Stochastic Inventory Routing with Time-based Shipment Consolidation

Danja R. Sonntag
Business School, University of Mannheim, Germany
dsonntag@mail.uni-mannheim.de

Albert H. Schrotenboer
Operations, Planning, Accounting, and Control Group, School of Engineering, Eindhoven University of Technology, Netherlands
a.h.schrotenboer@tue.nl

Gudrun P. Kiesmüller
TUM School of Management, TUM Campus Heilbronn, Technical University of Munich, Germany
gudrun.kiesmueller@tum.de

Inspired by the retail industry, we consider a stochastic inventory routing problem where retailers are replenished from a central warehouse using a time-based shipment consolidation policy. Such a time-based dispatching policy, where retailers facing stochastic demand are repetitively replenished at fixed times, is essential in practice. It allows for easy incorporation with dependent up- and downstream planning problems such as personal staffing and warehouse operations, and has become a standard part of transportation contracts. We provide a new chance-constrained model that determines an optimal clustering of retailers in groups, their associated routing and shipment interval, and each retailers' optimal inventory level. A newly developed branch-price-and-cut algorithm solves our model to optimality. Its efficiency comes from a tailored labeling algorithm for solving the pricing problem that relies, among others, on an optimality pruning criterion based on the approximate solution of a 0,1-knapsack problem. Computational experiments show that our exact method can solve instances of up to 60 retailers to optimality. Besides, we accommodate practitioners by providing a fast heuristic that provides excellent solutions with an optimality gap of less than 1%. Finally, we show that incorporating uncertainty already in the planning process is essential for stochastic inventory routing with time-based shipment consolidation, as it results in overall cost-savings of 7.7% compared to the current state-of-the-art.

Key words: transportation; stochastic inventory routing; inventory control; time-based shipment consolidation; branch-price-and-cut

1. Introduction

How to optimally replenish inventory at geographically dispersed retailers from a central warehouse is a fundamental question at the interface of inventory management and transportation. At this interface, operations managers need decision support to coordinate when shipments are sent to retailers, how much will be sent to each retailer, and how these shipments are consolidated in multiple vehicle routes to minimize inventory holding and transportation costs. Practical applications by Duffy (2004), Gaur and Fisher (2004), Coelho and Laporte (2014), and Van Anholt et al. (2016) have shown that an integrated view on inventory management and transportation leads to significant cost savings in practice. In particular, Van Anholt et al. (2016) estimated the expected business cost savings at about €10.1 million per year for a Dutch ATM operator and Blumenfeld et al. (1987) reported cost savings of $2.9 million a year at General Motors. The most-well known application of integrated inventory and transportation management has been presented

by Gaur and Fisher (2004), who describe the planning problem at Albert Heijn, a leading supermarket chain in the Netherlands. Albert Heijn faces the problem of replenishing all stores with stochastic demand several times during a week. Gaur and Fisher (2004) reported cost-savings of 4% during the first year by implementing a corresponding replenishment system.

For planning the retailer's inventory replenishments, it is necessary that shipments leave the warehouse and arrive at the retailers at fixed and predetermined points in time to allow for coordination with dependent up- and downstream planning processes such as material handling and staffing. The necessity for such time-based dispatching policies has been exemplified by case studies from Holzapfel et al. (2016) at a major European grocery retailer, Stenius et al. (2016) at a European metal sheet production company, and Campelo et al. (2019) for medicine distribution in the pharmaceutical sector. Time-based dispatching policies are also beneficial for logistics service providers. Namely, drivers can be utilized more efficiently because they become familiar with their routes, and, on a higher level, logistics providers can better plan the vehicles and personnel required. Consequently, companies can enter cheaper long-term contracts with logistics service providers (Gaur and Fisher 2004).

The resulting inventory replenishment strategy, where shipments are dispatched from the warehouse at fixed points in time to replenish inventory at predetermined groups of retailers, is known in the literature as a time-based shipment consolidation policy (see, e.g., Higginson and Bookbinder 1995, Marklund 2011, Ülkü and Bookbinder 2012, Stenius et al. 2016, Johansson et al. 2020). Optimizing inventory replenishments under time-based shipment consolidation is challenging because retailers face stochastic demand. In such a case, each replenishment takes place at fixed times while the amount to be shipped is stochastic. In the literature, the joint optimization of inventory replenishments and transportation is known as the inventory routing problem (Coelho et al. 2014b). However, researchers have been studying time-based shipment consolidation policies mainly when retailer demand is deterministic. To the best of our knowledge, the few studies on stochastic inventory routing problems with time-based shipment consolidation, which we discuss in detail in the literature review, all present heuristic solution approaches to this highly complex problem.

To solve our chance-constrained optimization model, we transform it into an integer program for which we present a tailored branch-price-and-cut algorithm. To solve the associated pricing problem, we introduce a new pruning criterion based on the approximate solution of a 0,1-knapsack problem. The approach appears to be efficient because it solves instances of up to 60 retailers to optimality. Notably, this is the first exact solution approach to the stochastic inventory routing problem that considers the joint optimization of retailer clusters, their corresponding shipment intervals, the routing within each cluster, and the base stock levels at all retailers. Moreover, we provide tailored constructive heuristics that also serve as input to the branch-price-and-cut algorithm and a hybrid heuristic that solves all considered problem instances with an optimality gap of typically less than 1% in negligible time. We show that including the uncertainty of retailer demand in the planning process is crucial for controlling costs and capacity utilization. Relying on expected demand

and buffer space in trucks yields an average cost increase of 7.7% and insufficient truck capacity in about 95% of the considered instances, despite that we re-optimize shipment intervals to prevent this to the best extent possible. This shows that our new approach, not suffering from this drawback, determines structurally different replenishment schemes that can better mitigate retailer demand uncertainty by making more efficient use of the available truck capacity.

In summary, the contributions of this paper are as follows: (1) We develop a tailored branch-price-and-cut algorithm to provide optimal solutions to this highly complex problem. (2) We provide tailored constructive heuristics and a hybrid heuristic that exhibit excellent performance in negligible time. (3) We show the importance of including uncertainty in jointly optimizing inventory replenishment and transportation decisions under time-based shipment consolidation policies.

The remainder of the paper is organized as follows. In Section 2, we describe the problem in detail. In Section 3, we review the relevant literature from the fields of inventory management, capacitated clustering, and inventory routing, and emphasize the research gap that this paper is closing. Based on the problem statement in Section 2, we formulate a new, integer chance-constrained optimization model in Section 4. In Section 5, we present an exact solution approach and in Section 5.3 several heuristics. In Section 6, we provide numerical results showing the performance of all solution approaches, the importance of considering uncertainty, and insights into the structure of the optimal solution. We conclude the paper and provide an outlook on future research opportunities in Section 7.

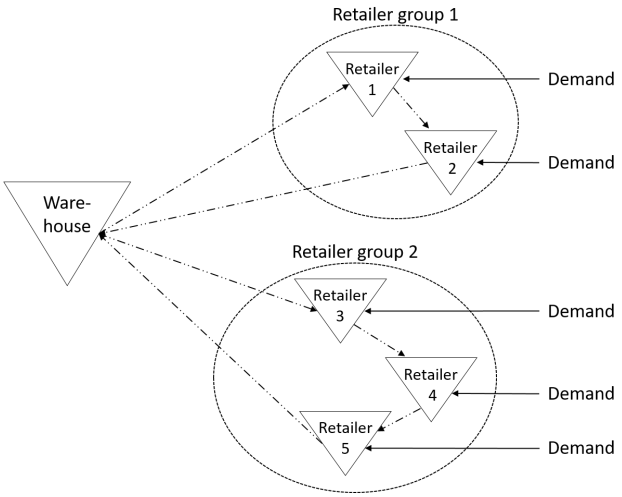
2. Problem Statement

We consider a periodic, infinite-horizon, single-warehouse multiple-retailer distribution inventory system as depicted in Figure 1. In the following, we provide a high-level description of our system and postpone the formulation of our optimization problem to Section 4, where we also present an overview of the main notation used (see Table 2).

The system consists of a set of retailers $\mathcal{N} = \{1,\dots,N\}$, where each retailer $i \in \mathcal{N}$ faces stationary but stochastic period demand D_i for a single item with known distribution and known expectation $\mathbb{E}[D_i]$ and variance $\mathrm{VAR}[D_i]$. Note that a single item model is suitable even in a retail setting with thousands of products, if the demand is aggregated to, e.g., units of volume as it is done by Albert Heijn (Gaur and Fisher 2004). Demand occurring at each retailer is either satisfied from stock on hand immediately or backlogged in case of stock-outs. Backorders are satisfied as soon as possible on a first-come, first-served basis once a new delivery arrives at the retailer. Each retailer i imposes a target service level α_i^* that refers to the non-stock-out probability at retailer i, which ensures a high level of customer satisfaction.

To replenish stock from the central warehouse, each retailer uses a periodic replenishment policy. More precisely, each retailer i places orders according to an (R_i, S_i)-policy comprising a review period R_i and a base-stock level S_i. Under an (R_i, S_i)-policy, retailer i places an order every R_i periods such that the inventory

Figure 1 Single-Warehouse Multiple-Retailer Distribution Inventory System with Shipment Consolidation



position after ordering is equal to the base-stock level S_i. We denote the vector of base-stock levels for all retailers $i \in \mathcal{N}$ by $\boldsymbol{S} = (S_1, S_2, \ldots, S_N)$. The central warehouse itself has ample supply and distributes the items ordered using a homogeneous fleet with an unlimited number of vehicles, each with capacity Q.

We consolidate all N retailers to K ($1 \leq K \leq N$) mutually exclusive and collectively exhaustive retailer groups. Note that the number of retailer groups K, as well as the allocation of the retailers to retailer groups, is part of our optimization problem. For readability, we denote the set of retailer groups by $\mathcal{K} = \{1, \ldots, K\}$. By using time-based shipment consolidation, all goods shipped to retail group $k \in \mathcal{K}$ are dispatched periodically with a constant shipment interval $T_k \in \mathbb{N}_{\geq 1}$, which refers to the time between two consecutive replenishments. Note that under time-based shipment consolidation, shipments leave the central warehouse independently of the vehicle's current utilization but in fixed time intervals, which leads to a fixed delivery pattern. For all retailer groups K collectively, we denote the vector of shipment intervals by $\boldsymbol{T} = (T_1, T_2, \ldots, T_K)$ and assume that the shipment interval is constant within each retailer group. This means that whenever a truck replenishes a retailer group, all retailers belonging to that group will be visited independently of the requested quantities.

The number of units that comprise a replenishment to retailer group k depends on the retailers' orders according to their current inventory situation. Because shipments to retailer group k leave the warehouse every T_k periods, all retailers i belonging to that retailer group place orders every T_k periods just before the vehicle departs from the central warehouse, and hence $R_i = T_k$. Thus, orders are placed by all retailers i belonging to retailer group k according to a (T_k, S_i)-policy. To raise the inventory position after ordering to S_i, the order quantity of each retailer belonging to group k equals the demand during the last T_k periods. Because demand is stochastic, order quantities are stochastic as well, which can lead to a situation where the capacity of a vehicle is not sufficient to transport all ordered units. In such a situation, an emergency shipment is performed by an external service provider to deliver excess units to the corresponding retailers at a fixed cost per unit. We assume that regular and emergency shipments arrive at the retailers immediately after departure

from the warehouse, which means that we neglect transportation, order processing, and material handling times, because they are usually very short compared to the length of the shipment intervals. However, these processing times can easily be included in the model by increasing the lead time from the warehouse to the corresponding retailers.

Henceforth, the sequence of events in each period is as follows: At the beginning of every period, shipments arrive at all retailers belonging to a retailer group k that receive a replenishment according to the group-specific shipment interval T_k. Following this, demand occurs at the retailers and is either satisfied from stock or backlogged. At the end of each period, retailers receiving an order in the next period place a new order based on the current inventory position, which comprises the stock-on-hand, backlogs, and all outstanding orders, and all costs are reported.

The aim is to find a cost-minimizing set of retailer groups, their associated shipment intervals, and the base-stock level at each retailer. The expected costs per period are composed of the sum of fixed and variable transportation costs, emergency shipment costs, and inventory holding costs for cycle and safety stock. Cycle stock is defined as the expected stock-on-hand required to satisfy the mean demand per replenishment cycle, whereas safety stock equals the expected stock-on-hand just before a replenishment arrives. A detailed description of the mathematical formulation follows in Section 4, after an overview of the relevant literature related to our problem, which is presented in the next section.

3. Literature Overview

Our study interfaces with scientific work in the fields of inventory management in single- and two-echelon inventory systems and inventory routing problems. Because of the large number of excellent papers in all of these fields, the following overview does not claim completeness, but instead emphasizes the research gap our paper is closing.

