrelation relates to the consistent use of a unique time-window for each pair customer-segment throughout the planning horizon. An ALNS framework is designed in this paper to cope with the characteristics of the problem regarding its two main groups of decisions: the routing problem with MCVs and the time-window assignment.

ALNS algorithms are used in the literature to solve different problem settings and showed to provide good results for distinct VRP variants, such as MCVRP (Derigs et al., 2011), VRPTW (Ropke and Pisinger, 2006), ConVRP (Kovacs et al., 2014b). The ALNS framework was first introduced by Ropke and Pisinger (2006). Its central idea is to sequentially improve an initial solution by destroying and rebuilding parts of it. In the VRP variants, the destroying phase uses a destroy operator to remove a given number of requests from the routes, which afterwards are re-inserted according to an insertion operator, in the rebuilding phase. In an ALNS framework, several destroy and insertion operators are available and selected during the search procedure in an adaptive way, dependent on their performance during the search (see Section 5.5.6).

The ALNS framework developed to solve the PTWA-MCVRP combines daily and weekly operators to tackle the different problem decisions. The daily operators focus on a particular day and try to improve the routing decisions of the problem, while the weekly operators have a broader scope, analyzing all the days at the same time, and aligning the time-window assignment decisions. The pseudocode of the ALNS framework developed is shown in Algorithm 2. The general framework is explained below and the main features are detailed in the subsequent subsections.

Algorithm 2 ALNS scheme for the product-oriented time-window assignment for MCVRP

```
generate a solution $S$                                                        ▷ Section 5.5.1
set $S_{best} := S$
repeat
    select a destroy-repair heuristic pair $(d, r)$ based on adaptive weights $(\rho_{dr})$   ▷ Section 5.5.6
    if $d$ is a daily operator then                                   ▷ Section 5.5.2-Section 5.5.4
        select randomly day $t$
        generate solution $S'$ by applying $(d, r)$ to $S$ on day $t$
    else                                                              ▷ Section 5.5.2-Section 5.5.4
        generate solution $S'$ by applying $(d, r)$ to $S$
    end if
    if $S'$ better than $S_{best}$ then                                      ▷ Section 5.5.5
        $S_{best} = S'$
        $S := S'$
    else if $S'$ complies with the acceptance criteria then                  ▷ Section 5.5.5
        $S := S'$
    end if
    update performance of destroy-repair heuristic pair $(d, r)$             ▷ Section 5.5.6
until maximum number of iterations is reached
return $S_{best}$
```

The algorithm starts with the generation of an initial solution S (see Section 5.5.1), not considering the *time-window consistency constraint*. This constraint will be enforced during the search by adding to the objective function ($f(S')$) a violation cost as described in Equation (5.19), creating an artificial objective function ($f_a(S')$) similar to Kovacs et al. (2014b).

$$f_a(S') = f(S') + \zeta \cdot inconPairs \tag{5.19}$$

The violation cost is set proportional to the number of pairs customer-segment with an inconsistent delivery plan (*inconPairs*), i.e. number of pairs with more than one time-window assigned, and on a consistency cost ζ. Parameter ζ is initialized at the beginning of the search and updated every 100 iterations ($\zeta = \exp^{(iterations/\delta)}$). Hence, the more advanced we are in the search, the more costly it is to violate the *time-window consistency constraint*.

In each iteration, a destroy-repair heuristic pair (d, r) is chosen by a roulette wheel selection, recurring to adaptive weights (see Section 5.5.6). The destroy operator can be selected from the group of daily or weekly operators (see Section 5.5.2). The removal step is followed by the reinsertion phase. The reinsertion is performed by the selected repair operator and each order is reinserted for the corresponding day on which the order is scheduled (see Section 5.5.3). After each remove and insertion, the arrival times of the orders are updated and the time-window assignments are reset (see Section 5.5.4). If the new solution S' meets the acceptance criteria, then it replaces S. If it improves the best solution found so far it replaces S_{best} (see Section 5.5.5).

5.5.1 Initial Solution

The initial solution is generated by applying the Clarke Wright savings heuristic (Clarke and Wright, 1964) to each individual day of the planning horizon. This approach starts by creating routes with single orders and afterwards iteratively combines routes according to a calculated saving in traveling distance, while satisfying the vehicles capacity and maximum duration constraints. This heuristic is commonly used in different VRP problems and was chosen because it provides a fast solution with a reasonable traveling distance.

With the routes defined, the arrival times to each customer are calculated, assuming the departures from the depot at time zero of each day. A time-window is assigned to each individual order based on their arrival time, guaranteeing no penalty cost, i.e. not performing early or late deliveries. Note that the orders of each pair customer-segment can be assigned to distinct time-windows. Hence, the first generated solution is most probably not feasible regarding the *time-window consistency constraint*, therefore generating a violation cost.

5.5.2 Destroy Operators

In this solution approach, the destroy operators are separated in daily and weekly operators. We use six destroy operators, three for each group. Each of the operators was developed to address a special characteristic of the problem. The daily destroy operators focus on a specific day of the planning horizon and, therefore focus on the routing decisions of the problem. The weekly destroy operators are the unique feature of our search procedure and were created to tackle the consistency aspect of the PTWA-MCVRP, thus focusing on the time-window assignment decisions. In contrast to the daily operators they analyze the entire planning horizon at once with the aim of aligning the time-window assignment for the customer-segment pairs.

5.5.2.1 Daily Operators

The daily operators remove r orders for a given day from its routes. The day-selection is random, however a day is set tabu after its selection until all other days have also been selected, independently of the quality of the solution generated. The number of removes r is chosen randomly from the interval $[0.1 \cdot N_{day}, 0.4 \cdot N_{day}]$, where N_{day} is the total number of orders to be delivered on the specific *day*. This interval is proposed by Ropke and Pisinger (2006) and used in different works of the VRP literature.

The three daily destroy operators used are: *random removal*, *Shaw removal* and *worst removal*. These operators were proposed by Shaw (1997) and Ropke and Pisinger (2006) and are frequently used in the ALNS for different VRP variants. The general idea of each operator is given below.

