A shortfall modelling-based solution approach for stochastic cyclic inventory routing

Birger Raa\textsuperscript{a,b,*}, Tarik Aouam\textsuperscript{c,d}

\textsuperscript{a}Department of Industrial Systems Engineering and Product Design, Ghent University
\textsuperscript{b}Antwerp Maritime Academy
\textsuperscript{c}Department of Business Informatics and Operations Management, Ghent University
\textsuperscript{d}Africa Business School, Mohammed VI Polytechnic University

Abstract

This paper studies the cyclic inventory routing problem with stochastic demand. A geographically dispersed set of retailers with stochastic demand rates is replenished from a single depot using vehicles with limited capacity. For an infinite horizon, a fixed-partition policy is adopted that partitions the retailers into subsets that are always replenished together in the same route being cyclically repeated. The objective is to provide cost efficient buffering of the demand variability within a cyclic distribution plan by providing carefully calibrated safety stock levels at the retailers. In doing so, the vehicle capacity needs to be taken into account, since cumulative demand during a cycle of the retailers in a route may exceed this capacity. In that case, shortfall remains at the retailer inventories because they are not fully replenished, which affects the service level (and cost balance) in the consecutive cycle(s). An approximate method is presented for determining the safety stock levels and is integrated into a state-of-the-art metaheuristic solution approach for cyclic inventory routing. An illustrative example and experiments on benchmark instances show (i) the effect of the vehicle capacity on the cost balance in a route, (ii) the accuracy of the approximation, and (iii) the added value of taking demand variability and shortfall due to limited vehicle capacity into account during the route design.

Keywords: Logistics, Distribution, Stochastic inventory routing, Cyclic planning, Safety stocks, Inventory shortfall

*Corresponding author: Technologiepark 46, 9052 Zwijnaarde, Belgium; tel: +32(0)9/264.5508
Email addresses: birger.raa@ugent.be (Birger Raa), tarik.aouam@ugent.be (Tarik Aouam)
1. Introduction

This paper studies the cyclic inventory routing problem (IRP) with stochastic customer demand. A fixed-partition policy (FPP) is adopted, meaning that customers are partitioned into subsets that are always replenished together in a dedicated route for each partition. Routes are repeated cyclically, i.e., with constant time between consecutive iterations. The cycle time of a route is optimized to balance the costs of making the route (loading, driving and unloading) with the inventory-related costs at the customers (expected holding and backlogging costs).

For the customers, being replenished by a cyclic route results in a periodic review, base-stock inventory control policy. Thus, a base-stock level has to be determined for each customer, up to which it should be replenished every time a delivery is made. However, since customer demand is stochastic, it could happen that the cumulative quantity for replenishing all customers in a route exceeds the capacity of the vehicle making the route. When this happens, the customers in the route start the next cycle with some shortfall, i.e., their inventory level after being replenished is still below their base-stock level. This results in a higher stockout probability in the next cycle. Therefore, the trade-off being made in determining customer base-stock levels between inventory holding costs and backlogging costs is distorted by the limited vehicle capacity via the occurrence of shortfall.

It is well-known that variability has to be buffered by providing slack in capacity, inventory and/or time. In the cyclic inventory routing problem with capacitated replenishments, the vehicle capacity is fixed and cyclic planning is adopted. This means that, in terms of capacity and time, the customer demand variability can be buffered by combining customers into a route and/or adjusting a route’s cycle time in such a way that the cumulative customer demand in the route per cycle is well below the vehicle capacity. Once a route and its cycle time are decided, customer demand variability must be buffered further by keeping safety stock at the customers, i.e., by increasing the base-stock level. The less slack there is in the vehicle capacity, the more likely it is that shortfall will occur and the more safety stock the customers need to keep. This problem of designing cyclic routes and deciding customer base-stock levels is the focus of this paper.