3.1. Inventory Management

Shipment consolidation in a single-warehouse multiple-retailer setting with stochastic demand has been studied extensively. We focus on studies that take an integrated view of consolidation and inventory management, and refer to Çetinkaya and Bookbinder (2003) and Mutlu et al. (2010) for an overview of studies that focus solely on consolidation policies, i.e., time-based, quantity-based, and time- and quantity-based shipment consolidation. As this paper considers a single-echelon system, in which ample supply is available at the central warehouse, we first review the existing literature in this field. Then, we briefly discuss the literature on two-echelon systems with limited instead of ample supply at the central warehouse.

In the stream of single-echelon models that integrate consolidation and inventory management, the grouping of retail stores is given, and the shipping capacity is unrestricted. Çetinkaya and Lee (2000) were the first to consider stochastic demand, i.e., a Poisson process, together with a time-based dispatching policy, for which they presented a heuristic solution approach. An exact solution procedure and an improved heuristic

procedure were presented by Axsäter (2001). A less practical alternative to time-based dispatching policies is represented by quantity-based or hybrid dispatching policies, which were extensively compared by Chen et al. (2005) and Çetinkaya et al. (2006), who found that quantity-based policies are, in terms of costs, at least as good as time-based policies. Cetinkaya et al. (2008) presented optimal quantity-based dispatching policies for the case in which both order arrivals and order sizes are stochastic.

Owing to the complexity of the problem, the literature on shipment consolidation in the context of two-echelon distribution inventory systems with stochastic demand is limited. It has focused mainly on time-based dispatching policies because of their practicability, for instance, in the retail industry (for quantity-based consolidation policies, we refer to Kiesmüller and De Kok 2005). Considering time-based dispatching policies, Marklund (2011) and Stenius et al. (2016) presented exact approaches to minimize the total system-wide costs under Poisson and compound Poisson demand, respectively. Johansson et al. (2020) used the same model as Stenius et al. (2016) and developed computationally attractive heuristics solving larger problem instances.

All papers on single- and two-echelon inventory systems have in common that they consider an unlimited fleet of vehicles with unlimited capacity. Furthermore, geographically close retailers are replenished according to externally given retailer groups. Thus, these papers focus solely on determining shipment intervals $\boldsymbol{T}$ and base-stock levels $\boldsymbol{S}$ at the retailers, and neglect the clustering and routing as well as the interaction between all these decisions. However, the importance and complexity of determining retailer groups have recently been recognized. For instance, Stenius et al. (2018) wrote that the "configuration of retailer groups can affect the performance of the system."

3.2. Inventory Routing Problem

The basic version of the inventory routing problem (IRP) considers one supplier and several geographically dispersed retailers. Each period, the supplier has to replenish stock at a subset of retailers using capacitated vehicles. The objective is to minimize the total inventory distribution cost while meeting the demand at each retailer (Coelho et al. 2014b).

For the replenishment of stock at the retailers, one distinguishes between static and dynamic allocation policies (Kumar et al. 1995). Under static allocation policies, the allocation of items on a vehicle to retailers is made for all retailers before the vehicle leaves the central warehouse. By contrast, under dynamic allocation policies, the allocation of vehicle inventory is postponed to the moment when the vehicle reaches each retailer. Because the solution procedure under dynamic allocation is very different from that for static allocation, which we consider in this paper, we do not review the literature on dynamic allocation but instead refer to, for instance, Trudeau and Dror (1992), Reiman et al. (1999), Jaillet et al. (2002), Schwarz et al. (2006), Huang and Lin (2010), and Ghiami et al. (2019) for papers considering this kind of problem.

Focusing on static allocation policies, the problem we consider in this paper differs from classical IRPs in the following ways:

1. Many stochastic IRP papers consider that in each period, the planners are concerned with deciding which customers to replenish and how much to deliver. In contrast to this, we consider a planning problem with fixed routes to a subset of retailers that are replenished repeatedly in fixed time intervals.

2. Those IRP papers focusing on fixed routes—aggregated in the literature under the topic *Cyclic Inventory Routing Problem (CIRP)*—mostly assume that demand rates are deterministic, whereas we consider stochastic demand, which makes it necessary to

(i) optimize safety stock at each retailer, which depends on the replenishment frequency;

(ii) deal with situations where not all ordered units fit on the capacitated vehicle by, for instance, considering emergency shipments.

3. Because of the complexity of the problem, most (C)IRP papers derive heuristic instead of exact algorithms. By contrast, we present an exact algorithm and several heuristics to solve the stochastic (cyclic) inventory routing problem.

Table 1 presents a summary of the IRP literature, including the most relevant papers as well as those most closely related to our problem. In this overview, we emphasize the research gap between this paper and previous work in this extensively studied field of research. For extended literature reviews, we refer interested readers to the excellent papers by Kleywegt et al. (2002), Moin and Salhi (2007), Andersson et al. (2010), and Coelho et al. (2014b). The column headings in Table 1 represent some key problem characteristics:

1. The time between two consecutive shipments—the shipment interval—can be either fixed or flexible over a particular planning horizon.

2. The demand per period at the retailers can be deterministic or stochastic.

3. This column states whether or not emergency shipments are being considered.

4. The next four columns refer to the decisions: Determining clusters of retailers that are always replenished jointly, determining fixed shipment frequencies instead of daily replenishment decisions, determining the routing between retailers and the warehouse, and determining inventory holding costs.

5. This paper has a special focus on the inventory component. Therefore, we identify papers in the IRP literature that optimize cycle stock and/or safety stock at the retailers.

6. Solution approaches can be exact or heuristic.

Table 1 divides the IRP literature into two groups. The first group focuses on cyclic planning problems under mainly deterministic demand, whereas the second group considers a planning problem with flexible shipment intervals under stochastic demand. Under flexible shipment intervals, the probability of stock-outs at the retailers determines whether or not a retailer is replenished in the current period. Flexible shipment intervals are motivated by practical applications in, for instance, the distribution of gases to customers (see, e.g., Trudeau and Dror 1992, Kleywegt et al. 2004). In such a setting, the customers "do not routinely call requesting replenishment nor are there regular pre-scheduled deliveries" (Berman and Larson 2001).

Table 1 Comparison of Our Contributions with Those of the Most Relevant Existing Studies

	1		2		3	4				5		6	
	Fixed intervals	Flexible intervals	Deterministic demand	Stochastic demand	Emergency shipments	Clustering	Shipment interval	Routing	Inventory	Optimize cycle stock	Optimize safety stock	Exact solution	Heuristic solution
Raa and Aouam (2021)	✓			✓	✓	✓	✓	✓	✓	✓	✓		✓
Malicki and Minner (2021)	✓			✓		✓	✓	✓	✓	✓	✓		✓
Aghezzaf (2008)	✓			(✓)		✓	✓	✓	✓	✓	✓		✓
Gaur and Fisher (2004)	✓		✓	(✓)	(✓)	✓	✓	✓					✓
Aghezzaf et al. (2012)	✓		✓			✓	✓	✓	✓	✓		✓	✓
Bertazzi et al. (2020)	✓		✓			✓	✓	✓	✓	✓		✓	✓
Diabat et al. (2021)	✓		✓			✓	✓	✓	✓	✓		✓	
Lefever et al. (2016)	✓		✓			✓	✓	✓	✓	✓		✓	
Bramel and Simchi-Levi (1995)	✓		✓			✓	✓	✓	✓	✓			✓
Chan et al. (1998)	✓		✓			✓	✓	✓	✓	✓			✓
Vansteenwegen and Mateo (2014)	✓		✓			✓	✓	✓	✓	✓			✓
Aghezzaf et al. (2006)	✓		✓			✓	✓	✓	✓	✓			✓
Zhao et al. (2008)	✓		✓			✓	✓	✓	✓	✓			✓
Raa and Dullaert (2017)	✓		✓			✓	✓	✓	✓	✓			✓
Chitsaz et al. (2016)	✓		✓			✓	✓	✓	✓	✓			✓
Ekici et al. (2015)	✓		✓			✓	✓	✓	✓	✓			✓
Raa and Aghezzaf (2009)	✓		✓			✓	✓	✓	✓	✓			✓
Burns et al. (1985)	✓		✓				✓		✓	✓		✓	
Coelho et al. (2014a)		✓		✓	✓			✓	(✓)				✓
Solyalı et al. (2012)		✓		✓				✓	✓			✓	
Adelman (2004)		✓		✓			✓	✓	✓	✓	✓		✓
Bertazzi et al. (2013)		✓		✓			✓	✓	✓	✓			✓
Hvattum and Løkketangen (2009)		✓		✓				✓	✓				✓
Juan et al. (2014)				✓				✓	✓	✓			✓
Kleywegt et al. (2002)		✓		✓			✓	✓	✓	✓	✓		✓
Kleywegt et al. (2004)		✓		✓			✓	✓	✓	✓	✓		✓
Bard et al. (1998)		✓		✓				✓					✓
Crama et al. (2018)		✓		✓				✓					✓
Campbell and Savelsbergh (2004)		✓	✓			✓		✓					✓
This paper	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Therefore, no coordination with dependent planning processes is necessary, and day-to-day planning of deliveries is reasonable resulting in a very different planning problem.

Focusing on fixed shipment intervals, the work by Raa and Aouam (2021) is most closely related to our problem. The authors use a similar model as in this paper, including stochastic demand and expedited

shipments in case of insufficient vehicle capacity, with the objective to minimize total average transportation and inventory related costs. Raa and Aouam (2021) present a heuristic approach to solve the problem under the assumption of normally distributed demand, whereas we present an exact algorithm under a more general gamma distributed demand structure, which allows for larger demand variability and skewed demand distributions. The second study in the field of stochastic cyclic IRPs is that of Malicki and Minner (2021), who consider a slightly different model with a finite planning horizon and formulate it as a cyclic lot sizing problem. To determine the replenishment frequencies and quantities, a multi-start adaptive search and an adaptive large neighborhood search heuristic are proposed. Despite the studies by Raa and Aouam (2021) and Malicki and Minner (2021), Aghezzaf (2008) and Gaur and Fisher (2004) are the only authors considering the inventory routing problem under periodic deliveries and stochastic demand. However, Aghezzaf (2008) and Gaur and Fisher (2004) use a decomposition approach and first solve a deterministic formulation of the problem, which is based on expectations of the random variables. In the second step the solution is adjusted using, e.g., simulations to hedge against the uncertainty of the demand.

Summarizing, there have been only a few studies considering fixed replenishment intervals for retailers facing stochastic customer demand and we are not aware of any study solving this complex planning problem exactly, which highlights the contribution of our work.

4. Model Formulation

In the following, we first formulate the mathematical model that can be used to solve our inventory routing problem with stochastic demand. We formulate it as a chance-constrained integer optimization problem. We then introduce an integer formulation by applying a Dantzig–Wolfe reformulation to the chance-constrained integer optimization problem. An overview of the notation is given in Table 2.

4.1. Mathematical Model

First, let $\mathscr{N}^0 := \mathscr{N} \cup \{0\}$, where $\{0\}$ represents the central warehouse and $\mathscr{N}$ the set of retailers. For readability, we refer to all $i \in \mathscr{N}^0$ as retailers. Our problem is then defined on the graph $G = (\mathscr{N}^0, \mathscr{A})$, where $\mathscr{A}$ is the complete edge set. We set the maximum number of retailer groups $K = N$ to ensure that we allow for all possible retailer groups in our model, because retailer groups are allowed to be empty. Recall the vectors with decision variables $\boldsymbol{T} = (T_1, \ldots, T_K)$, denoting the shipment intervals of the K retailer groups, and $\boldsymbol{S} = (S_1, \ldots, S_N)$, denoting the base-stock levels at each retailer. Furthermore, let $\boldsymbol{Y} = (y_{ijk})_{(N+1)\times(N+1)\times K}$ denote binary decision variables that equal 1 if retailer $j \in \mathscr{N}^0$ is directly visited after retailer $i \in \mathscr{N}^0$ in retailer group $k \in \mathscr{K}$, and equal 0 otherwise. Finally, for readability, we introduce $\boldsymbol{X} = (x_{ik})_{N\times K}$ as binary decision variables that equal 1 if retailer group k contains retailer $i \in \mathscr{N}$, and equal 0 otherwise. It holds that $\sum_{j\in\mathscr{N}^0} y_{ijk} = x_{ik}$ for all $i \in \mathscr{N}$.