The *random removal* operator removes randomly the orders from the set of routes of the day selected. The *Shaw removal* removes the orders based on a similarity measure. The similarity between two orders (z,m) is calculated based on four terms: distance, capacity, arrival time and segment affiliation. These terms are weighted using the weights ϕ, ψ, φ and ω, respectively. The complete similarity measure is given by Equation (5.20). d_{zm} represents the distance between corresponding customers of the orders, $q(z)$ the order size, $a(z)$ the arrival time and s_{zm} the orders segment affiliation, i.e. $s_{zm} = 1$ if they are from the same segment, 0 otherwise. The parameters d_{max}, q_{max} and a_{max} indicate the maximum distance between two customers and the maximum quantity and arrival time difference between any two orders across all available orders.

$$R_{zm} := \phi \cdot \frac{d_{zm}}{d_{max}} + \psi \cdot \frac{|q(z) - q(m)|}{q_{max}} + \varphi \cdot \frac{|a(z) - a(m)|}{a_{max}} + \omega \cdot s_{zm} \tag{5.20}$$

The smaller R_{zm} gets, the more similar the orders are. In addition to the similarity measure, a randomization is used according to Shaw (1997) to diversify the search and ensure that not the most similar order is chosen during the search. During the procedure, the first order for removal is chosen randomly, all further orders are based on the similarity to one of the already removed ones. Finally, the *worst removal* removes the orders that seem to be in a costly position in the solution. The cost of an order is the difference between the current solution cost and the solution cost if the order was removed (not having any additional cost of not being delivered). In this approach the solution cost is evaluated by the artificial objective function (f_a), which comprises the routing and penalty cost plus the consistency violation cost. A randomized process is also integrated in this operator to ensure that not always the order with the worst cost is removed.

5.5.2.2 Weekly Operators

The weekly operators remove the orders of r pairs customer-segment from all days of the planning horizon. The number of removes r is chosen randomly from the interval $[2, 0.1 \cdot n]$, where n represents the number of customers. Since all the orders of the r pairs are removed the value of r has to be more restrictive than for the daily operators. By way of example, if two pairs are chosen with 5 orders for each pair throughout the planning horizon, this results already in 10 orders for removal.

Three weekly destroy operators are designed specifically to tackle the assignment decisions of the problem: *product-based removal*, *worst time-window removal*, *worst arrival removal*. The *product-based removal* is a variant of a random removal. The operator selects randomly r pairs of customer-segments and removes all orders of that pair from the solution, i.e., all orders a customer placed of a segment in the planning horizon are removed. The aim of this operator is to diversify the assignment of time-windows. The last two operators are variants of a *worst removal*, based on the operators designed by Kovacs et al. (2015a). The *worst time-window removal* operator calculates the number of time-window assigned to each customer-segment pair along the planning horizon and removes all the orders of the r pairs with the highest number of assignments. This operator aims at reducing the number of time-windows used for each pair, favoring more consistent deliveries. The *worst arrival removal* considers the maximum arrival time difference between two orders of a customer-segment pair. It identifies the r pairs with the highest arrival time deviation and removes the two corresponding orders of each pair. It also aims at reducing the number of time-windows but it is less disruptive since it does not remove all the orders from the same pair customer-segment.

5.5.3 Repair Operators

Once a destroy operator is applied and orders are removed, the repair operator selected rebuilds the solution by reinserting the orders into the routes of their delivery days. If the removals were made by a daily removal operator, all the orders removed are from the same day of delivery and the repair operator only considers that day. Otherwise, for weekly removals, the repair operator will focus on each day separately. From the list of days from which orders have been removed, a day is selected at random and all the orders of the corresponding day are reinserted according to the repair operator chosen. The process is repeated until all days are rebuilt.

Following most of the VRP literature that uses ALNS as a solution method, four insertion heuristics are applied as repair operators: one greedy insertion and three regret insertions. These operators are based on Ropke and Pisinger (2006) and account for all costs of the problem, including consistency violation cost. The *greedy insertion* operator calculates for each order removed the cheapest feasible position for reinsertion and the order with the lowest cost increase is selected to be inserted. The process is repeated until all orders are inserted. The *regret insertion* operators improve the greedy insertion by analyzing not only the best option for each order but the k best. This procedure integrates ahead information and calculates the regret of postponing an insertion. Let Δ_z^j denote the change in the objective value for inserting order z at its best feasible position on the $j-th$ cheapest route. The regret value is calculated according to Equation (5.21) for all the orders removed.

$$regret_k^z := \sum_{j=2}^{k} (\Delta_z^j - \Delta_z^1) \tag{5.21}$$

The order z with the highest regret value is selected to be inserted at its best feasible position. In each insertion, the regret value is recalculated for the set of orders remaining on the removal list, until all orders are inserted. Three regret insertion operators are used with $k \in \{2,3,4\}$.

5.5.4 Update of Arrival Times and Time-Windows Assignments

Since waiting time between deliveries is not allowed, every remove or insertion in a route impacts the arrival time of the successive orders and thus times need to be updated. We assume that all routes start at the depot at time 0 and therefore the arrival time of each order is calculated by consecutively adding the travel times between the customers visited (tt_{ij}) and their correspondent service time. As previously mention, the service time at a customer has a variable component proportional to the size of the orders delivered (st_{is}^d) and a fixed component per stop (sf, see also Section 5.2). Note that a customer can receive more than one order across different segments. In this case, the same arrival time is set for the correspondent customer orders and the service time at the customer is the total variable service time of the distinct orders plus the fixed service time.

Once the arrival times are updated, the new arrival time of an order might lie outside the bounds of the assigned time-window. However, since the problem considers time-windows bounds as soft constraints, the solution is still feasible, but it yields a penalty cost for an early or late delivery. Due to the solution approach features, we can decide to change the time-window assigned to the order to avoid penalty cost which could cause an inconsistent assignment or accept the penalty cost that maintains the consistent assignment. Therefore, it is necessary to evaluate if the time-window assignment should be altered or maintained in order to achieve the minimal cost assignment, i.e. the best option between accepting a penalty cost or an additional consistency violation cost. Note that during the search, the time-window assignment will be more restrictive as the consistency cost ζ increases.