The remainder of this work is organized as follows. Section 2 presents a review of the relevant literature on cyclic inventory routing and on inventory shortfall. It also highlights the contributions this paper makes to the literature. Next, Section 3 provides a formal description of the problem and presents the derivations for optimizing base-stock levels. Section 4 presents an extensive illustrative example. The solution approach is explained in Section 5, followed by computational results using a set of benchmark instances in Section 6. Finally, conclusions and further research suggestions are given in Section 7.
2. Literature review

This work is the first to consider the issue of inventory shortfall in the context of the cyclic inventory routing problem. The relevant literature on these two topics is discussed here.

*Cyclic inventory routing*

In the literature, many variants of the inventory routing problem (IRP) are being studied, with varying characteristics such as the network structure, the customer demand pattern and the planning time horizon \[5, 13\]. Most papers consider a time horizon of a limited number of periods, while the literature on the infinite-horizon IRP is less extensive. As a result, IRP papers that consider stochastic demand are also more prevalent for the multi-period IRP. These papers are mainly based on a MILP formulation of the problem and then reformulate the stochastic problem either as a deterministic model in which stockouts probabilities are limited by using chance constraints \[30\], as a robust model with a budget-of-uncertainty approach \[27, 3\], or as a two-stage stochastic program \[4, 18\].

As mentioned above, this paper considers an infinite horizon rather and adopts a cyclic planning approach under a so-called ‘fixed partition policy’ (FPP) \[6\]. Most of the literature on this cyclic inventory routing problem assumes that the demand rates at the customers are constant. Aghezzaf et al. \[2\] and Raa and Aghezzaf \[20\] tackle the partitioning with a column-generation approach, where the columns correspond to routes. The subproblem of designing routes for a single vehicle is solved heuristically. This single-vehicle subproblem is being studied separately in much more detail as well \[32, 29\]. Another assumption that is often made, is that route cycle times have to be integers (e.g., a number of days). Powerful local-search based metaheuristic solution approaches for this discrete-time cyclic IRP are presented by Raa and Aghezzaf \[21\], Chitsaz et al. \[12\], Raa and Dullaert \[23\]. Some special cases of the cyclic IRP appear in the literature as well. E.g., Zenker et al. \[31\] study the problem complexity when all customers are located along a line, while Dai et al. \[14\] consider perishable products with price-dependent demand. Other papers study the cyclic inventory routing problem without adopting a fixed partition policy. Bertazzi et al. \[8\] study a deterministic, cyclic inventory routing problem for the collection of components to a manufacturing plant. They present a branch-and-cut algorithm with new valid inequalities. Bertazzi et al. \[7\] propose a matheuristic solution approach for the Periodic IRP, in which a route-based formulation of the problem is solved to optimality for a given subset of routes. They also show how to design effective subsets of routes.

Research on the stochastic cyclic IRP is very limited so far. Aghezzaf \[1\] presents a first attempt by considering uncertain demand rates and travel times. He tries to make the cyclic schedules more robust by imposing a service level constraint at the customers (e.g., less than 5% stockout...
probability per cycle) and by adding a chance constraint w.r.t. the route cycle times. This ensures, with a predetermined probability (e.g., 98%), that the total demand of the customers in a route during the duration of the route is less than the vehicle capacity. However, it does not specify what should be done in the remaining 2% of cycles where this constraint is violated. Solutions are evaluated using Monte-Carlo simulation and the safety stock levels are iteratively adjusted to reach the required service level if necessary. Given this computational burden, the approach cannot be applied to large instances within limited CPU times and is merely illustrated on an illustrative 12-customer example.

A recent paper by Raa and Aouam [22] also studies the cyclic IRP with stochastic demand. They explicitly consider the trade-off between inventory holding costs and backlogging costs at the customers. They deal with the limited vehicle capacity and the occurrence of shortfall by providing expedited shipments whenever customer demand in a cycle exceeds the vehicle capacity. As such, replenishments up to the base stock are guaranteed for all customers. In designing the routes, the cheapest way of providing expediting and the probability of requiring an expedited shipment (and incurring the corresponding cost) are taken into account and carefully balanced with routing, inventory holding, and backlogging costs. The state-of-the-art solution approach of [23] is adopted and fine-tuned and the impact of variability and the different cost parameters on the intricate cost trade-off is evaluated using extensive experiments.