Table 2 Overview of notation used

Sets:	
$\mathscr{N}$	set of retailers
$\mathscr{N}^0$	set including all retailers and the central warehouse
$\mathscr{K}$	set of retailer groups
$\mathscr{A}$	complete edge set
Indices:	
N	number of retailers
i, j	index for retailers
K	number of retailer groups
k	index for retailer groups
Parameters:	
D_i	random variable for the demand per period at retailer i
$\mathbb{E}[D_i]$	mean demand per period at retailer i
$\mathrm{VAR}[D_i]$	variance of the demand per period at retailer i
$D_i(T_k)$	random variable for the demand for retailer i during T_k periods
α_i^*	target service level at retailer i
γ^*	target probability that the vehicle's capacity for a replenishment is not exceeded
Q	capacity of all vehicles
$\mathrm{IL}_i^+(t, S_i)$	stock on hand at the end of period t at retailer i depending on the base-stock level S_i
G	graph
Decision variables:	
$\boldsymbol{S}$	vector of base-stock levels
S_i	base-stock level for retailer i
$\boldsymbol{T}$	vector of shipment intervals to all retailer groups
T_k	shipment interval for group k
R_i	review period for retailer i belonging to group k ($R_i = T_k$)
$\boldsymbol{Y}$	$(N \times N \times K)$ matrix containing binary decision variables y_{ijk}
y_{ijk}	binary decision variable that equals 1 if retailer j is visited after retailer i in group k, and 0 otherwise
$\boldsymbol{X}$	$(N \times K)$ matrix containing binary decision variables x_{ik}
x_{ik}	binary decision variable that equals 1 if retailer group k contains retailer i, and equal 0 otherwise
Cost components:	
TC	total expected cost per period
W	fixed shipment costs per shipment
c_{ij}	travel cost between retailer i and j
w_k	variable shipment cost per shipment to retailer group k
e	emergency shipment cost per emergency shipment to one retailer group
h	holding cost per unit and time unit

We first present the complete objective function. We then discuss the components of the objective function and the associated variables and parameters one-by-one. The objective function is formulated as

$$\mathrm{TC}(\boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{k \in \mathscr{K}} \frac{W + w_k(\boldsymbol{Y})}{T_k}$$

$$+\sum_{k\in\mathscr{K}}\mathbb{E}\left[\left(\sum_{i\in\mathscr{N}}x_{ik}D_i(T_k)-Q\right)^+\right]\frac{e}{T_k}$$
$$+h\sum_{i\in\mathscr{N}}\sum_{k\in\mathscr{K}}\frac{1}{T_k}\sum_{t=1}^{T_k}x_{ik}\mathbb{E}[\mathrm{IL}_i^+(t,S_i)]. \tag{1}$$

The first term in Equation (1) corresponds to the regular shipment costs per period for each retailer group k. These are composed of a fixed term W per replenishment and a variable component $w_k(\boldsymbol{Y})$, which corresponds to the variable transportation costs required to replenish all retailers in group k with a single vehicle tour. The variable costs are defined as

$$w_k(\boldsymbol{Y})=\sum_{i\in\mathscr{N}^0}\sum_{j\in\mathscr{N}^0}c_{ij}y_{ijk}\quad\forall k\in\mathscr{K}, \tag{2}$$

and depend on the routing related decision variables y_{ijk} and the travel cost c_{ij} between retailer i and j.

The second term in Equation (1) refers to the expected emergency shipment costs per period determined based on the expected number of units that exceed the vehicle's capacity Q for retailer group k. Here, $D_i(T_k)$ refers to the demand at retailer i during T_k periods and $(\cdot)^+$ is the positive-part operator, i.e., $(u)^+=\max\{0,u\}$. We assume that all excess items are delivered to the corresponding retailers by an external service provider for a fixed unit price e. Because emergency shipments can occur only in periods where regular shipments take place, emergency shipment costs to retailer group k are divided by the group-specific shipment interval T_k to obtain the expected emergency shipment costs per period. Therefore, the emergency shipment costs per period are not only affected by the composition of retailer groups but also the corresponding shipment interval, which emphasizes the difficulty of the planning problem because of an interrelation of all decision variables and the resulting cost terms.

The third term in Equation (1) accumulates the expected holding costs per period over all retailers. Holding costs for a single retailer i belonging to retailer group k are calculated by multiplying the unit holding cost parameter h with the expected stock on hand $\mathbb{E}[\mathrm{IL}_i^+(t,S_i)]$ at the end of each period t during the replenishment cycle of length T_k. Note that the expected stock on hand depends on the base-stock level S_i, which in turn is dependent on the length of the corresponding shipment interval and, therefore, the time between the delivery of two consecutive orders.

While minimizing the total expected costs per period TC, several constraints need to hold such that we can find a feasible solution to our problem. First, the retailers' target service levels α_i^* must be satisfied:

$$P(\mathrm{IL}_i^+(T_k,S_i)>0)\geq\alpha_i^*\quad\forall i\in\mathscr{N},k\in\mathscr{K}. \tag{3}$$

By including a service level constraint, we ensure that the probability for costly stock-outs leading to customer dissatisfaction does not exceed $1-\alpha_i^*$, $\forall i\in\mathscr{N}$.

Second, the capacity of all vehicles should be respected. We model this via the chance constraint

$$P\left(\sum_{i\in\mathcal{N}} x_{ik}D_i(T_k) \leq Q\right) \geq \gamma^* \quad \forall k \in \mathcal{K}. \tag{4}$$

This ensures that the probability, that the total demand for retailer group k during T_k periods is less than or equal to the vehicle's capacity Q, is at least γ^*. Chance constraint (4) is added in addition to the expected emergency costs in the objective function for two reasons: First, including it avoids situations where emergency shipments with very little volume are expected to be required for every retailer group and every replenishment cycle. This is from a practical perspective highly desirable because each emergency shipment that needs to be scheduled on an operational level requires time and effort. The chance constraint overcomes the problem of frequent emergency shipments containing little volume. Second, the emergency cost term in the objective function limits only the expected number of emergency shipments, while chance constraint (4) also includes its variability.

Summarizing, the chance-constrained nonlinear optimization model that minimizes the total expected costs per period, as defined in Equation (1), is given by

$$\begin{aligned}
\min \quad & \mathrm{TC}(\boldsymbol{S},\boldsymbol{T},\boldsymbol{X},\boldsymbol{Y}) && && (5)\\
\text{s.t.} \quad & P(\mathrm{IL}_i^+(T_k,S_i)>0) \geq \alpha_i^* && \forall i\in\mathcal{N}, k\in\mathcal{K}, && (6)\\
& P\left(\sum_{i\in\mathcal{N}} x_{ik}D_i(T_k) \leq Q\right) \geq \gamma^* && \forall k\in\mathcal{K}, && (7)\\
& x_{ik} = \sum_{j\in N^0} y_{ijk} && \forall i\in\mathcal{N}, k\in\mathcal{K}, && (8)\\
& \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{N}^0\setminus\{i\}} y_{jik} = 1 && \forall i\in\mathcal{N}, && (9)\\
& \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{N}^0\setminus\{i\}} y_{ijk} = 1 && \forall i\in\mathcal{N}, && (10)\\
& \sum_{i\in U}\sum_{j\in U} y_{ijk} \leq |U|-1 && \forall k\in\mathcal{K}, \forall U\subset\mathcal{N}, |U|>2, && (11)\\
& S_i \in \mathbb{R} && \forall i\in\mathcal{N}, && (12)\\
& T_k \in \mathbb{N}_{\geq 1} && \forall k\in\mathcal{K}, && (13)\\
& x_{ik} \in \{0,1\} && \forall i\in\mathcal{N}, k\in\mathcal{K}, && (14)\\
& y_{ijk} \in \{0,1\} && \forall i\in\mathcal{N}^0, j\in\mathcal{N}, k\in\mathcal{K}. && (15)
\end{aligned}$$

In this formulation, the constraints (6) and (7) are the chance constraints discussed previously, the constraints (8) link the x_{ik} variables to the y_{ijk} variables, the constraints (9) and (10) ensure that each retailer is visited exactly once, the constraints (11) enforce the condition that there is only a single tour among all retailers belonging to a single retailer group k, and the constraints (12)–(15) indicate the domain of the decision variables. Note that the presented model results in a stochastic mixed-integer nonlinear problem, due to constraints (6) and (7). Therefore, we present an integer linear reformulation in the next subsection and discuss how to handle constraints (6) and (7).

4.2. An Integer Linear Reformulation

In the remainder of this paper, we will work with an integer reformulation of the model (5)–(15). We obtain this via a Dantzig-Wolfe reformulation on the constraints related to the feasibility of individual retailer clusters. The resulting formulation is a set-partitioning model, where we select retailer *clusters* that together partition the set of retailers $\mathscr{N}$. Each cluster comprises a set of retailers with associated vehicle route, shipment interval, and corresponding individual base-stock levels in order to meet the service levels α_i^*. Furthermore, for each cluster, it is ensured that with probability γ^* the truck capacity is sufficiently large, as modeled via the chance constraints (7). How to explicitly calculate this is discussed in Lemmas 2 and 3, and Corollary 1 in Section 5.

The formulation comprises an exponentially large set of clusters. Let $\mathscr{R}$ be a collection of retail clusters, where each cluster $r \in \mathscr{R}$ describes a single vehicle route along all retail stores in the cluster. Let β_r^i be equal to 1 if retailer i is contained in cluster r, and 0 otherwise. Furthermore, we define T_r as the corresponding optimal shipment interval and S_r^i as the optimal base-stock level of retailer i with $\beta_r^i = 1$ (see Section 5.1.1 for their calculation).

The costs c_r of cluster $r \in \mathscr{R}$ are defined as

$$c_r = \frac{1}{T_r}(w_r + W) + \frac{e}{T_r}\mathbb{E}\left[\left(\sum_{i=1}^{N}\beta_r^i D_i(T_r) - Q\right)^+\right] + h\sum_{i=1}^{N}\beta_r^i \frac{1}{T_r}\sum_{t=1}^{T_r}\mathbb{E}[\mathrm{IL}_i^+(t, S_r^i)], \tag{16}$$

where w_r is equal to the variable transportation costs within cluster r. Let z_r be binary decision variables equaling 1 if cluster r is selected and 0 otherwise. Then, the inventory routing problem with stochastic demand can be formulated as

$$\mathrm{MP}(z) = \min \quad \sum_{r\in\mathscr{R}} c_r z_r \tag{17}$$

$$\text{s.t.} \quad \sum_{r\in\mathscr{R}} \beta_r^i z_r = 1 \qquad \forall i \in \mathscr{N}, \tag{18}$$

$$z_r \in \{0,1\} \qquad \forall r \in \mathscr{R}. \tag{19}$$

Here, the objective (17) minimizes the costs of the selected retail clusters. The constraints (18) ensure that each customer is assigned to exactly one retail cluster, and the constraints (19) indicate the domain of the variables.

In the above formulation, the set of clusters $\mathscr{R}$ is of exponentially large size and cannot be enumerated. Instead, we consider the above master formulation restricted to a subset $\bar{\mathscr{R}} \subset \mathscr{R}$ of the variables. We call this the restricted master problem (RMP) and denote its solution value by $\mathrm{RMP}(z)$. The linear relaxation of the RMP (i.e., the replacement of (19) by $z_r \geq 0 \; \forall \; r \in \bar{\mathscr{R}}$) is solved to optimality using column generation (see, e.g., Barnhart et al. 1998, Lübbecke and Desrosiers 2005).

In column generation, one iteratively generates retail clusters $r \in \mathscr{R} \setminus \bar{\mathscr{R}}$ of negative reduced cost, adds these clusters to the RMP, and consequently (re)solves the linear relaxation of the RMP. If there are no retail clusters of negative reduced cost, then the linear relaxation of the RMP is provably optimal for the linear relaxation of the MP. The major challenge in column generation is therefore to determine if there are no retail clusters of negative reduced cost left after solving the linear relaxation of the RMP. This problem is called the pricing problem, and formally asks for a solution of

$$\min_{r \in \mathscr{R} \setminus \bar{\mathscr{R}}} \hat{c}_r := c_r - \sum_{i \in N} \beta_r^i \pi_i. \tag{20}$$

Here, $\hat{c}_r$ is called the reduced cost of cluster r, and $\pi_i \in \mathbb{R}$ are the dual variables corresponding to the constraints (18). If the pricing problem returns a cluster r with $\hat{c}_r > 0$, then we are certain that there are no retail clusters of negative reduced cost left, and the linear relaxation of the RMP is optimal for the linear relaxation of the MP. For our problem, in particular because of the chance constraints (6) and (7), this pricing problem is a new variant of the elementary resource-constrained shortest-path problem (ERCSPP).