The time-window assignment update procedure is performed for each order separately, after the arrival time is determined. We define as $current_{cost}$ the current penalty and consistency violation cost induced by the pair customer-segment disregarding the order under analysis. This means, we calculate the penalty and consistency costs for all orders of the customer-segment pair without the order which is currently under consideration. When deciding on the new time-window assignment for the order under analysis, there are two possible situations, as previously mentioned:

- Update only the time-window assignment of the order in analysis, maintaining the previous assignment of the pair remaining orders. With this decision the total assignment cost of the pair customer-segment is calculated by adding to the $current_{cost}$ the penalty and consistency violation cost associated with the time-window assignment for the new order. The total cost of this situation is denoted as $singleUpdate_{cost}$, corresponding to a single assignment update.

- Update the assignment of all the orders from the pair customer-segment to the same time-window, ensuring a consistent delivery throughout the planning horizon. This decision guarantees a null consistency violation cost (zero cost, as only one time-window is assigned to the pair). Therefore, the total assignment cost of the pair comprises only the penalty costs that all the orders incur due to the new time-window assigned. The total cost of this situation is denoted as $groupUpdate_{cost}$ and a new assignment is performed for the group of orders.

During the procedure, both costs $singleUpdate_{cost}$ and $groupUpdate_{cost}$ are calculated for each of the time-window available of the pair customer-segment (TW_{is}) and the cheapest assignment of all is chosen. The penalty cost that an order incurs is calculated by comparing the order arrival time with the bounds of the time-window assigned. If the arrival time lies outside the time-window bounds it causes a penalty cost proportional to the deviation. The consistency violation cost of updating the time-window of a single order is determined by checking if the time-window assigned is already used by one of the other orders of the pair. If the time-window has not been used the consistency violation cost is increased proportionally to the current consistency cost (ζ).

5.5.5 Acceptance Criteria

The solution approach proposed uses a simulated annealing framework to evaluate and accept the solutions generated. A new solution S' is accepted as S_{best} if it improves the best solution. Otherwise, it is compared against the incumbent solution S by means of the probability $e^{-(f_a(S')-f_a(S))/\hat{t}}$. The parameter $\hat{t}$ denotes the current temperature. It is initialized at the beginning of the search and decreased in every iteration with a cooling rate γ ($\hat{t} = \hat{t} \cdot \gamma$). Solutions are compared regarding the artificial objective function f_a, allowing infeasible solutions to be accepted during the search. The ALNS algorithm stops after a given number of iterations.

5.5.6 Selection of a Destroy-Repair Heuristic Pair (d,r)

The pairwise selection of the destroy-repair heuristic pair is applied in each iteration and based on a roulette wheel selection principle, as proposed in Ropke and Pisinger (2006). The probability Φ_{dr} of a pair (d,r) to be chosen is given by Equation (5.22), where ρ_{dr} denotes the weight of the heuristic pair.

$$\Phi_{dr} := \frac{\rho_{dr}}{\sum_{d'=1}^{n_d} \sum_{r'=1}^{n_r} \rho_{d'r'}} \tag{5.22}$$

The weights ρ_{dr} are set to one at the beginning of the procedure, and updated dynamically during the search. Each pair (d,r) is associated with a score Ψ_{dr} that is updated each time the heuristic pair is applied according to the following criteria:

- Ψ_{dr}+σ_1, if the heuristic pair generates a new best solution;
- Ψ_{dr}+σ_2, if the heuristic pair generates a solution that has not been visited before, and is accepted as the new incumbent solution S.

As in Ropke and Pisinger (2006), the scores are initialized to zero and updated at each iteration according to the previous criteria. Every 100 iterations, the weights ρ_{dr} are updated according to the recursive Equation (5.23) and the scores are reset to zero for the next round.

$$\rho_{dr} := (1-\alpha)\rho_{dr} + \alpha\frac{\Psi_{dr}}{max(1,\Theta_{dr})} \tag{5.23}$$

The parameter α is a reaction factor that controls how the weights are influenced by past and recent performances. In this way it guides the search by controlling how sensitive the operator choice reacts to changes during the search. The reaction factor α is initialized with a small value that is increased once a feasible solution is saved as best solution.

5.6. Numerical Experiments

Extensive numerical experiments are needed to examine (i) the effectiveness and efficiency of the solution approach presented in Section 5.5, (ii) the impact of introducing consistent deliveries and (iii) the impact and implications of the introduced product-oriented time-window assignment. First, the performance of the ALNS is analyzed in Section 5.6.2, where solutions from related problem formulations from the literature are compared and an analysis of the operators execution is performed. Extensive analysis concerning the impact of our new model for the grocery distribution are then presented in Section 5.6.3. The data sets used for these analyzes are described in Section 5.6.1.

The computational results presented in this section were obtained on a 3.60 GHz PC with 16 GB memory. The algorithm was implemented in C++ and ran 10 times per instance, in all the tests performed, stopping after 60 000 iterations. The search parameters used are specified in Table 5.1. Each parameter was set to either a value reported in the literature (Ropke and Pisinger, 2006; Derigs et al., 2011; Hübner and Ostermeier, 2018) or to an initial guess that was tuned. All other parameters, e.g. number of removals, are as given in Section 5.5.

Table 5.1: Search parameters setting

δ	ϕ	ψ	φ	ω	γ	σ_1	σ_2	α
300	0.28	0.16	0.28	0.28	0.99975	33	13	0.1

5.6.1 Overview of the data sets tested

In a first analysis, our problem and solution approach are compared to another VRP variant that considers a consistent delivery over multiple periods. This comparison is made to benchmark results provided by Kovacs et al. (2014b) for the ConVRP. This experiment was chosen as Kovacs et al. (2014b) consider a consistency problem with a departure time at the depot at time zero. The further analysis tackle the specific problem characteristics and are therefore performed on data based on grocery distribution. The data sets used in both cases are described in the following.

5.6.1.1 ConVRP data set

The data sets used as benchmark were proposed by Groër et al. (2009) and Kovacs et al. (2014b), which were based on Christofides and Eilon (1969) instances for VRP considering a visit frequency of 70% (Groër et al., 2009), as well as 50% and 90% (Kovacs et al., 2014b). The visit frequency indicates the likelihood of a customer placing an order for each day in the planning horizon. As our paper deals with a real-life problem in grocery distribution, we focus on the set of instances with a given maximum duration for tours and provided service times. We tested the instances with 50 to 100 customers. However, the problem tackled by Kovacs et al. (2014b) is a ConVRP that considers besides the arrival-time consistency a driver consistency. Their goal is to minimize the traveling time while satisfying the two consistencies, i.e. approach each customer by the same driver every day with a maximum arrival-time deviation (l_{max}) smaller than a pre-defined width L. These two aspects are not considered in our problem formulation and therefore modifications had to be applied to enable a fair comparison of both solution approaches.