By providing expediting shipments like Raa and Aouam [22], the occurrence of shortfall does not spill over to the next cycle(s). The effect of shortfall is thus absorbed by the backup capacity that the expediting provides. In the present paper, we aim to study how shortfall can be absorbed by providing additional buffering with inventory at the customers rather than with backup vehicle capacity, i.e., by increasing the base-stock levels. In this case, shortfall does spill over from one cycle to the next. To model this, the shortfall needs to be characterized further. Before doing so, the relevant literature on inventory shortfall is examined.

**Inventory shortfall**

In the inventory theory literature, different approaches to modelling shortfall emerge, using Markov chain modelling, using approximations based on asymptotic behavior, or using Monte-Carlo simulation.

Kijima and Takimoto [16] consider a periodic-review inventory/production model with a limited production capacity. They construct a Markov chain model of the inventory shortfall, and derive the time-dependent distributions of the leadtime and the customer waiting time. Similarly, Rappold and Muckstadt [24] adopt a Markov-chain approach to model the shortfall distribution in developing an approximate solution method for determining inventory levels in a capacitated two-echelon
production-inventory system. Markov Chain analysis is also adopt by Caceres et al. [11] when they evaluate the shortfall distribution in a multi-period production/inventory base-stock system with two suppliers, where demand sizes and supplier lead time are stochastic and correlated. Tang [28] uses insights from queueing theory and random-walk processes to provide a two-level iterative procedure for a discrete-demand shortfall distribution and then uses this to manage finished-goods inventory under capacitated postponement.

Based on the work of Glasserman [15] on the asymptotic behaviour of the shortfall distribution, Roundy and Muckstadt [25] first adopt a mass exponential distribution to approximate the shortfall distribution and then develop an improved, but computationally more demanding approximation that works extremely well as long as the coefficient of variation of demand is less than two. Acknowledging that the improved approximation of Roundy and Muckstadt [25] is computationally expensive, Betts [9] approximates the inventory shortfall distribution by sampling from a single simulation run and then sets target inventory levels using iterative search. This is an efficient alternative to simulation-based optimisation.

In this paper, we adopt the approximation of Roundy and Muckstadt [25] but generalize it to the situation where multiple customers being served in a single route are confronted with a joint capacity constraint, i.e., the limited loading capacity of the vehicle making the route. Furthermore, like Betts [9], we present an iterative search to optimize the base-stock levels. However, we present a customized iterative approach rather than a generic line search method.

**Contribution**

In summary, this paper makes the following contributions to the literature:

- It is the first paper to explicitly take into account inventory shortfall in the context of the cyclic inventory routing problem.

- Based on an approximation of the shortfall distribution adopted from the literature, an expression is derived for the expected cost rate of a set of customers served together in a cyclic route by a vehicle with a limited capacity.

- This derivation and the underlying demand distribution are used in a tailored iterative search procedure for setting the base-stock levels of the customers in a route.

- A state-of-the-art solution approach for cyclic inventory routing is adopted and extended (with this iterative procedure a.o.) to tackle stochastic customer demand and the resulting inventory shortfall.
The extended solution approach is applied to benchmark instances in computational experiments that confirm its strong performance and that result in relevant insights on the intricate cost trade-offs being made.

3. Formal problem description

The stochastic cyclic inventory routing problem with capacitated replenishments is described in detail in this section. We first introduce the assumptions and notation before formally defining the problem. Then, necessary derivations are done for determining the base-stock levels when there is shortfall.