Integer optimality is then obtained by embedding column generation in a branch-and-bound scheme (called branch-and-price), and its computational efficiency is enhanced by including valid inequalities. The resulting approach is called branch-price-and-cut, which we discuss in detail in Section 5.

5. Branch-Price-and-Cut Algorithm

The branch-price-and-cut algorithm to solve the formulation (17)–(19) consists of three parts that will be discussed in the following. First (see Section 5.1), we describe a tailored labeling algorithm to solve the pricing problem (20). We give structural insights into our problem and present a novel bounding method based on 0,1-knapsack relaxations. Second (see Section 5.2), we outline the remaining ingredients of the branch-and-price-and-cut algorithm by discussing the valid inequalities used, the branching decisions, and the node selection rules. Third, we introduce several upper-bounding methods that are used to initialize the set of retail clusters but can also serve as constructive heuristics to solve the problem in negligible time.

5.1. Solving the Pricing Problem

The pricing problem differs from traditional ERCSPPs because in addition to the truck routes—determined as in traditional ERCSPPs— we also need to determine the associated shipment intervals for the clusters and the base-stock levels at all retailers belonging to each cluster.

We propose a tailored labeling algorithm to solve our pricing problem. A labeling algorithm is a dynamic-programming-based method in which we iteratively extend partial retailer clusters, called *labels*, with a new retailer and check its feasibility. The core of its efficiency lies in the ability to prune labels based on dominance criteria. These criteria specify the conditions under which a label dominates another label, meaning that the dominated label does not need to be considered, because it will not lead to an improved solution. Besides dominance criteria, pruning based on optimality criteria is gaining popularity. Optimality

criteria consider whether or not a label and all its possible extensions can potentially improve the current best-known solution. If not, the label can be pruned. We consider both dominance and optimality criteria, as we will detail in Section 5.1.2. Before that, we first define the elements of a label and the resource-extension functions (REFs) that describe how label elements change after extending it by a retailer in Section 5.1.1.

In the remainder of this section, we assume that customer demand D_i per period at retailer i is gamma-distributed with retailer-dependent shape parameter κ_i and equal scale parameter θ. The advantage of considering a gamma distribution is that the probability density function can depict very different shapes and thereby cover a broad variety of possible demand distributions. Although we use a gamma distribution to model demand at the retailers, our algorithm can be adapted easily for other demand distributions (e.g., normal or Poisson distributions).

5.1.1. Label Definition, Feasibility, and Extension. To track feasibility and to apply dominance criteria, we define a label L by the following elements:

1. the current partial path of retail stores $\vec{v}(L)$, with $i(L)$ denoting the retail store at the end of $\vec{v}(L)$;
2. a sorted set of unreachable retailers $\mathscr{U}(L) \subset \mathscr{N}$ denoting the retailers that can no longer be visited or that are visited already in the label;
3. the sum $\kappa(L)$ of shape parameters κ_i of all retailers i contained in L;
4. the incurred dual costs $\hat{c}(L)$, i.e., the sum of dual variables π_i for all retailers i contained in L;
5. the total transportation costs $w(L)$;
6. the optimal shipment interval $t(L)$;
7. the truck capacity used $q(L)$, defined as the $\gamma\%$ percentile of the joint demand distribution associated with shipment interval $t(L)$ of all the retailers in label L;
8. the objective value $c(L)$ composed of regular and expected emergency transportation and holding costs minus the incurred dual costs.

A label L is initialized by a single warehouse visit, i.e., $\vec{v}(L) = (0), \mathscr{U}(L) = \emptyset, \kappa(L) = \hat{c}(L) = t(L) = q(L) = c(L) = 0$, and $w(L) = W$. The basic action of a labeling algorithm is to extend a label L with a new retailer $i \in \mathscr{N}^0$ into an extended label L'. Before we detail the resource extension functions that update the elements of L upon extension, we explain how the optimal shipment interval $t(L)$ can be computed for a given label. For better comprehensibility, we first summarize some properties of the gamma distribution. After this, we detail some properties of retailers included in a label, i.e., how to determine the base-stock level and the expected cycle and safety stock. Then, we indicate how to calculate the expected amount of emergency shipments and how truck capacity should be respected. We present this in a sequence of three lemmas.

LEMMA 1 **(Gamma distribution)**. *Consider a given label L. For readability, assume that $\mathscr{U}(L)$ equals the set of retailers contained in $\vec{v}(L)$. Let $t(L)$ be the associated shipment interval. Assume that each retailer $i \in \mathscr{U}(L)$ faces gamma-distributed demand $\Gamma(\kappa_i, \theta)$ with shape parameters κ_i and equal-scale parameter θ.*

- *The total demand at retailer i during a replenishment cycle of length $t(L)$, $D_i(t(L))$, is gamma-distributed with shape parameter $t(L)\kappa_i$ and scale parameter θ: $D_i(t(L)) \sim \Gamma(t(L)\kappa_i, \theta)$.*
- *The total demand for label L per replenishment cycle of length $t(L)$, $D_L(t(L))$, is also gamma-distributed with parameters $t(L)\kappa(L)$ and θ: $D_L(t(L)) \sim \Gamma(t(L)\kappa(L), \theta)$.*

Using this lemma on the properties of the demand distribution faced by a label L, we can calculate the base-stock levels at the contained retailers and determine the expected cycle and safety stock as shown in Lemma 2.

LEMMA 2 **(Retailer properties)**. *For any given label L, the following statements are true for a retailer $i \in \mathcal{N}$ contained in L:*

- *The optimal base-stock level at retailer i as a function of $t(L)$ equals $S_i(t(L)) = F^{-1}(\alpha_i^*; t(L)\kappa_i, \theta)$, where $F^{-1}(x; a_1, a_2)$ refers to the inverse cumulative distribution function of a gamma-distributed random variable with shape parameter a_1 and scale parameter a_2 at point x. The formula follows directly from constraint (6). This result is general for any inverse cumulative distribution function F^{-1}.*
- *The cycle stock per period under a replenishment cycle of length $t(L)$ at retailer i equals $I_i^{cs}(t(L)) = \frac{1}{2}\bar{d}_i(t(L))$, where $\bar{d}_i(t(L))$ equals the mean demand at retailer i during $t(L)$ periods. Note that this result is independent of the considered demand distribution. Under gamma distributed demand, this can be calculated as $\bar{d}_i(t(L)) = t(L)\kappa_i\theta$.*
- *The safety stock per period at retailer i equals the expected positive amount of stock that is available at the end of each replenishment cycle just before a new replenishment is received: $I_i^{ss}(t(L)) = \mathbb{E}[(S_i(t(L)) - D_i(t(L)))^+]$. This term can be simplified under gamma-distributed demand to*

$$S_i(t(L)) - \bar{d}_i(t(L)) + \bar{d}_i(t(L))(1 - F(S_i(t(L)); t(L)\kappa_i + 1, \theta)) - S_i(t(L))(1 - F(S_i(t(L)); t(L)\kappa_i, \theta)),$$

where $F(x; a_1, a_2)$ denotes the cumulative distribution function of a gamma-distributed random variable with shape parameter a_1 and shape parameter a_2 at point x (see, e.g., Silver et al. (1998)). Note that the calculation of the safety stock as $I_i^{ss}(t(L)) = \mathbb{E}[(S_i(t(L)) - D_i(t(L)))^+]$ is independent of the considered demand distribution. However, the simplification denoted above depends on the used demand distribution.

Besides the properties of each retailer, we need to determine the required amount of emergency shipments of label L that is directly related with the truck capacity. We summarize this in the following lemma:

LEMMA 3 **(Label properties)**. *For each label L, the following holds:*

- *The expected emergency costs for label L are calculated by multiplying the unit emergency cost with the expected amount of units exceeding truck capacity on the occasion of a replenishment, given by*

$$E_L(t(L)) = \mathbb{E}[(D_L(t(L)) - Q)^+]$$

$$= \int_{Q}^{\infty} (u-Q) f(u; t(L)\kappa(L), \theta)\, \mathrm{d}u,$$

where $f(x;a_1,a_2)$ is defined as the probability density function of a gamma-distributed random variable with shape parameter a_1 and scale parameter a_2. The integral in this equation is called the first-order loss function and can be simplified according to Silver et al. (1998) to

$$\bar{d}_L(1-F(Q;t(L)\kappa(L)+1,\theta)) - Q(1-F(Q;t(L)\kappa(L),\theta)),$$

where $\bar{d}_L(t(L)) = t(L)\kappa(L)\theta$ is defined as the mean demand per replenishment cycle of label L.

- *The probability of exceeding capacity Q in a replenishment cycle with shipment interval $t(L)$, $P(D_L(t(L)) > Q)$, equals $1-F^{-1}(\gamma^\star; t(L)\kappa(L), \theta)$.*

Based on Lemma 2 and 3, we can calculate the optimal shipment interval $t(L)$ of label L given gamma-distributed demand at each retailer, as shown in the following corollary:

COROLLARY 1. *For a given label L, the optimal shipment interval $t(L)$ under gamma-distributed demand is given by*

$$t(L) := \underset{u\in\mathbb{N}^+}{\arg\min}\{\Phi(u,L) \mid P(D_L(u) \le Q) \ge \gamma^\star\}, \tag{21}$$

$$\Phi(u,L) = \frac{w(L)}{u} + \frac{e}{u}E_L(u) + h\sum_{i\in\vec{v}(L)}\left[I_i^{\mathrm{cs}}(u) + I_i^{\mathrm{ss}}(u)\right]. \tag{22}$$

Note that $I_i^{\mathrm{cs}}(u)$ and $I_i^{\mathrm{ss}}(u)$ refer to the optimal expected cycle and safety stock for any given shipment interval u and are defined in Lemma 2.

Lemmas 1-3 and Corollary 1 define how the different cost components of a label can be calculated in case demand is gamma distributed. We continue our exposition by introducing the resource-extension functions that describe how the label elements should be updated in case a label L is extended with an arbitrarily retailer $j \in \mathcal{N}^0$ into a new label L'. That is, label L' is obtained as follows:

1. $\vec{v}(L') = (\vec{v}(L), j)$;
2. $\mathscr{U}(L') = \mathscr{U}(L) \cup \{j\}$;
3. $k(L') = k(L) + \kappa_j$;
4. $\hat{c}(L') = \hat{c}(L) + \pi_j$;
5. $w(L') = w(L) + c_{ij}$, where i is the last node of $\vec{v}(L)$.
6. $t(L') = \arg\min_{u\in\mathbb{N}^+}\{\Phi(u,L') | P(D_{L'}(u) \le Q) \ge \gamma^\star\}$;
7. $q(L') = F^{-1}(\gamma^\star; t(L')k(L'); \theta)$;
8. $c(L') = \Phi(t(L'), L') - \hat{c}(L')$.

Note that in the case $j = 0$, we only update the path $\vec{v}(L')$ and leave the other label elements untouched. Feasibility of a label follows trivially from the label definition; i.e., extending label L with a retailer j into label L' is feasible if the truck capacity constraints are respected [$q(L') \le Q$] and if $j \notin \mathscr{U}(L)$. When a label is extended with the central warehouse, i.e., if $j = 0$, we check whether $c(L') < 0$ and transform the label with negative reduced cost into a new retailer cluster and add it to the set $\bar{\mathscr{R}}$.

5.1.2. Dominance and Optimality Criteria. Here, we introduce dominance and optimality criteria in order to prune labels during the labeling algorithm. This is crucial for the efficiency of the algorithm. If these criteria are not included, the labeling algorithm simply enumerates all possible clusters. The dominance and optimality criteria are valid under the assumption that we fix the shipment interval during execution of the labeling algorithm. Let $t^{\text{lb}} \leq t(L) \leq t^{\text{ub}}$ for all labels L. In the remainder of this section, we assume that the shipment interval is fixed (i.e., $t^{\text{lb}} = t^{\text{ub}}$). In Section 5.1.3, we indicate how we ensure optimality of the complete labeling algorithm.

We now discuss how a label L in the labeling algorithm can be disregarded. First, we consider dominance criteria that disregard a label because there provably exists another label L' that results in the same label extensions as L but at lower cost. The dominance criteria are similar to those of the capacitated vehicle routing problem (see, e.g., Costa et al. 2019). For completeness, we restate these dominance criteria. A label L is said to be dominated by label L' if the following three conditions hold:

$$\mathscr{U}(L') \subseteq \mathscr{U}(L), \quad c(L') \leq c(L), \quad q(L') \leq q(L). \tag{23}$$

From left to right, these state that L' dominates L if L' contains a subset of retailers only, has lower reduced cost, and less vehicle capacity used.