The driver consistency is added and treated like the *time-window consistency constraint*. That means that we penalize the use of multiple drivers by adding to the artificial objective function ($f_a(S')$) a violation cost that is increased during the search, as explained in Section 5.5. The best l_{max} found by Kovacs et al. (2014b) is used as input to set the width of our set of time-windows. Following this approach, the earliest given time-window starts at time 0 with a l_{max} width and, from there, new time-windows are available with a shift of one time-unit. Since Kovacs et al. (2014b) perform a set of tests with the instances, our comparison is made considering the l_{max} achieved for the tests with a maximum arrival time bound denoted as L_1, which the authors defined by running their algorithm without bounding the arrival time differences.

Further, to ensure deliveries take place within the given time-windows we need to consider hard time-windows and therefore we set the penalization costs for early/late deliveries to a very large number. Lastly, as only traveling times are considered, loading and unloading costs are set to 0.

5.6.1.2 Grocery distribution data set

To analyze the problem setting of grocery distribution we used randomly generated instances (20 instances), which comprise a planning horizon of 7 delivery days and 50 stores served from a given DC. The order structure is based on a visit frequency of 70% as proposed by Groër et al. (2009). If a store is flagged to be visited on a given day, it will place orders for all four available segments. The remaining data settings are based on a case study with a major European retailer. Following the insights from the study, loading/unloading costs have been derived and are used in the experiments. The loading costs are presented in Table 5.2 and depend on the number of compartments per vehicle. Unloading costs accrue with every customer stop and are set to 2.20 currency units (CU). The transportation costs are based on the travel distance between any two locations i and j, $i, j \in N$. All experiments assume a vehicle capacity of 33 transportation units (TU). Further, an early or late delivery is penalized with 0.17 CU per minute, covering the working cost of a store employee.

Table 5.2: Applied costs for loading MCV

# Compartments	1	2	3	4
Loading (CU/shipping gate)	2.70	5.57	8.27	10.97

The demand for each segment maps the actual demand structure in practice. For each segment the individual order size is randomly chosen between a given minimum and maximum order quantity. For the first segment, the order quantity ranges between 1 and 5 TUs, for the second segment between 1 and 10 TUs. Furthermore, to map segments with a higher sales volume, segment 3 ranges between 5 and 20 TUs and finally segment 4 between 10 and 25 TUs. The distance information is based on the VRP instances by Christofides and Eilon (1969). As the vehicle capacity in the application by Christofides and Eilon (1969) is far higher than our given capacity, we modified the distances to suffice our requirements in practice and still achieve a comparable route length. We therefore multiply the given distances by four to increase the travel distances, but kept the same density.

5.6.2 Algorithm performance

The average variation coefficient (standard deviation/mean) for the tests performed with the ALNS is 0.012 for the ConVRP data sets and 0.003 for the grocery distribution ones. This last value demonstrate that the ALNS proposed is able to provide stable solutions for the PTWA-MCVRP. Regarding the computational effort, the solution approach takes between one to one and half hours to provide a solution for the grocery distribution problem. As this is a tactical decision planning, these values are acceptable in practice. The runtime required to solve the ConVRP instances is more variable, increasing proportional to the number of customer and visiting frequency. For the instances with 50 and 75 customers the runtime ranges from one hour for 50% visit frequency to four hours for 90% visit frequency. The computational effort for the 100 customers instances ranges from four to twenty hours. The data sets from the ConVRP require more time to solve because it has to analyze, during the search, a much higher number of time-windows. As the number of customers and time-windows available increase, the update of arrival times and time-windows assignment procedure (see Section 5.5.4) runtime increases. In the following sections, we analyze the results for the ConVRP benchmark instances and the application frequency of the proposed

ALNS operators.

5.6.2.1 Results comparison to ConVRP benchmark instances

These numerical experiments confirm the ability of our algorithm to solve related problems efficiently. We tested nine instances with our ALNS framework and compared the results with the ones achieved by Kovacs et al. (2014b) with the template ALNS (TALNS). Table 5.3 shows the gaps of the best solution reached for each instance (Best Gap) and the average solution obtained (Avg Gap). The total travel time plus service times (TT) for each instance are the comparison measure.

Table 5.3: Results of comparison to TALNS best solution by Kovacs et al. (2014b)

# Customers	Visit frequency	Best Gap (%)	Avg Gap (%)
50	50%	1.4%	1.5%
	70%	0.0%	0.1%
	90%	0.0%	0.7%
75	50%	0.2%	1.9%
	70%	0.1%	1.4%
	90%	-0.1%	1.7%
100	50%	0.3%	1.3%
	70%	0.2%	1.2%
	90%	0.1%	1.1%
Average	-	0.3%	1.2%

In all tests, the solution approach was able to converge to consistent solutions, in terms of driver and time-windows. The results show that the proposed ALNS reaches solutions close to the TALNS best, with the Best Gap close to 0%. Only for the instance with 50 customers and a visit frequency of 50% the best solution reached by the ALNS was 1.4% worse. Additionally, for the remaining instances the average gap lies below 1.5% for most instances.

These results demonstrate that our solution approach is able to find a consistent solution with a good traveling time. Please note that this approach was not developed to incorporate driver consistency.

5.6.2.2 ALNS operators analysis

Since we propose a new problem and developed an ALNS solution approach to solve it, an analysis of the operators considered is presented. In the solution approach, we propose weekly destroy operators that are adjusted from the literature to cope with our problem characteristics. In this section, we compare the application frequency (AF) of each operator, i.e. the proportion of iterations each operator is called. This comparison is made for the runs with the ConVRP and the grocery distribution data sets.

Figures 5.2 and 5.3 present the application frequency (AF) of the destroy operators (see Section 5.5.2) for both data sets, and Figures 5.4 and 5.5 present the same information, but for the repair operators (see Section 5.5.3).