3.1. Assumptions and notation

In the cyclic inventory routing problem, a set of geographically dispersed customers is replenished from a single depot over an infinite time horizon. The set of customers is denoted $C$ and is indexed by $i$ and $j$. A fixed-partition policy is adopted such that customers are partitioned into disjoint subsets and each subset is served in a separate route. The set of routes (or partitions) is denoted $R$ and indexed by $r$. The subset of customers in route $r$ is denoted $C_r$. Every route $r$ is repeated cyclically and has its own cycle time $T_r$. Thus, the time between consecutive visits to any customer $i \in C_r$ is always the same and given by the cycle time of the route, $T_r$. As a result, the customers adopt a periodic-review inventory policy, since they are only being replenished at periodic intervals. Route cycle times are discretized and have to be an integer number of days ($T_r \in \mathbb{N}_{>0}$).

Making a route once every $T_r$ days takes time and incurs costs. The duration and cost of making route $r$ (denoted $D_r$ and $F_r$, respectively) obviously depend on the sequence in which the customers are visited. Designing a route thus corresponds to finding the minimum-cost sequence and is therefore a Traveling Salesperson Problem on $C_r + \{0\}$ ($C_r$ plus the depot 0). There is a time $\tau_0$ and cost $\varphi_0$ for loading the vehicle at the depot, a time $t_{ij}$ and cost $c_{ij}$ for driving between locations $i$ and $j$, and a time $\tau_i$ and cost $\varphi_i$ for the delivery at each customer $i \in C_r$. The route duration $D_r$ is limited to the maximum working hours of a driver in one day, denoted $H$.

We denote the successor of node $i$ in the sequence by $s(i)$. (The first customer in the route is $s(0)$ and the depot is the successor of the last customer in the route.) The duration $D_r$ and the cost $F_r$
of making the route are then as follows:

\[ D_r = \sum_{i \in C_r} (\tau_i + t_{i,s(i)}) \]  \hspace{1cm} (1)

\[ F_r = \sum_{i \in C_r} (\varphi_i + c_{i,s(i)}) \] \hspace{1cm} (2)

Every customer observes a stochastic daily demand. The daily demand rate is assumed to be i.i.d. across days and to be normally distributed with a customer-specific mean \( \mu_i \) and standard deviation \( \sigma_i \). The periodic-review inventory policy means that a base-stock level \( S_i \) must be determined for every customer. At every visit, a replenishment up to this base-stock level is supposed to be made, such that demand for the next cycle is covered. In case a customer runs out of stock before being replenished, its demand is backlogged and fulfilled at the start of the next cycle. The base-stock level must be chosen such that expected inventory holding costs and expected backlogging costs are balanced. The expected average on-hand inventory at customer \( i \) in a cycle is denoted \( E[I_i] \) and the expected backlog level at the end of a cycle, just before being replenished, is denoted \( E[B_i] \). For the on-hand inventory, a holding cost per unit per period \( \eta \) is incurred. When there is a stockout, a backlog cost per unit short \( \beta \) is incurred. The inventory-related cost per day at customer \( i \) in route \( r \) is thus \( \eta E[I_i] + \beta \frac{1}{T_r} E[B_i] \). As a result, the total cost rate \( CR_r \) of route \( r \) is the following:

\[ CR_r = F_r \cdot \frac{1}{T_r} + \sum_{i \in C_r} \left( \eta \cdot E[I_i] + \beta \cdot \frac{E[B_i]}{T_r} \right) \] \hspace{1cm} (3)

This cost rate is minimized by choosing the right cycle time \( T_r \) and optimizing the base-stock levels \( S_i \) at the customers.

When a route is made, it might happen that the cumulative demand of the customers in that route exceeds the limited vehicle capacity, denoted \( \kappa \). When this happens, the customers in the route cannot be replenished up to their base-stock level and some shortfall remains after delivery. When there is shortfall at the beginning of a cycle, the inventory level is lower than it should be, and hence the risk of stocking out during the next cycle is higher. The shortfall thus disturbs the balance between holding costs and backordering costs.