Second, we provide an optimality criteria based on a so-called completion bound. This is a lower bound on the reduced cost that can be obtained by all possible extensions from a particular label L. If we keep track of the best known solution so far to the pricing problem, we can disregard L if we can show that L (or its extensions) cannot improve that current best-known solution. Note, we also disregard L if the completion bound shows us that only a positive reduced cost can be achieved from further extending this label.

The completion bound exploits the fact that the dual costs are only incurred at the retailers (and not on the actual arcs chosen) and that holding costs can be calculated directly for a fixed shipment interval. Consequently, our completion bound considers all retailers $i \notin \mathscr{U}(L)$ and "collects" the sum of positive dual costs and the minimum holding costs at i associated with the shipment interval of label L. We do this so that we obtain a lower bound on the reduced cost of any cluster that results from extending label L.

To obtain a valid completion bound, we need an easy-to-compute lower bound on the increase of $q(L)$, independent of the actual demand distributions of retailers contained in L. Our lower bound uses the following property of gamma-distributed demand:

LEMMA 4 **(Gamma-distribution quantile)**. *Let $D_i \sim \Gamma(\kappa_i, \theta)$ and $D_j \sim \Gamma(\kappa_j, \theta)$ be gamma-distributed independent random variables, representing the demand per period of retail stores i and j. Then, for any $i, j \in \mathbb{N}, \kappa_i, \kappa_j \in \mathbb{R}^+$, $\theta \in \mathbb{R}^+$ and for any parameter constellation $(\kappa_i, \kappa_j, \theta)$, there exists an $\tilde{a}(\kappa_i, \kappa_j, \theta)$ such that for all $a > \tilde{a}(\kappa_i, \kappa_j, \theta)$, the following inequality holds:*

$$F^{-1}(a; \kappa_i + \kappa_j, \theta) \geq F^{-1}(a; \kappa_i, \theta) + \kappa_j \theta. \tag{24}$$

The proof is provided in the Online Appendix. Lemma 4 provides a lower bound on the right tail quantile of the sum of two gamma-distributed random variables, which is useful for constructing our completion bound because it allows us to work with lower bounds on the increase of $q(L)$ by extensions of some label L. In particular, it implies that $q(L)$ increases by less than $\kappa_j\theta$ if some label L is extended with retailer $j \in \mathcal{N}$.

We define for each retailer i its so-called completion costs CC_i:

$$\mathrm{CC}_i = \frac{w}{t^{\mathrm{lb}}} \min_{j \in \mathcal{N}^0} c_{ij} - \pi_i + h[I_i^{\mathrm{cs}}(t^{\mathrm{lb}}) + I_i^{\mathrm{ss}}(t^{\mathrm{lb}})]. \tag{25}$$

The completion costs CC_i consist of the shortest outgoing arc from i, its associated dual costs π_i, and the holding cost at node i with shipment interval t^{lb}. For any label L and retailer $i \notin \mathscr{U}(L)$, the completion costs CC_i are a lower bound on the increase in $c(L)$, because CC_i does not account for the emergency shipment costs and the shipment interval is fixed to t^{lb}. In other words, it holds that the extended label $L \cup \{i\} := L'$ has costs $c(L') \geq c(L) + \mathrm{CC}_i$.

The completion bound, defined for any label L, then consists of selecting the most "profitable" retailers $i \notin \mathscr{U}(L)$ according to their completion costs CC_i. Using Lemma 4, we define the completion bound as the solution to the linear optimization problem

$$\bar{z}(L) = \min \quad \mathrm{CC}_i y_i \tag{26}$$

$$\text{s.t.} \quad \sum_{i \notin \mathscr{U}(L)} \theta \kappa_i y_i \leq Q - q(L), \tag{27}$$

$$0 \leq y_i \leq 1 \quad \forall i \notin \mathscr{U}(L). \tag{28}$$

This is a linear knapsack problem ($\mathrm{CC}_i \in \mathbb{R}$), which is solved by sorting all retailers $i \notin \mathscr{U}(L)$ in non-increasing order according to $\mathrm{CC}_i/(\theta\kappa_i)$ and assigning the $Q - q(L)$ remaining capacity according to the sorted list of retailers. This sorting operation is done before the labeling algorithm starts and does not add complexity to solving the pricing problem. However, the assignment of remaining capacity $Q - q(L)$ to retailers $i \notin \mathscr{U}(L)$ depends on L and is performed when the optimality pruning criteria are invoked. This can be done quickly as we can sort the retailers based upon $\mathrm{CC}_i/(\theta\kappa_i)$ before invoking the labeling algorithm as these weights are independent from the constructed partial retailer clusters in the labeling algorithm. We summarize these optimality pruning criteria by our completion bound in the following lemma:

LEMMA 5 **(Completion bound)**. *Let L be a given label, and let $\bar{z}(L)$ be defined as above. Let z^0 be the best solution found so far during the execution of the labeling algorithm. If $c(L) + \bar{z}(L) > z^0$, then any extension of label L will result in a label of cost larger than z^0, and label L can be pruned.*

5.1.3. Labeling Algorithm Procedure. The labeling algorithm procedure works as follows. After observing the LP relaxation to the RMP, we iteratively run the labeling algorithm for the parameter $\bar{t} \in \{\bar{t}^{\max}, \bar{t}^{\max} - 1, \ldots, 1\}$. For $\bar{t} = \bar{t}^{\max}$, we set $t^{\mathrm{ub}} = \infty$ and $t^{\mathrm{lb}} = \bar{t}$. For other values of $\bar{t}$, we impose $t^{\mathrm{lb}} = t^{\mathrm{ub}} = \bar{t}$.

Note that in the case $t^{\text{ub}} = \infty$, our optimality pruning criteria are not valid, which is not a problem, because setting $\bar{t}^{\max}$ sufficiently large ensures that the vehicle capacity is very restricted in this case. In our experiments, we set $\bar{t}^{\max} = 4$. Note, this does not imply that we consider only shipment intervals of at most 4, but is rather the value from which disregarding the optimality pruning criteria does not harm computational performance. We abort the enumeration over $\bar{t}$ if for some $\bar{t}$ we identified retailer clusters of negative reduced cost. We then solve the linear relaxation of the RMP and restart our iterative labeling algorithm. We terminate a single call to the labeling algorithm procedure after we have found 5000 clusters of negative reduced cost, after which we resolve the RMP and restart the labeling algorithm procedure.

5.2. Branching and Valid Inequalities

Valid branching rules are required to obtain a working exact solution method. We make use of two branching rules. First, we branch on an integral number of vehicles being used. For a given LP relaxation to the RMP, we define $z^*_{ij} := \sum_{r \in \bar{\mathscr{R}}} \delta^r_{ij} z^*_r$, where δ^r_{ij} is a binary parameter indicating if retailer j is visited directly after retailer i in cluster r, and z^*_r is the value of z_r in the LP relaxation of the RMP. Then, if $\sum_{j \in \mathscr{N}} z^*_{0j}$ is fractional, we create two child nodes where we impose $\sum_{r \in \bar{\mathscr{R}}} \delta^r_{0j} z_r \leq \lfloor \sum_{j \in \mathscr{N}} z^*_{0j} \rfloor$ and $\sum_{r \in \bar{\mathscr{R}}} \delta^r_{0j} z_r \geq \lceil \sum_{j \in \mathscr{N}} z^*_{0j} \rceil$, respectively. If an integral number of vehicles is used, we continue with branching on individual arcs. That is, we select the arc (i, j) for which z^*_{ij} is closest to 0.5, and create two child nodes: one child node where we enforce arc (i, j) to be traversed by at least a single cluster, and one in which we do not allow arc (i, j) to be traversed. Note that the branching rule on the number of vehicles imposes cuts on the model formulation, which we take into account in our pricing problem by initializing the dual cost component of our labeling algorithm with the dual values associated with the branching cut, i.e., we subtract them from the outgoing arcs of the depot. Note that this is equal to initializing the label reduced cost with the dual costs associated with these constraints. Furthermore, the branch rule on individual arcs requires to dynamically adjust the set of generated clusters during the branch-and-bound search by setting local upper bounds of 0.0 in the corresponding master variables. For arcs set to zero, this is done by trivially selecting the retailer clusters that visit that arc. If an arc is set to one, it implies that other outgoing or incoming arcs (except for arcs leaving or entering the depot) are set to zero, and we then select the associated retailer clusters. Node selection is done using the default node selection method from the constraint programming environment SCIP 6.0.2 (Gleixner et al. 2018), which we use to code our branch-price-and-cut algorithm.

In addition, we add subset-row inequalities (Jepsen et al. 2008) to further strengthen our LP relaxation. Subset-row inequalities are defined for an $\mathscr{O} \subset \mathscr{N}$ of retailer nodes and an integer $1 \leq \phi \leq |\mathscr{O}| - 1$, and are given by

$$\sum_{r \in \bar{\mathscr{R}}} \left\lfloor \frac{1}{\phi} \sum_{i \in \mathscr{O}} \beta^i_r \right\rfloor z_r \leq \left\lfloor \frac{|\mathscr{O}|}{\phi} \right\rfloor. \tag{29}$$

Preliminary experiments have shown that only subset-row inequalities considering $|\mathscr{O}| = 3$ and $\phi = 2$ have a significant effect on the quality of the root node relaxation. We have therefore only included those in our branch-price-and-cut algorithm. Separation is done via complete enumeration and is considered in every node of the branch-and-bound tree, because it requires negligible time compared with solving the pricing problems. Finally, including these inequalities requires some standard adaptations of our labeling algorithm, because the subset-row inequalities considered in this paper are so-called "non-robust cuts." We have made similar changes to those outlined in, for instance, Costa et al. (2019), Schrotenboer et al. (2019), and thus introduce new label elements for each included subset-row cut to keep track on the number of visited retailers associated with the cut. This is needed because a variable only enters a subset-row inequality if at least two of the three associated retailers are visited.

5.3. Upper-bounds

In this section, we briefly introduce three constructive heuristics to solve the considered problem. These heuristics serve two purposes. First, they provide relatively high-quality upper bounds very quickly, which is of practical relevance. Second, all distinct clusters evaluated during execution of the heuristics are used as an initial set of clusters for the branch-price-and-cut algorithm. This speeds up the overall run-time of the branch-price-and-cut algorithm by potentially reducing the number of columns to be generated and by providing a better starting point for primal heuristics included by default in SCIP. A detailed description of the heuristics including pseudo codes is provided in the Online Appendix.

KM: The first upper bounding method is a tailored version of a K-means algorithm, which is the "best-known clustering algorithm" (Geetha et al. 2009). Using this algorithm, all retailers are iteratively assigned to the closest feasible cluster according to Euclidean distances, which entails the consideration of vehicle capacities throughout the clustering procedure.

SAV: The second heuristic is based on ideas from the classical savings algorithm by Clarke and Wright (1964) for the vehicle routing problem. The algorithm starts by initializing a set of clusters consisting of single retailers that are replenished using direct deliveries as in Kleywegt et al. (2002). Then, the algorithm iteratively merges two clusters until no further cost savings can be achieved.

KMSAV: The third heuristic is a combination of the first two algorithms. In an initial step, the algorithm clusters all retailers into "regions" based on a K-means algorithm that ignores capacity restrictions. Afterwards, we apply the savings-based heuristic within each generated "region". Numerical experiments have shown that such a two-step procedure can result in significant cost reductions induced by relaxing the order in which retailer clusters are merged, i.e., if the retailers to be merged in the classical SAV algorithm belong to two different regions, they cannot be merged, which can lead to better decisions in the long run when more clusters are consolidated.

6. Computational Results

This section evaluates the performance of the branch-price-and-cut algorithm. Besides the complete branch-price-and-cut algorithm that we abbreviate as MIP-BPC, we also introduce two MIP-based heuristics based on the set partitioning formulation in Section 4.2 that we refer to as MIP-CG and MIP-H. MIP-CG is a heuristic variant of MIP-BPC and considers column generation in the root node only, which reduces computation times significantly and allows for solving larger instances close to optimality. MIP-H builds upon the exploration of retailer clusters by the three constructive heuristics (KM, SAV, and KMSAV) discussed in Section 5.3. During the execution of each of the three heuristics, we consider many different retailer clusters, all of which we store in memory together with their corresponding optimal shipment intervals and base-stock levels. The MIP-H heuristic then solves the set-partitioning formulation RMP subject to this fixed set of clusters without invoking column generation.

All algorithms are implemented in C++17, using the framework for constraint programming SCIP 6.0.2 and CPLEX 12.8. All the implementations are completely single-threaded. The experiments are performed on an Intel Xeon E5 2680v3 2.5 GHz CPU with 40 GB of RAM allocated. In the following, we first provide details on the benchmark instances used in Section 6.1. Then, in Section 6.2, we evaluate the performance of our exact and heuristic methods on the benchmark instances. In Section 6.3, we illustrate the importance of considering stochastic customer demand, by studying the value of the stochastic solution. In Section 6.4, we show how the different cost components steer the structure of the optimal solution.