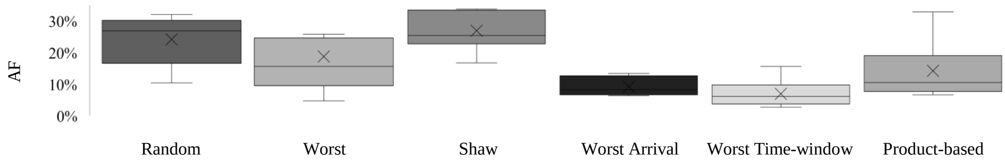


Figure 5.2: Application frequency of destroy operators for ConVRP data sets

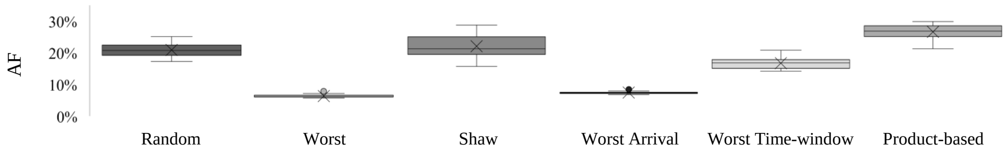


Figure 5.3: Application frequency of destroy operators for grocery distribution data sets

When running our solution approach for the ConVRP data sets, results indicate that all the operators proposed are called during the search, having the daily operators a share of 70%. Nevertheless, each of the weekly operators is also called around 10% of the times, being the product-based removal operator the most called due to the diversification that it allows.

The application frequency of these operators changes when the solution approach is used for the grocery distribution problem. Daily and weekly operators share 49%/51% application frequency, respectively. The results also indicate that the random and shaw removals maintain an application frequency around 20%, having the product-based removal operator an increase of calls. The worst and worst arrival removal are constantly called 7% of the times. The weekly operators included in our solution approach clearly help the search for the PTWA-MCVRP.

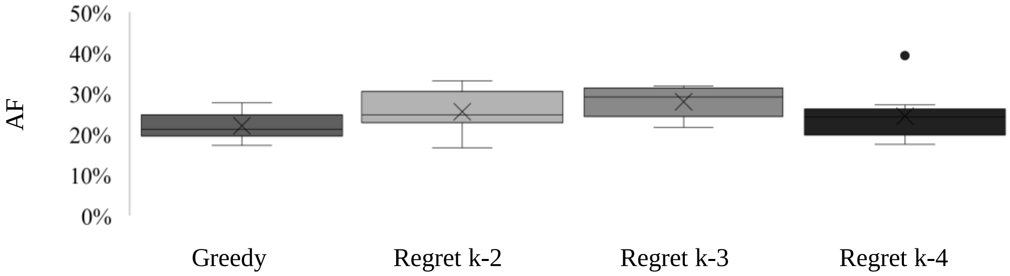


Figure 5.4: Application frequency of repair operators for ConVRP data sets

Figure 5.4 shows that for the ConVRP data set the four repair operators are similarly used by the ALNS, which is not the case for the grocery distribution data set. For the last, the greedy insertion operator has a much smaller application frequency than the remaining operators. Furthermore, the regret insertion with $k = 2$ seems to be the operator contributing the most with an average application frequency of 40% (see Figure 5.5).

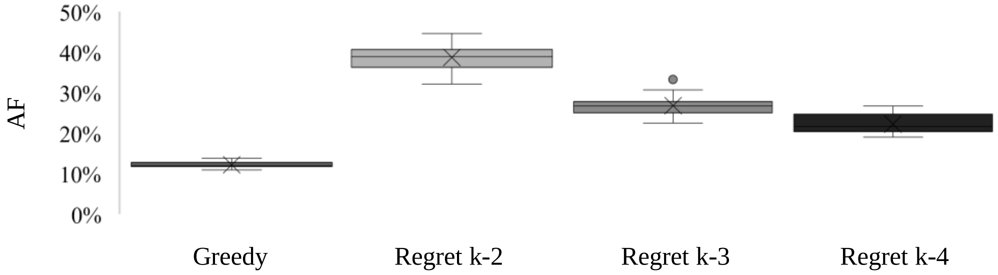


Figure 5.5: Application frequency of repair operators for grocery distribution data sets

5.6.3 Impact Analysis of Time-Window Assignment in Grocery Distribution

In this second part of our numerical experiments, we analyze the impact and benefits of introducing consistent deliveries and product-oriented time-window assignment in grocery distribution. We start with the simplest case of introducing time-windows assignment within the MCVRP with all time-windows available and analyze the impact of consistent deliveries. Afterwards, the complexity is gradually increased by restricting the set of time-windows for each pair customer-segment. In this way we can analyze the different characteristics of the problem.

5.6.3.1 Analysis of consistent time-windows deliveries

In this analysis we want to evaluate the impact of performing consistent deliveries on time, i.e. deliver each segment to stores within a unique time-window assigned. In this experiment, a set of eight time-windows with one hour width is considered, being the full set available to all customer-segment pairs. Two tests are performed for each instance: (1) the ALNS is run for each day individually, considering only the daily operators with no time restrictions, and (2) the full ALNS is run for the complete planning horizon. The first test only attempts to minimize the routing costs, including the loading, traveling and unloading costs. Considering the resulted arrival times of the orders, a time-window is assigned to each customer-segment pair in order to reduce the penalty costs for early or late deliveries. The expected total cost for the final solution is calculated by summing up the routing costs of each day with the penalty cost per customer-segment pair. This first test is named unrestricted planning and is based on the common practice. The second test aims at minimizing routing and penalty costs while ensuring consistent deliveries. The final solution provides a consistent time-window assignment and therefore the test is referred to as consistent delivery planning. The deviations between the expected total cost for the consistent delivery planning and unrestricted planning are presented in Figure 5.6. Three cost deviations are shown related to the comparison between the best, average and worst solutions found in each test, for each instance. All tests provided consistent solutions.