To make up for the increased stock-out risk due to shortfall, the base-stock level should be increased compared to the situation where there is no shortfall. Without shortfall, the base-stock level can be determined for each customer separately, based on the distribution of a customer’s demand during a cycle. However, the cumulative demand of all customers in a route may result in shortfall and hence the base-stock level at a customer must be adjusted. It depends not only on the distribution of a customer’s own demand, but also on the distribution of this cumulative demand. This makes the base stock derivation much more complicated, as further explained below.
The assumption is made that, on the day a route is made, all customers in the route share their order quantity with the distributor, i.e., the difference between their base stock $S_i$ and their current inventory position $IP_i$: $q_i = S_i - IP_i$. If the cumulative replenishment quantity of all customers in the route exceeds the vehicle capacity $\kappa$, the distributor knows there will be a shortfall, denoted $y$: $y = \left[ \sum_{i \in C_r} q_i - \kappa \right]^+$. The actual replenishment quantities then have to be less than the order quantities and a fair-share policy for dividing the available vehicle capacity across the customers has to be applied.

An obvious policy is to spread the available vehicle capacity across the customers by adjusting the replenishment quantities as follows: $q_i \cdot \frac{\kappa}{\sum_{j \in C_r} q_j}$. However, another, more appropriate fair-share policy is introduced below, in which a certain fraction $\delta_i$ of the shortfall is allocated to each customer and the actual replenishment quantities are $q_i - \delta_i y$ (with $\sum_{i \in C_r} \delta_i = 1$).

The vehicle capacity imposes a maximum on the route’s cycle time $T_r \leq T_r^{\text{max}} = \kappa_0 / \sum_{i \in C_r} \mu_i$. The slack in the vehicle capacity $\kappa_0 - T_r \sum_{i \in C_r} \mu_i$ helps buffer demand uncertainty: the more slack, the lower the risk of shortfall.

Every $T_r$ periods, the route is made and each customer in the route is replenished. Since demands per period are i.i.d., and a cycle has length $T_r$, the customer demand per cycle has an average of $T_r \mu_i$ and a standard deviation of $\sqrt{T_r} \sigma_i$. To deal with demand variability and shortfall, safety stock is kept at the customers and the base-stock level is the average demand per cycle plus the safety stock: $S_i = T_r \mu_i + Z_i \sqrt{T_r} \sigma_i$, with $Z_i$ the so-called safety factor.

### 3.2. Shortfall derivations for a single customer

As explained above, possible shortfall needs to be buffered, which complicates the matter of finding the optimal base-stock level. For ease of notation, the derivation is done here for a single route with a single customer, hence the customer index $i$ and the route index $r$ can be omitted.

The customer demand per cycle is denoted $x$. It has probability density function $f(x)$, cumulative density function $F(x)$ and complementary CDF $F(x)$, with mean $\mu_X$ and standard deviation $\sqrt{T_r} \sigma_i$. The shortfall at the beginning of a cycle is denoted $y$ and has probability density function $g(y)$, with mean $\mu_Y$.

In any cycle, the starting inventory position of the customer, denoted $IP^0$, is $S - y$, and its ending inventory position, denoted $IP^T$, is $S - y - x$. We assume that $IP^0$ is always positive, such that backlog at the start of a cycle is always zero (i.e., backlog at the end of a cycle is always smaller than the vehicle capacity $\kappa$). The backlog at the end of a cycle, denoted $B$, depends on the demand per cycle $x$. If $x > S - y$, then $B$ is positive and a backlogging cost per unit of backlog $\beta$ is incurred.
The expected backlog at the end of a cycle is:

\[ E[B] = \int_0^\infty \left( \int_{S-y}^\infty (x - S + y) f(x) dx \right) g(y) dy \quad (4) \]