6.1. Instance Characteristics

We use two different benchmark sets. Benchmark Set A is based on the deterministic instances provided by Raa (2006) and reflects applications where the mean demand per period is relatively low compared to the truck capacity, which leads to long shipment intervals and, therefore, infrequent replenishments. Because discussions with supermarkets and companies in the retail sector have shown that replenishment frequencies are usually much higher (up to several times a day) due to relatively high mean demand per period in relation to the truck capacities, we introduce a new Benchmark Set B that captures this characteristic. In the following, we briefly describe both benchmark sets and refer to the accompanying data for full descriptions of the individual instances.

Similar to Raa and Aouam (2021), we adapt for Benchmark Set A a deterministic IRP benchmark set of Raa (2006). Raa and Aouam (2021) assume a normally distributed demand and adjust the benchmark set of Raa (2006) by adding demand variability in terms of different standard deviations. We are using the same approach of adding variability to the demand. However, because we are considering gamma distributed demand to be able to model higher demand variability, we set the scale parameter θ to either 0.8 or 1.0, while ensuring at the same time that the mean of the gamma distributed customer demand $\kappa_i\theta$ equals the deterministic demand rate used by Raa (2006).

As already mentioned, the instances of Raa (2006) have relatively low mean demand compared to the truck capacity, which would result without additional constraints to in practice unreasonably long routes. Therefore, Raa (2006) add a constraint on the maximum route duration and set it to 8 hours. Moreover, there is a fixed service time of 15 minutes per retailer included. For solving the instances in Benchmark Set A, we include these required and straightforward adaptions to our methods to ensure feasibility with respect to the maximum route duration.

Similar to Raa (2006), we set the fixed vehicle cost W to 100 and selected the instances with holding costs 0.8 per unit and time unit. To adjust the deterministic benchmark set, we add emergency shipment costs e equal to 50, retailers' target service levels α_i^* of 95%, and a service level γ^* on the truck capacity of 90%. Finally, we selected both instances of Raa (2006) with a small and large geographical area and define the following instance types based on the size of the geographical area and the scale parameter $\theta \in \{0.8, 1.0\}$: Type 0 implies a large geographical area with $\theta = 0.8$, Type 1 implies a large geographical area with $\theta = 1.0$, Type 2 implies a small geographical area with $\theta = 0.8$, and Type 3 implies a small geographical area with $\theta = 1.0$. The original benchmark set comprises 10 instances with at least 70 customers each. For each combination of number of retailers and instance type, we solve all 10 instances but select the first $N \in \{25, 30, 35, 40\}$ customers to be able to solve the instances to optimality.

Benchmark set B is a newly constructed benchmark set that covers—according to our discussions with businesses—practically more relevant instances with high demand rates and, therefore, frequent replenishments. More precisely, we set the truck capacity Q to either 70 or 90 and the parameters of the gamma distributed demand for all $N \in \{20, 25, 30, 35, 40, 45, 50, 55, 60\}$ retailers as follows: the scale parameter θ is set to $15/16$ and we randomly draw the shape parameter κ_i for each retailer i between 10 and 22. This results in an average shape parameter $\bar{\kappa}$ equal to 16, which corresponds to an average coefficient of variation of $\text{CoV} = \sqrt{\bar{\kappa}\theta}/\bar{\kappa}\theta = 1/\sqrt{\bar{\kappa}} = 0.25$. All retailers are randomly located according to a uniform distribution in a 100×100 box and we fix the location of the warehouse to coordinates $(50, 50)$. The cost parameters are set as follows: holding costs $h \in \{0.5, 1.0\}$, fixed vehicle cost $W = 100$, emergency shipment costs $e = 50$, and travel cost c_{ij} equal to the Euclidean distance between nodes i and j. The service levels α_i^* and γ are set equal to those in Benchmark Set A, i.e., $\alpha_i^* = 0.95 \forall i = 1, ..., N$ and $\gamma = 0.90$. For each combination of these parameters, we randomly create 10 instances. The randomness thereby lies in the demand at the retailers and the location of retailers.

6.2. Performance of MIP-BPC, MIP-CG, and MIP-H

In Tables 3 and 4, we provide an overview of the performance of MIP-BPC, MIP-CG, and MIP-H on Benchmark Sets A and B, respectively. Each row represents averaged values for the particular parameter combinations over all instances solved. We refer for the solutions of the individual instances to the supplementary materials that can be downloaded from the publishers website. The column "$\#_{\text{rep}}$" denotes the

Table 3 Performance of MIP-BPC, MIP-CG, and MIP-H on Benchmark Set A

			MIP-BPC				MIP-CG					MIP-H	
N	Type	$\#_{rep}$	$\#_{sol}$	$\Delta(\%)$	$\max\Delta(\%)$	$time^{opt}(s)$	$\#_{sol}$	$\Delta_{UB}(\%)$	$\Delta_{LB}(\%)$	time(s)	$time^{opt}(s)$	$\Delta_{UB}(\%)$	$\Delta_{LB}(\%)$
25	0	10	10	-	-	60.90	10	0.00	0.00	13	13	0.01	0.01
	1	10	10	-	-	99.03	10	0.00	0.00	16	16	0.01	0.01
	2	10	10	-	-	645.56	10	0.00	0.00	205	205	0.04	0.04
	3	10	9	0.20	0.20	505.82	10	0.00	0.02	300	299.60	0.08	0.10
30	0	10	9	0.05	0.05	222.93	10	0.00	0.01	61	61	0.07	0.08
	1	10	10	-	-	94.50	10	0.02	0.02	63	63	0.07	0.07
	2	10	9	0.92	0.92	3557.30	10	-0.05	0.04	1715	1715	0.03	0.12
	3	10	9	1.18	1.18	2834.48	10	-0.07	0.05	1716	1717	-0.02	0.10
35	0	10	10	-	-	315.89	10	0.00	0.00	166	166	0.06	0.06
	1	10	10	-	-	406.95	10	0.01	0.01	172	172	0.02	0.02
	2	9	3	1.77	8.63	3930.82	8	-0.11	0.13	4657	4657	-0.86	0.25
	3	9	4	1.04	3.76	8638.73	7	-0.40	0.16	8412	4690	-0.40	0.16
40	0	10	10	-	-	1194.89	10	0.00	0.00	668	668	0.08	0.08
	1	10	10	-	-	1486.75	10	0.00	0.00	946	946	0.04	0.04
	2	5	2	4.88	8.23	8905.49	5	-2.60	0.17	8684	8684	-2.51	0.27
	3	5	0	4.81	11.21	-	3	-3.99	0.52	14143	11015	-4.17	0.31

Table 4 Performance of MIP-BPC, MIP-CG, and MIP-H on Benchmark Set B

				MIP-BPC				MIP-CG					MIP-H	
N	Q	h	$\#_{rep}$	$\#_{sol}$	$\Delta(\%)$	$\max\Delta(\%)$	$time^{opt}(s)$	$\#_{sol}$	$\Delta_{UB}(\%)$	$\Delta_{LB}(\%)$	time(s)	$time^{opt}(s)$	$\Delta_{UB}(\%)$	$\Delta_{LB}(\%)$
20	70	0.5	10	10	-	-	2	10	0.00	0.00	1	1	0.10	0.10
		1.0	10	10	-	-	30	10	0.04	0.04	2	2	0.50	0.50
	90	0.5	10	10	-	-	10	10	0.03	0.03	8	8	0.19	0.19
		1.0	10	10	-	-	36	10	0.07	0.07	16	16	0.69	0.69
25	70	0.5	10	10	-	-	38	10	0.00	0.00	6	6	0.16	0.16
		1.0	10	10	-	-	65	10	0.00	0.00	7	7	0.64	0.64
	90	0.5	10	10	-	-	415	10	0.00	0.00	44	44	0.40	0.40
		1.0	10	10	-	-	2976	10	0.00	0.00	70	70	1.13	1.13
30	70	0.5	10	10	-	-	32	10	0.02	0.02	11	11	0.25	0.25
		1.0	10	10	-	-	80	10	0.00	0.00	14	14	0.85	0.85
	90	0.5	10	10	-	-	554	10	0.01	0.01	108	108	0.48	0.48
		1.0	10	8	0.22	0.29	917	10	0.17	0.21	151	118	0.84	0.88
35	70	0.5	10	10	-	-	69	10	0.00	0.00	27	27	0.18	0.18
		1.0	10	10	-	-	3221	10	0.04	0.04	32	32	0.82	0.82
	90	0.5	10	9	0.10	0.10	1064	10	0.09	0.10	425	392	0.56	0.57
		1.0	10	6	0.27	0.67	1953	10	0.09	0.20	550	388	1.34	1.45
40	70	0.5	10	10	-	-	246	10	0.00	0.00	44	44	0.30	0.30
		1.0	10	9	0.16	0.16	3642	10	0.11	0.12	49	48	0.65	0.67
	90	0.5	10	6	0.05	0.09	5387	10	0.01	0.03	597	603	0.51	0.53
		1.0	10	5	0.70	1.88	8214	10	-0.08	0.26	865	707	0.32	0.67
45	70	0.5	10	10	-	-	502	10	0.06	0.06	90	90	0.32	0.32
		1.0	10	4	0.13	0.19	4027	10	0.08	0.15	97	95	0.66	0.73
	90	0.5	10	6	0.44	0.88	6325	10	-0.07	0.11	1653	1137	0.30	0.48
		1.0	10	3	2.38	14.70	4435	10	-1.32	0.27	2056	1703	-0.57	1.03
50	70	0.5	10	9	0.10	0.10	2319	10	0.02	0.03	125	128	0.38	0.39
		1.0	10	6	0.27	0.42	3284	10	0.11	0.22	133	146	0.64	0.75
	90	0.5	10	5	0.33	0.92	8833	10	-0.10	0.06	2102	2120	0.40	0.57
		1.0	10	3	1.05	4.41	5060	10	-0.48	0.25	2409	2861	0.07	0.80
55	70	0.5	10	7	0.08	0.10	3017	10	0.02	0.05	234	170	0.29	0.32
		1.0	10	4	0.13	0.23	6195	10	0.09	0.16	269	219	0.51	0.59
	90	0.5	10	1	3.79	9.50	16572	9	-3.03	0.19	4730	3564	-2.50	0.74
		1.0	8	1	10.10	13.10	6635	7	-7.55	0.63	6376	5322	-7.10	1.12
60	70	0.5	10	7	0.11	0.15	2302	10	0.02	0.05	313	299	0.40	0.43
		1.0	10	3	0.25	0.52	7194	10	0.01	0.18	306	355	0.50	0.67
	90	0.5	10	0	2.84	7.31	-	10	-2.50	0.24	7054	-	-2.02	0.74
		1.0	8	0	3.99	9.83	-	8	-3.40	0.49	9459	-	-2.73	1.18

number of instances (out of 10) whose root node could be processed within our prespecified time limit of 18,000 seconds and our memory limit of 40GB. We considered only those instances for which we obtained a root node solution in our comparison of the average performance of MIP-BPC, MIP-CG, and MIP-H in each row of tables 3 and 4. The column "Type" in Table 3 reflects the type as outlined before. Other than that, the subsequent columns in Table 3 and 4 are the same: The column "$\#_{\text{sol}}$" indicates how many of the $\#_{\text{rep}}$ instances are solved to optimality for MIP-BPC, or until convergence for MIP-CG. The column "Δ (%)" shows the average optimality gap for the instances not solved to optimality, the column "maxΔ (%)" shows the maximum optimality gap, and the column "time$^{\text{opt}}$ (s)" shows the average computation time of the instances solved to optimality by MIP-BPC. For MIP-CG and MIP-H, we compare the resulting upper bound with the upper bound of MIP-BPC (the column "$\Delta_{\text{UB}}(\%)$") and the lower bound of MIP-BPC (the column "$\Delta_{\text{LB}}(\%)$"). Note that negative values for $\Delta_{\text{UB}}(\%)$ indicate improvements over the upper bounds found by the exact method, which can happen in case the solution is not solved to optimality. Finally, the column "time (s)" gives the average computation time on instances solved to optimality or until convergence by MIP-CG. The computation time of MIP-H ranges from one to a few seconds, and is therefore omitted from the table.