The results show that consistent delivery planning enables better overall solutions than the unrestricted planning just focused on the routing. The best consistent solution is able to improve the unrestricted best solution by around 0.7%. Furthermore, we see that the solutions generated with a consistent planning are always better than the unrestricted planning,

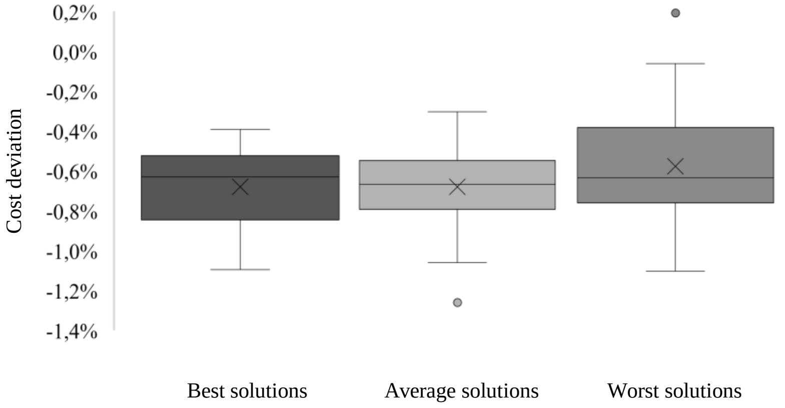


Figure 5.6: Results comparison of consistent delivery planning and unrestricted planning

with the cost deviation of the worst solution having an average improvement of 0.5%.

Although the consistent planning provides solutions with a better overall cost, the cost deviation between the two plannings is small (below 1%). However, the solutions are very different. The routing cost of the consistent planning is between 1.1% and 1.7% higher than that for the unrestricted planning, which is compensated by a 68% to 76% improvement of the overall penalty cost. A further analysis of the delivery time of each order for both plans shows that the consistent planning originates more on-time deliveries. Table 5.4 presents for both plans the average percentage of orders delivered "X" minutes outside the time-window assigned bounds.

Table 5.4: Average percentage of orders delivered "X" minutes earlier or later than the time-window assigned bounds

Interval of time (X in minutes)	Unrestricted planning	Consistent planning
0	69%	84%
0 - 10	9%	8%
10 - 30	11%	6%
30 - 60	5%	2%
60 - 120	3%	1%
120 - 240	3%	1%
> 240	2%	0%

From the results of Table 5.4, we see that the consistent planning reduces the amount of deliveries performed outside the bounds of the time-window or with a very small deviation. While the unrestricted planning contains 7% of the orders delivered with at least one hour deviation, the consistent planning reduces this percentage to 3%. Furthermore, note that the time deviations from the time-window bounds are penalized in the solution overall cost by 0.17 CU per minute, covering the working cost of a store employee. However, in practice for some of the deliveries the costs can be much higher as spoilage or stock outs situations can occur, for example.

5.6.3.2 Analysis of product-oriented time-window assignment

Having discussed the influence of consistent delivery planning, we now analyze the impact of defining a product-oriented time-window assignment. In the previous tests, all pairs customer-segment had the full set of time-windows available for assignment. However, in practice the stores might prefer to receive some segments in a more restricted set of time-windows, as it was described in Section 3.3. Therefore, we tested three different scenarios:

1st scenario (Fresh TW): Only a limited number of time-windows can be used for one of the segments (representing the fresh products), having the remaining segments the full time-window set available. As the fresh products have usually to be delivered at the beginning of the day, and we assume departures at time 0, the set of time-windows available for the fresh segment is set to the three earliest from the overall set.

2nd and 3rd scenarios: Other segments might have additional restrictions in different stores, depending on their operations. Therefore, we used two random sets of time-windows for the remaining segments. In the second scenario, named *Random TW (4-8)*, it is randomly selected the number and time-windows available, between four and eight, for each customer-segment pair. For the third scenario (*Random TW (4)*), the number of time-windows available is fixed to four, being the time-windows randomly selected. In both scenarios, the set of time-windows available for the fresh segment is the same of the first scenario (*Fresh TW*).

The solutions obtained for the three scenarios are compared with the solutions from the consistent delivery planning with all time-windows available (obtained in Section 5.6.3.1). The results are presented in Figure 5.7. Once again, the best, average and worst solutions reached by each of the tests for each instance are compared.

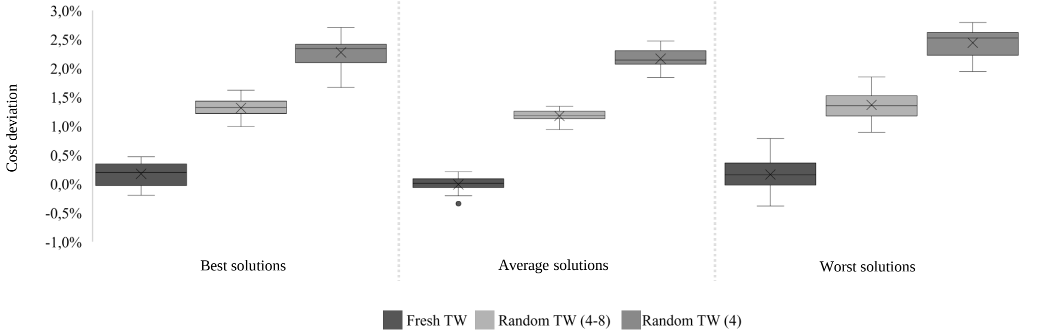


Figure 5.7: Results comparison of consistent delivery planning with distinct sets of time-windows available

Naturally, these results show that restricting the set of time-windows available for each pair customer-segment increases the solution overall cost. The deviations between the best, average and worst solutions of each scenario are very similar. Analyzing the cost deviation of the best solutions for each scenario, we see that the *Fresh TW* scenario originates a small increase in costs (below 1%) in the solution cost compared with the full time-window set available. This deviation reaches higher levels when all segments have random time-windows available for assignment.

A further analysis of the two cost contributions (routing and penalty), indicates that the routing cost is very similar between all scenarios, with an average deviation below 0.4%, pointing the penalty cost as the main driver for the cost increase. Similarly to the previous section, Table 5.5 presents for each scenario the average percentage of orders delivered "X" minutes outside the time-window assigned bounds. The results from the previous unrestricted planning and consistent delivery planning analysis are also presented for comparison. The average number of time-windows assigned per customer are presented for all scenarios in the bottom line.