The on-hand inventory, given by the inventory position plus the backlog, incurs a holding cost per unit per day \( \eta \). The on-hand inventory at the start of a cycle, denoted \( I^0 \), corresponds to \( IP^0 \) because backlog is assumed to be zero at the start of a cycle, and the on-hand inventory at the end of the cycle, denoted \( I^T \), is \( IP^T + B \). The demand \( x \) occurs continuously throughout the whole cycle, so the average on-hand inventory during a cycle, denoted \( I \), is \( (I^0 + I^T)/2 \). The expected average on-hand inventory during a cycle is therefore:

\[ E[I] = E_Y \left[ E_X \left[ \frac{I^0 + I^T}{2} \right] \right] = E_Y \left[ E_X \left[ S - y - \frac{x}{2} + \frac{B}{2} \right] \right] = S - \mu_Y - \frac{\mu_X}{2} + \frac{E[B]}{2} \quad (5) \]

The customer thus incurs the following average cost per day:

\[ TC = \eta E[I] + \beta E[B] = \eta \left( S - \mu_Y - \frac{\mu_X}{2} \right) + \left( \frac{\beta}{T} + \frac{\eta}{2} \right) E[B] \quad (6) \]

The objective is to find the base-stock level \( S \) that minimizes this cost rate, i.e., the \( S \) for which:

\[ \frac{\partial TC}{\partial S} = \eta + \left( \frac{\beta}{T} + \frac{\eta}{2} \right) \frac{\partial E[B]}{\partial S} = 0 \]

Using

\[ \frac{\partial E[B]}{\partial S} = \int_0^\infty \left( \int_{S-y}^\infty -f(x)dx \right) g(y)dy = -\int_0^\infty F(S - y)g(y)dy, \]

the optimal base-stock level \( S^* \) thus satisfies:

\[ \int_0^\infty F(S^* - y)g(y)dy = \frac{\eta}{\frac{T}{2} + \frac{\eta}{2}} = \frac{2T\eta}{2\beta + T\eta}, \]

or

\[ \int_0^\infty F(S^* - y)g(y)dy = \frac{2\beta - T\eta}{2\beta + T\eta} \quad (7) \]

If \( S^* \) is chosen such that the probability of not stocking out during a cycle corresponds to the critical ratio \( \frac{2\beta - T\eta}{2\beta + T\eta} \), then the balance between holding and backlogging costs is optimized and the total cost rate is minimized. That probability of not stocking out is denoted the cycle service level (CSL).

If there is never any shortfall at the customer (i.e., \( y \equiv 0 \)), this simplifies to \( F(S^*) = \frac{2\beta - T\eta}{2\beta + T\eta} \) and the classical inventory theory result is obtained that the base-stock level should match the critical percentile of the demand per cycle distribution \( F(x) \), i.e., the percentile that corresponds to the ratio of backlogging and holding costs at the customer.
Derivation of approximate $E[B]$

In general, no closed formula is available for the convolution of the customer demand distribution $f(x)$ and the shortfall distribution $g(y)$ in Equation (4). Glasserman [15] showed that the complementary CDF of the shortfall distribution is asymptotically exponential. This inspired Roundy and Muckstadt [25] to approximate the shortfall distribution with a mass exponential function: the shortfall is zero with a probability $P_0$ and distributed exponentially with rate $\gamma$ when it is non-zero (which happens with probability $\overline{P}_0 = 1 - P_0$):

$$g(y) \approx \begin{cases} P_0, & y = 0 \\ \overline{P}_0 \gamma e^{-\gamma y}, & y > 0 \end{cases} \quad (8)$$

Assuming this (approximate) shortfall distribution, the expected shortfall $\mu_Y$ immediately follows:

$$\mu_Y = \frac{P_0}{\gamma} \quad (9)$$

Substituting the approximation of $g(y)$ from (8) into Equation (4) results in the following:

$$E[B] = P_0 \cdot E[B \mid 0] + \overline{P}_0 \gamma \cdot \int_0^\infty E[B \mid y] e^{-\gamma y} dy \quad (10)$$

where the notation $E[B \mid y]$ indicates the expected backlog for a given shortfall $y$.