Table 3 shows that our solution approach is able to solve all instances up to 30 retailers in Benchmark Set A to optimality. For larger instances of Type 2 and 3 an optimal solution could not be found for all instances, which is explained by the small geographical area, resulting in relatively low transportation costs and travel times and, therefore, larger clusters. The results indicate that the average percentage deviation of the heuristic approach MIP-CG to the lower bound ($\Delta_{\text{LB}}(\%)$) is rather small. Furthermore, the heuristic MIP-H based on the upper bounding procedures presented in Section 5.3 shows satisfactory results for all instances with an average deviation from the lower bound of at most 0.31%.

Because this paper's focus is on the practically more relevant instances in the retail industry with higher average demand and, therefore, more frequent replenishments, we focus in the following on the analysis of Benchmark Set B. Table 4 shows that MIP-BPC can solve instances to optimality up to $N = 60$ retailers, although computation times increase and the performance of MIP-BPC is affected by vehicle capacity and holding costs. Higher capacity and holding costs lead to an increase in cluster sizes, and therefore to a larger solution space, which increases the computational efforts. The reported average optimality gaps are, if not solved to optimality, rather small for MIP-BPC, and also the maximum optimality gaps are within a 14.7% range. Notice that for the larger instances, solving the pricing problem in each branch-and-bound node is a rather time-consuming aspect of MIP-BPC, which may impair its ability to search for high-quality upper bounds.

In comparison, the performance of MIP-CG is outstanding. It converges on all instances except two, with the average deviation from the lower bound of MIP-BPC being only 0.11%. On the smaller instances, it typically finds the same solution as MIP-BPC, and on larger instances it exploits its relatively fast branch-and-bound process compared with MIP-BPC to achieve significantly better upper bounds, although at the expense

of missing the optimal solution with a small probability. On the instances with more than 50 customers and high vehicle capacity, the upper bounds of MIP-CG outperform those of MIP-BPC by a few percentages. Comparing the computation times on the instances solved to optimality by MIP-BPC, we observe that MIP-CG only requires 251 s on average, compared with 1951 s for MIP-BPC.

Comparing the generated number of variables in the root node (i.e., the total number of variables of MIP-CG) versus the heuristically generated columns via the constructive heuristics (i.e., the total number of variables of MIP-H), we observe that only few extra variables are generated in the root node. The average number of variables for MIP-CG and MIP-H equals 1930 and 1846 for the instances of at most 35 customers, respectively. For the instances with more than 35 customers, these numbers are on average 7626 and 7414. This implies, that the constructive heuristics search the solution space rather efficiently, though not efficient enough to close the optimality gap.

Summarizing, we can solve medium size instances up to 30 retailers for almost all instances to optimality. Because the performance of MIP-BPC deteriorates with increasing problem size, MIP-CG can be used to generate near-optimal solutions for large instances in reasonable time. The hybrid heuristic MIP-H returns excellent solutions in negligible time, and is therefore an attractive method solving practically sized problem instances.

6.3. The Value of the Stochastic Solution

To study the impact of demand uncertainty on the composition of retailer groups and costs, we provide a deterministic counterpart to the solution of MIP-CG. This so-called expected value solution is obtained by considering deterministic retailer demand equal to the mean demand in our stochastic model. We set the available truck capacity to $\gamma^* Q$, which is equivalent to considering $(1-\gamma^*)Q$ buffer space on each vehicle. Note that emergency shipments and safety stocks are not required if demand is deterministic which reduces the complexity significantly.

To overcome the unrealistically low service levels at the retailers in case one applies the expected value solution to our stochastic setting, we enhance the expected value solution by re-calculating the corresponding optimal shipment intervals and base-stock levels. We do that by considering the actual stochastic demand distributions while keeping the retailer clustering fixed. Although we re-optimize shipment intervals, truck capacity might still be insufficient leading to relatively high emergency shipment costs. This is caused by clusters with low shipment intervals (especially those with shipment interval 1), which can become infeasible with regards to the chance-constraint on truck capacity because these clusters would require a delivery more than once per period which we do not allow by assumption (note that $T_k \in \mathbb{N}_{\geq 1}$).

The results are presented in Table 5 and show that the adjusted expected value (AEV) solution, where shipment intervals and base stock levels are re-optimized, is on average 7.7 % worse than MIP-CG. Besides, 94.7% of the instances contain at least one cluster were the probability that truck capacity is sufficient is

below γ^*, which also leads to an increase of emergency shipment costs from 3.1% to 6.9% of the total costs. Hence, accounting for uncertainty in the demand distribution in our joint optimization problem is of crucial importance to control costs.

Table 5 Comparison of the Adjusted Expected Value (AEV) Solution and MIP-CG Solution Characteristics, Averaged over all Instances of the Benchmark Sets*

	MIP-CG	AEV
Average number of clusters	11.2	11.9
Average length of shipment intervals	1.5	1.6
Average cost increase	–	7.7%
Cost composition		
travel cost	77.9%	75.2%
emergency cost	3.1%	6.9%
holding cost	19.1%	17.9%
Average expected truck fill rate	80.5%	78.4%
Percentage instances with ≥ 1 infeasible cluster	–	94.7%

*: The instance categories with $N \geq 55, Q = 90$ were excluded.

We further study the impact of γ^* on the difference between the MIP-CG and the AEV. Note that an increase in γ^* reduces the expected amount of emergency shipments, because the probability that not all units fit on the truck $(1-\gamma^*)$ is reduced. This can be desirable from a practical perspective to avoid planning effort for organizing these emergency shipments. The results are presented in Figure 2. Comparing the total costs (left panel), we clearly see that MIP-CG outperforms the simple allocation of buffer capacity in AEV. It can be seen that the amount of buffer capacity has a small effect on the costs of MIP-CG because stochastic demand is directly taken into account in the planning process. Instead, under the AEV solution, planners have to set the right amount of buffer capacity to minimize costs. For high values of γ^*, emergency shipment costs under AEV increase as a result of insufficient buffer capacity, whereas for low values of γ^* AEV incurs high transportation cost as a result of inefficient routes. The results also show that our approach is able to increase truck capacity reliability up to 95% while similar costs are obtained. This might be relevant especially in the context of practical applications. In the right panel of Figure 2, we report the average load of the trucks with increasing γ^*. It confirms that it is crucial to consider uncertain demands already in the planning process because it increases average truck loads and reduces costs.

6.4. Impact of Different Cost Components.

We continue the analysis by providing further insights in the structure of the optimal solution for varying cost components. In Figure 3, we investigate the effect of ignoring holding and emergency costs. On the left of Figure 3, the structure of the optimal solution when incorporating holding and emergency costs is shown. The middle and right solutions correspond to setting $h = 0$ and $e = 0$, respectively. Zooming in on

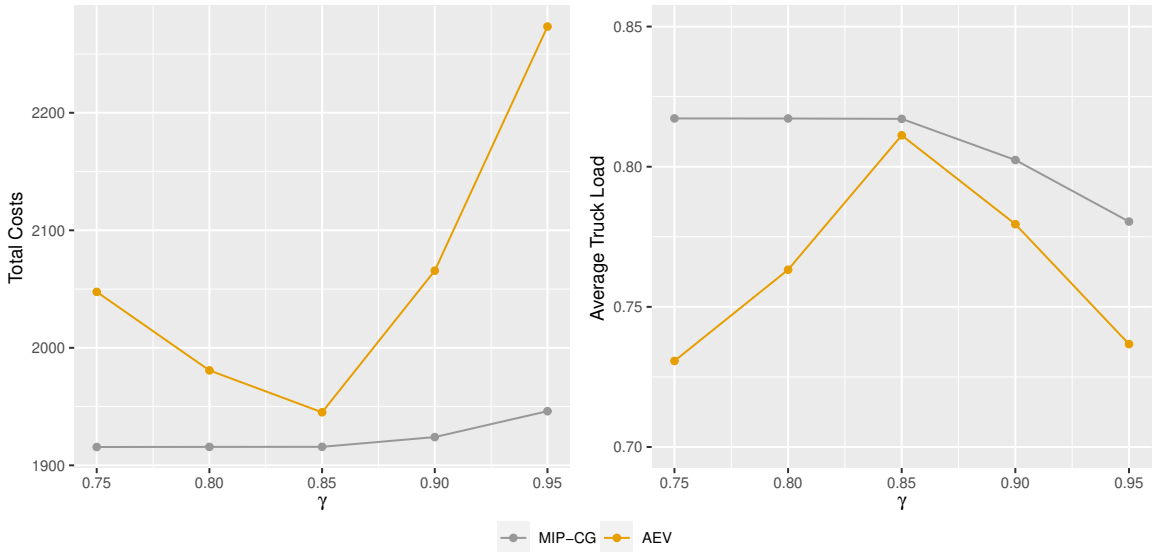


Figure 2 **Impact of varying values of γ and on the performance of MIP-CG and AEV**

the solution without holding costs (middle panel), we observe that direct deliveries become favorable for most retailers. This is because the reduction in fixed shipment costs achieved by replenishing retailers as infrequently as possible is higher than the effect of less variable shipment costs resulting from merging retailers into larger clusters. Moreover, when emergency shipment costs are ignored (right panel), we observe that retailer groupings increase. Note that the chance constraint on truck capacity still limits the size of the retailer groupings, and excess items not fitting on the truck are still delivered to the retailers by an external service provider but at zero costs. If excess items can be transported at zero costs, then less buffer space needs to be considered on the vehicle, which increases cluster sizes.

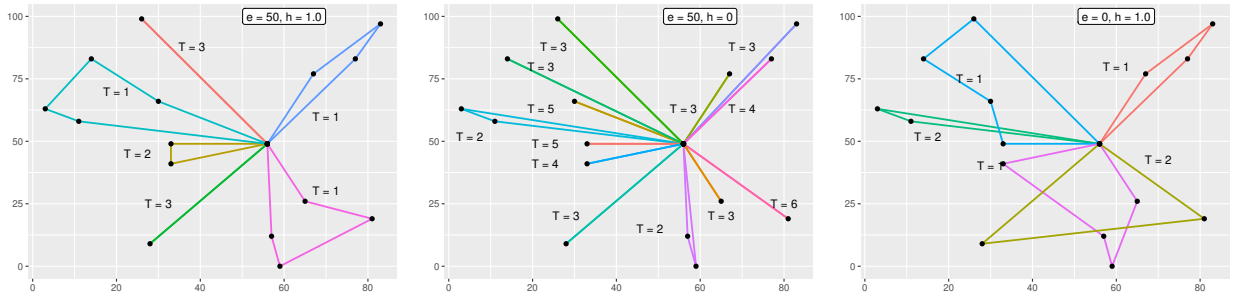


Figure 3 **Solution Structures of the Optimization Problem (left), in the Case Where Holding Costs Are Ignored (Middle), and in the Case Where Emergency Costs Are Ignored (Right)**

7. Conclusions and Future Research

We have presented a new approach to the stochastic inventory routing problem taking into account fixed replenishment intervals to groups of retailers mandated by, e.g., the retail industry to allow coordination with

up- and downstream planning operations. Both our optimal solution and our heuristic solutions balance fixed as well as variable transportation costs, emergency shipment costs, and holding costs by allocating retailers to clusters that are replenished in fixed intervals.

We transformed the naturally integer chance-constrained model to a linear integer model, which we solve using column generation. We have provided an exact branch-price-and-cut method and several upper bounding procedures that are also used as an initial set of columns for the exact method. The efficiency of the exact method relies on a labeling algorithm tailored toward our setting and utilizes an approximate stochastic knapsack solution as a pruning mechanism. Whereas the exact method is able to solve, depending on the chosen parameter values, instances with up to 60 retailers, the hybrid heuristic achieves a performance that is within 1% from optimality. This shows that for large-scale practical instances, the combination of our exact method (which can provide a lower bound) with the use of our heuristic methods shows an excellent performance that is provably near-optimal. An analysis of MIP-CG compared to the AEV solution has shown that it is important to include stochastic demand already in the planning process. Furthermore, we have illustrated how different cost components impact the structure of the optimal solution, which reveals the importance of considering all these cost components—transportation costs, emergency shipment costs, and holding costs—in the context of stochastic inventory routing problems.

The opportunities for further research are numerous. Many of these opportunities would not require structural adaptations of the presented model: for instance, one could consider delivery lead times that depend on the customer order in a tour and the corresponding transportation times, which requires modeling on a continuous time horizon. We believe that such extensions are valuable on its own, and our presented model can be used as a starting point.