Table 5.5: Percentage of orders "X" minutes earlier or later than the time-window assigned in each scenario

Interval of time (X in minutes)	Unrestricted	All TW	Fresh TW	Random TW (4-8)	Random TW (4)
0	69%	84%	82%	72%	65%
0-10	9%	8%	9%	10%	10%
10-30	11%	6%	6%	11%	12%
30-60	5%	2%	2%	5%	8%
60-120	3%	1%	1%	2%	4%
120-240	3%	1%	1%	1%	1%
>240	2%	0%	0%	0%	0%
Avg # TW per customer	1.32	1.32	1.42	1.92	2.22

The results show a reduction of on-time deliveries from the consistent scenario with all time-windows available to the more restricted scenarios. We can see that when all time-windows are available for all customer-segment pairs, the best solutions try to assign the same time-window to the full range of products, having both the unrestricted and consistent planning scenarios an average of 1.32 time-windows assigned to each customer. As the scenarios constrain the time-window set, the average number of time-windows assigned per customer increases. This would lead to separate deliveries which would increase the routing cost. However, it seems that by reducing the on-time deliveries, we maintain similar routing cost, having better overall costs.

5.7. Conclusion

This work extended the research on multi-compartment vehicle routing problems (MCVRPs) by tackling a multi-period setting with a product-oriented time-window assignment. The resulting PTWA-MCVRP was studied for the grocery distribution application, which has particular characteristics due to the multiple products it distributes with distinct temperature requirements. The aim of the PTWA-MCVRP is to define a unique time-window that should be used consistently throughout the planning horizon for each type of product of a store, taking into account the possibility of delivering the full product range jointly or separated with the use of MCVs. However, in practice the time-windows are not hard constraints and, therefore, deliveries outside the time-windows bounds are allowed with a negative impact for the in-store operations. Hence, the objective of the problem proposed is to minimize the routing costs inherent to the use of MCVs and the penalty cost related with missing the time-windows assigned.

An ALNS framework was designed in this paper to cope with the characteristics of the PTWA-MCVRP, combining daily and weekly operators to tackle the different problem decisions. The daily operators focus on a particular day and try to improve the routing decisions of the problem, while the weekly operators have a broader scope, aligning the time-window assignment decisions across all days.

The algorithm was tested on benchmark instances for the ConVRP, which is closely related with our problem due to the arrival time consistency constraint, and generated instances based on a grocery distribution problem. The effectiveness and efficiency of the ALNS framework proposed was validated as the solution approach was able to converge to consistent solutions close to the best solutions of the ConVRP. Furthermore, an analysis of the application frequency of the operators used showed that the PTWA-MCVRP requires a high search diversification, and that both daily and weekly destroy operators help guiding the search to better solutions.

An impact analysis of time-window assignment in grocery distribution was also conducted. At a first stage, we showed that performing a consistent delivery planning provides better overall solutions than an unrestricted planning, just focused on the routing costs. Although the deviation cost achieved between both plans was small (0.7% average improvement), it was shown that there was a significant difference on the percentage of orders delivered outside the time-window bounds, which could lead to higher costs due to spoilage or stock-outs situations. Finally, the implication of using a product-oriented time-window assignment was analyzed by restricting the number of time-windows available for the assignment to the different products. We concluded that if all products have the same time-windows available, most of the stores will receive the full range of products within the same time-window. Therefore, as we restrict the set of time-windows the number of time-windows used per customer increases, as well as the overall solution cost. This last effect is originated by the increase in the percentage of orders delivered outside the time-window bounds, indicating that it is less costly to miss the time-window than changing the routing.

Further extensions of this work can be made by considering lower and upper bounds for time-windows violation to prevent excessive penalties, as proposed by Ioannou et al. (2003). These bounds can be defined per customer and product, differentiating the cases that would not be so much affected by the situation. For instance, stores with small backrooms would require more on-time deliveries than the others. As we assumed departures from depot at time zero, a logic extension is to consider different departure times for the vehicles, as already considered by Kovacs et al. (2015b) for ConVRP. Additionally, considerations regarding DC docks capacity and fleet size could be included. Not all vehicles can departure at the same time due to loading docks capacity restrictions. Moreover, the fleet size is dependent on the number of simultaneous deliveries, so having different departure times allows for a smaller fleet. The development of an exact approach, such as a branch-and-price, would also be a future research direction in order to achieve near optimal solutions for comparison and evaluate in more detail the ALNS performance.

Acknowledgements The research was financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The first author was also supported by Grant SFRH/BD/102013/2014 from FCT. The second author was founded by the German Ministry of Education and Research and the Hanns Seidel Foundation. The initialization of this international research cooperation was financially supported by the Bavarian Research Alliance.

Bibliography

Abdulkader, M. M., Gajpal, Y., and ElMekkawy, T. Y. (2015). Hybridized ant colony algorithm for the multi compartment vehicle routing problem. *Applied Soft Computing*, 37:196–203.

Avella, P., Boccia, M., and Sforza, A. (2004). Solving a fuel delivery problem by heuristic and exact approaches. *European Journal of Operational Research*, 152(1):170–179.

Belhaiza, S., Hansen, P., and Laporte, G. (2014). A hybrid variable neighborhood tabu search heuristic for the vehicle routing problem with multiple time windows. *Computers & Operations Research*, 52:269–281.

Chajakis, E. D. and Guignard, M. (2003). Scheduling deliveries in vehicles with multiple compartments. *Journal of Global Optimization*, 26(1):43–78.

Christofides, N. and Eilon, S. (1969). An algorithm for the vehicle-dispatching problem. *OR*, 20(3):309–318.

Clarke, G. and Wright, J. W. (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research*, 12(4):568–581.

Coelho, L. C. and Laporte, G. (2015). Classification, models and exact algorithms for multi-compartment delivery problems. *European Journal of Operational Research*, 242(3):854–864.

Cornillier, F., Boctor, F. F., Laporte, G., and Renaud, J. (2008). A heuristic for the multi-period petrol station replenishment problem. *European Journal of Operational Research*, 191(2):295–305.

Derigs, U., Gottlieb, J., Kalkoff, J., Piesche, M., Rothlauf, F., and Vogel, U. (2011). Vehicle routing with compartments: Applications, modelling and heuristics. *OR Spectrum*, 33:885–914.

El Fallahi, A., Prins, C., and Calvo, R. W. (2008). A memetic algorithm and a tabu search for the multi-compartment vehicle routing problem. *Computers & Operations Research*, 35(5):1725–1741.

Feillet, D., Garaix, T., Lehuédé, F., Péton, O., and Quadri, D. (2014). A new consistent vehicle routing problem for the transportation of people with disabilities. *Networks*, 63(3):211–224.

Golden, B. L., Raghavan, S., and Wasil, E. A. (2008). *The Vehicle Routing Problem: Latest Advances and New Challenges*. Operations Research/Computer Science Interfaces Series. Springer.

Groër, C., Golden, B., and Wasil, E. (2009). The consistent vehicle routing problem. *Manufacturing & service operations management*, 11(4):630–643.

Henke, T., Speranza, M. G., and Wäscher, G. (2015). The multi-compartment vehicle routing problem with flexible compartment sizes. *European Journal of Operational Research*, 246(3):730–743.

Holzapfel, A., Hübner, A., Kuhn, H., and Sternbeck, M. G. (2016). Delivery pattern and transportation planning in grocery retailing. *European Journal of Operational Research*, 252(1):54–68.

Hübner, A. and Ostermeier, M. (2018). A multi-compartment vehicle routing problem with loading and unloading costs. *Transportation Science, forthcoming*.

Hübner, A. H., Kuhn, H., and Sternbeck, M. G. (2013). Demand and supply chain planning in grocery retail: An operations planning framework. *International Journal of Retail & Distribution Management*, 41(7):512–530.

Ioannou, G., Kritikos, M., and Prastacos, G. (2003). A problem generator-solver heuristic for vehicle routing with soft time windows. *Omega*, 31(1):41–53.

Jabali, O., Leus, R., Van Woensel, T., and De Kok, T. (2015). Self-imposed time windows in vehicle routing problems. *OR Spectrum*, 37(2):331–352.

Kaabi, H. and Jabeur, K. (2015). Hybrid algorithm for solving the multi-compartment vehicle routing problem with time windows and profit. In *Informatics in Control, Automation and Robotics (ICINCO), 2015 12th International Conference on*, volume 1, pages 324–329. IEEE.

Kabcome, P. and Mouktonglang, T. (2015). Vehicle routing problem for multiple product types, compartments, and trips with soft time windows. *International Journal of Mathematics and Mathematical Sciences*, 2015.

Koskosidis, Y. A., Powell, W. B., and Solomon, M. M. (1992). An optimization-based heuristic for vehicle routing and scheduling with soft time window constraints. *Transportation science*, 26(2):69–85.

Kovacs, A. A., Golden, B. L., Hartl, R. F., and Parragh, S. N. (2014a). Vehicle routing problems in which consistency considerations are important: A survey. *Networks*, 64(3):192–213.

Kovacs, A. A., Golden, B. L., Hartl, R. F., and Parragh, S. N. (2015a). The generalized consistent vehicle routing problem. *Transportation Science*, 49(4):796–816.

Kovacs, A. A., Parragh, S. N., and Hartl, R. F. (2014b). A template-based adaptive large neighborhood search for the consistent vehicle routing problem. *Networks*, 63(1):60–81.

Kovacs, A. A., Parragh, S. N., and Hartl, R. F. (2015b). The multi-objective generalized consistent vehicle routing problem. *European Journal of Operational Research*, 247(2):441–458.

Kuhn, H. and Sternbeck, M. G. (2013). Integrative retail logistics: An exploratory study. *Operations Management Research*, 6(1-2):2–18.

Lian, K., Milburn, A. B., and Rardin, R. L. (2016). An improved multi-directional local search algorithm for the multi-objective consistent vehicle routing problem. *IIE Transactions*, 48(10):975–992.

Martins, S., Amorim, P., and Almada-Lobo, B. (2017). Delivery mode planning for distribution to brick-and-mortar retail stores: discussion and literature review. *Flexible Services and Manufacturing Journal*, pages 1–28.

Muyldermans, L. and Pang, G. (2010). On the benefits of co-collection: Experiments with a multi-compartment vehicle routing algorithm. *European Journal of Operational Research*, 206(1):93–103.

Pires, M., Pratas, J., Liz, J., and Amorim, P. (2017). A framework for designing backroom areas in grocery stores. *International Journal of Retail & Distribution Management*, 45(3):230–252.

Reed, M., Yiannakou, A., and Evering, R. (2014). An ant colony algorithm for the multi-compartment vehicle routing problem. *Applied Soft Computing*, 15:169–176.

Ropke, S. and Pisinger, D. (2006). An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. *Transportation Science*, 40(4):455–472.

Shaw, P. (1997). A new local search algorithm providing high quality solutions to vehicle routing problems. *APES Group, Dept of Computer Science, University of Strathclyde, Glasgow, Scotland, UK.*

Spliet, R., Dabia, S., and van Woensel, T. (2017). The time window assignment vehicle routing problem with time-dependent travel times. *Transportation Science*.

Spliet, R. and Desaulniers, G. (2015). The discrete time window assignment vehicle routing problem. *European Journal of Operational Research*, 244(2):379–391.

Spliet, R. and Gabor, A. F. (2014). The time window assignment vehicle routing problem. *Transportation Science*, 49(4):721–731.

Subramanyam, A. and Gounaris, C. E. (2016). A branch-and-cut framework for the consistent traveling salesman problem. *European Journal of Operational Research*, 248(2):384–395.

Subramanyam, A. and Gounaris, C. E. (2017). Strategic allocation of time windows in vehicle routing problems under uncertainty.

Sungur, I., Ren, Y., Ordóñez, F., Dessouky, M., and Zhong, H. (2010). A model and algorithm for the courier delivery problem with uncertainty. *Transportation Science*, 44(2):193–205.

Tarantilis, C. D., Stavropoulou, F., and Repoussis, P. P. (2012). A template-based tabu search algorithm for the consistent vehicle routing problem. *Expert Systems with Applications*, 39(4):4233–4239.

Toth, P. and Vigo, D. (2014). *Vehicle Routing: Problems, Methods, and Applications, Second Edition*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics.

van Zelst, S., van Donselaar, K., van Woensel, T., Broekmeulen, R., and Fransoo, J. (2009). Logistics drivers for shelf stacking in grocery retail stores: Potential for efficiency improvement. *International Journal of Production Economics*, 121(2):620 – 632.