More structural changes are required when, for instance, there is limited supply at the central warehouse, correlated demand between retailers and/or items are considered, vehicle tours are scheduled, or different shipment intervals in each retailer group are allowed. Limited supply at the central warehouse requires an allocation of available inventory to retailers as well as consideration of waiting times due to stock-outs at the central warehouse. Correlated demand increases the complexity of utilizing vehicle capacity and therefore also increases the complexity of clustering retailers into groups. In this paper, we have not scheduled vehicles and thus have assumed that each tour is performed by a different vehicle. By scheduling the tours, one can determine the optimal number of required vehicles, which is important if the shipments are performed by the company itself rather than by an external service provider. One can allow for different shipment intervals in each retailer group by, for example, considering so-called power-of-two policies described in the joint replenishment literature (see, e.g., Federgruen and Zheng 1992, Jackson et al. 1985), where each retailer is replenished in constant intervals that are power-of-two multiples of some base shipment interval. Thus, the number of parameters that needs to be optimized increases significantly.

References

Adelman D (2004) A price-directed approach to stochastic inventory/routing. *Operations Research* 52(4):499–514.

Aghezzaf EH (2008) Robust distribution planning for supplier-managed inventory agreements when demand rates and travel times are stationary. *Journal of the Operational Research Society* 59(8):1055–1065.

Aghezzaf EH, Raa B, Van Landeghem H (2006) Modeling inventory routing problems in supply chains of high consumption products. *European Journal of Operational Research* 169(3):1048–1063.

Aghezzaf EH, Zhong Y, Raa B, Mateo M (2012) Analysis of the single-vehicle cyclic inventory routing problem. *International Journal of Systems Science* 43(11):2040–2049.

Andersson H, Hoff A, Christiansen M, Hasle G, Løkketangen A (2010) Industrial aspects and literature survey: Combined inventory management and routing. *Computers & Operations Research* 37(9):1515–1536.

Axsäter S (2001) A note on stock replenishment and shipment scheduling for vendor-managed inventory systems. *Management Science* 47(9):1306–1310.

Bard JF, Huang L, Jaillet P, Dror M (1998) A decomposition approach to the inventory routing problem with satellite facilities. *Transportation Science* 32(2):189–203.

Barnhart C, Johnson EL, Nemhauser GL, Savelsbergh MW, Vance PH (1998) Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46(3):316–329.

Berman O, Larson RC (2001) Deliveries in an inventory/routing problem using stochastic dynamic programming. *Transportation Science* 35(2):192–213.

Bertazzi L, Bosco A, Guerriero F, Lagana D (2013) A stochastic inventory routing problem with stock-out. *Transportation Research Part C: Emerging Technologies* 27:89–107.

Bertazzi L, Laganà D, Ohlmann JW, Paradiso R (2020) An exact approach for cyclic inbound inventory routing in a level production system. *European Journal of Operational Research* 283(3):915–928.

Blumenfeld DE, Burns LD, Daganzo CF, Frick MC, Hall RW (1987) Reducing logistics costs at general motors. *Interfaces* 17(1):26–47.

Bramel J, Simchi-Levi D (1995) A location based heuristic for general routing problems. *Operations Research* 43(4):649–660.

Burns LD, Hall RW, Blumenfeld DE, Daganzo CF (1985) Distribution strategies that minimize transportation and inventory costs. *Operations Research* 33(3):469–490.

Campbell AM, Savelsbergh MW (2004) A decomposition approach for the inventory-routing problem. *Transportation Science* 38(4):488–502.

Campelo P, Neves-Moreira F, Amorim P, Almada-Lobo B (2019) Consistent vehicle routing problem with service level agreements: A case study in the pharmaceutical distribution sector. *European Journal of Operational Research* 273(1):131–145.

Çetinkaya S, Bookbinder JH (2003) Stochastic models for the dispatch of consolidated shipments. *Transportation Research Part B: Methodological* 37(8):747–768.

Çetinkaya S, Lee CY (2000) Stock replenishment and shipment scheduling for vendor-managed inventory systems. *Management Science* 46(2):217–232.

Çetinkaya S, Mutlu F, Lee CY (2006) A comparison of outbound dispatch policies for integrated inventory and transportation decisions. *European Journal of Operational Research* 171(3):1094–1112.

Cetinkaya S, Tekin E, Lee CY (2008) A stochastic model for joint inventory and outbound shipment decisions. *IIE Transactions* 40(3):324–340.

Chan LMA, Federgruen A, Simchi-Levi D (1998) Probabilistic analyses and practical algorithms for inventory-routing models. *Operations Research* 46(1):96–106.

Chen FY, Wang T, Xu TZ (2005) Integrated inventory replenishment and temporal shipment consolidation: A comparison of quantity-based and time-based models. *Annals of Operations Research* 135(1):197–210.

Chitsaz M, Divsalar A, Vansteenwegen P (2016) A two-phase algorithm for the cyclic inventory routing problem. *European Journal of Operational Research* 254(2):410–426.

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

Coelho LC, Cordeau JF, Laporte G (2014a) Heuristics for dynamic and stochastic inventory-routing. *Computers & Operations Research* 52:55–67.

Coelho LC, Cordeau JF, Laporte G (2014b) Thirty years of inventory routing. *Transportation Science* 48(1):1–19.

Coelho LC, Laporte G (2014) Optimal joint replenishment, delivery and inventory management policies for perishable products. *Computers & Operations Research* 47:42–52.

Costa L, Contardo C, Desaulniers G (2019) Exact branch-price-and-cut algorithms for vehicle routing. *Transportation Science* 53(4):946–985.

Crama Y, Rezaei M, Savelsbergh M, Woensel TV (2018) Stochastic inventory routing for perishable products. *Transportation Science* 52(3):526–546.

Diabat A, Archetti C, Najy W (2021) The fixed-partition policy inventory routing problem. *Transportation Science* 55(2):353–370.

Duffy M (2004) How gillette cleaned up its supply chain. *Supply Chain Management Review* 8(3):20–27.

Ekici A, Özener OÖ, Kuyzu G (2015) Cyclic delivery schedules for an inventory routing problem. *Transportation Science* 49(4):817–829.

Federgruen A, Zheng YS (1992) The joint replenishment problem with general joint cost structures. *Operations Research* 40(2):384–403.

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

Geetha S, Poonthalir G, Vanathi P (2009) Improved k-means algorithm for capacitated clustering problem. *INFOCOMP* 8(4):52–59.

Ghiami Y, Demir E, Van Woensel T, Christiansen M, Laporte G (2019) A deteriorating inventory routing problem for an inland liquefied natural gas distribution network. *Transportation Research Part B: Methodological* 126:45–67.

Gleixner A, Bastubbe M, Eifler L, Gally T, Gamrath G, Gottwald RL, Hendel G, Hojny C, Koch T, Lübbecke ME, Maher SJ, Miltenberger M, Müller B, Pfetsch ME, Puchert C, Rehfeldt D, Schlösser F, Schubert C, Serrano F, Shinano Y, Viernickel JM, Walter M, Wegscheider F, Witt JT, Witzig J (2018) The SCIP Optimization Suite 6.0. ZIB-Report 18-26, Zuse Institute Berlin, URL `http://nbn-resolving.de/urn:nbn:de:0297-zib-69361`.

Higginson JK, Bookbinder JH (1995) Markovian decision processes in shipment consolidation. *Transportation Science* 29(3):242–255.

Holzapfel A, Hübner A, Kuhn H, Sternbeck MG (2016) Delivery pattern and transportation planning in grocery retailing. *European Journal of Operational Research* 252(1):54–68.

Huang SH, Lin PC (2010) A modified ant colony optimization algorithm for multi-item inventory routing problems with demand uncertainty. *Transportation Research Part E: Logistics and Transportation Review* 46(5):598–611.

Hvattum LM, Løkketangen A (2009) Using scenario trees and progressive hedging for stochastic inventory routing problems. *Journal of Heuristics* 15(6):527.

Jackson P, Maxwell W, Muckstadt J (1985) The joint replenishment problem with a powers-of-two restriction. *IIE Transactions* 17(1):25–32.

Jaillet P, Bard JF, Huang L, Dror M (2002) Delivery cost approximations for inventory routing problems in a rolling horizon framework. *Transportation Science* 36(3):292–300.

Jepsen M, Petersen B, Spoorendonk S, Pisinger D (2008) Subset-row inequalities applied to the vehicle-routing problem with time windows. *Operations Research* 56(2):497–511.

Johansson L, Sonntag DR, Marklund J, Kiesmüller GP (2020) Controlling distribution inventory systems with shipment consolidation and compound poisson demand. *European Journal of Operational Research* 280(1):90–101.

Juan AA, Grasman SE, Caceres-Cruz J, Bektaş T (2014) A simheuristic algorithm for the single-period stochastic inventory-routing problem with stock-outs. *Simulation Modelling Practice and Theory* 46:40–52.

Kiesmüller G, De Kok A (2005) *A multi-item multi-echelon inventory system with quantity-based order consolidation* (Beta, Research School for Operations Management and Logistics).

Kleywegt AJ, Nori VS, Savelsbergh MW (2002) The stochastic inventory routing problem with direct deliveries. *Transportation Science* 36(1):94–118.

Kleywegt AJ, Nori VS, Savelsbergh MW (2004) Dynamic programming approximations for a stochastic inventory routing problem. *Transportation Science* 38(1):42–70.

Kumar A, Schwarz LB, Ward JE (1995) Risk-pooling along a fixed delivery route using a dynamic inventory-allocation policy. *Management Science* 41(2):344–362.

Lefever W, Aghezzaf EH, Hadj-Hamou K (2016) A convex optimization approach for solving the single-vehicle cyclic inventory routing problem. *Computers & Operations Research* 72:97–106.

Lübbecke ME, Desrosiers J (2005) Selected topics in column generation. *Operations Research* 53(6):1007–1023.

Malicki S, Minner S (2021) Cyclic inventory routing with dynamic safety stocks under recurring non-stationary interdependent demands. *Computers & Operations Research* 131:105247.

Marklund J (2011) Inventory control in divergent supply chains with time-based dispatching and shipment consolidation. *Naval Research Logistics* 58(1):59–71.

Moin NH, Salhi S (2007) Inventory routing problems: a logistical overview. *Journal of the Operational Research Society* 58(9):1185–1194.

Mutlu F, Çetinkaya Sil, Bookbinder JH (2010) An analytical model for computing the optimal time-and-quantity-based policy for consolidated shipments. *IIE Transactions* 42(5):367–377.

Raa B (2006) *Models and algorithms for the cyclic inventory routing problem*. Ph.D. thesis, Ghent University.

Raa B, Aghezzaf EH (2009) A practical solution approach for the cyclic inventory routing problem. *European Journal of Operational Research* 192(2):429–441.

Raa B, Aouam T (2021) Multi-vehicle stochastic cyclic inventory routing with guaranteed replenishments. *International Journal of Production Economics* 234:108059.

Raa B, Dullaert W (2017) Route and fleet design for cyclic inventory routing. *European Journal of Operational Research* 256(2):404–411.

Reiman MI, Rubio R, Wein LM (1999) Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Transportation Science* 33(4):361–380.

Schrotenboer AH, Ursavas E, Vis IFA (2019) A branch-and-price-and-cut algorithm for resource-constrained pickup and delivery problems. *Transportation Science* 53(4):1001–1022.

Schwarz LB, Ward JE, Zhai X (2006) On the interactions between routing and inventory-management policies in a one-warehouse n-retailer distribution system. *Manufacturing & Service Operations Management* 8(3):253–272.

Silver EA, Pyke DF, Peterson R, et al. (1998) *Inventory management and production planning and scheduling*, volume 3 (Wiley New York).

Solyalı O, Cordeau JF, Laporte G (2012) Robust inventory routing under demand uncertainty. *Transportation Science* 46(3):327–340.

Stenius O, Karaarslan AG, Marklund J, De Kok A (2016) Exact analysis of divergent inventory systems with time-based shipment consolidation and compound poisson demand. *Operations Research* 64(4):906–921.

Stenius O, Marklund J, Axsäter S (2018) Sustainable multi-echelon inventory control with shipment consolidation and volume dependent freight costs. *European Journal of Operational Research* 267(3):904–916.

Trudeau P, Dror M (1992) Stochastic inventory routing: Route design with stockouts and route failures. *Transportation Science* 26(3):171–184.

Ülkü MA, Bookbinder JH (2012) Optimal quoting of delivery time by a third party logistics provider: The impact of shipment consolidation and temporal pricing schemes. *European Journal of Operational Research* 221(1):110–117.

Van Anholt RG, Coelho LC, Laporte G, Vis IF (2016) An inventory-routing problem with pickups and deliveries arising in the replenishment of automated teller machines. *Transportation Science* 50(3):1077–1091.

Vansteenwegen P, Mateo M (2014) An iterated local search algorithm for the single-vehicle cyclic inventory routing problem. *European Journal of Operational Research* 237(3):802–813.

Zhao QH, Chen S, Zang CX (2008) Model and algorithm for inventory/routing decision in a three-echelon logistics system. *European Journal of Operational Research* 191(3):623–635.