The latter integral can be rewritten using integration by parts:

$$\int_0^\infty E[B \mid y] e^{-\gamma y} dy = \left[ E[B \mid y] \frac{e^{-\gamma y}}{-\gamma} \right]_0^\infty - \int_0^\infty \frac{d}{dy} E[B \mid y] \frac{e^{-\gamma y}}{-\gamma} dy$$

$$= \frac{1}{\gamma} E[B \mid 0] + \frac{1}{\gamma} \int_0^\infty \left( \int_{S-y}^\infty f(x) dx \right) e^{-\gamma y} dy \quad (11)$$

As a result

$$E[B] = (P_0 + \overline{P}_0) \cdot E[B \mid 0] + \overline{P}_0 \cdot \int_0^\infty \left( \int_{S-y}^\infty f(x) dx \right) e^{-\gamma y} dy$$

$$= E[B \mid 0] + \overline{P}_0 \cdot \int_0^\infty F(S-y) e^{-\gamma y} dy \quad (12)$$

The integral in equation (12) can be rewritten using integration by parts as well:

$$\int_0^\infty F(S-y) e^{-\gamma y} dy = \left[ F(S-y) \frac{e^{-\gamma y}}{-\gamma} \right]_0^\infty - \int_0^\infty \frac{d}{dy} F(S-y) \frac{e^{-\gamma y}}{-\gamma} dy$$

$$= \frac{F(S)}{\gamma} + \frac{1}{\gamma} \int_0^\infty f(S-y) e^{-\gamma y} dy \quad (13)$$

This gives us

$$E[B] = E[B \mid 0] + \mu_Y \left( F(S) + \int_0^\infty f(S-y) e^{-\gamma y} dy \right) \quad (14)$$
Approximate $E[B]$ and $S^*$ for normally distributed demand

As mentioned above, customer demand per cycle $x$ is assumed to be normally distributed with mean $\mu_X = T\mu$ and standard deviation $\sigma_X = \sqrt{T} \sigma$, so $f(x) = \frac{1}{\sigma_X} \varphi(z)$ and $F(x) = \Phi(z)$, with $z = \frac{x - \mu_X}{\sigma_X}$ the normalized value of $x$, $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ the standard normal distribution density function and $\Phi(z)$ the standard normal CDF. Remember that the base-stock $S = T\mu + Z\sqrt{T} \sigma$, with $Z$ the critical factor.

So far, our derivations are valid for any distribution $f(x)$ of the demand per cycle $x$. From now on, however, we assume that $x$ is normally distributed with mean $\mu_X$ and standard deviation $\sigma_X$, so $f(x) = \frac{1}{\sigma_X} \varphi(z)$ and $F(x) = \Phi(z)$, with $z = \frac{x - \mu_X}{\sigma_X}$ the normalized value of $x$, $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ the standard normal distribution density function and $\Phi(z)$ the standard normal CDF. The base-stock can then be written as $S = \mu_X + Z\sigma_X$, with $Z$ the so-called critical factor. This assumption of a normal demand distribution is very common in the literature, especially in the situation where a stationary demand over an infinite horizon is considered.

For normally distributed demand per cycle, the parameters of the approximate shortfall distribution are the following [17]: the probability of non-zero shortfall $\mathcal{P}_0 = \mathcal{F}(\kappa)/\mathcal{F}(\kappa)$ and the rate of the exponential distribution $\gamma$ is $\frac{2X}{\sigma_X}$, with $Z_K = \frac{\kappa - \mu_X}{\sigma_X}$ the normalized value for the vehicle capacity $\kappa$.

Using this assumption and notation, the remaining integral in (14) can be rewritten as follows:

$$\int_0^\infty f(S - y) e^{-\gamma y} dy = \frac{1}{\sqrt{2\pi} \sigma_X} \int_0^\infty \exp \left( -\frac{1}{2} \left( \frac{Z - y}{\sigma_X} \right)^2 + \frac{1}{2} (2\gamma y) \right) dy \tag{15}$$

The terms in the exponential can be rewritten as follows:

$$\left( \frac{Z - y}{\sigma_X} \right)^2 + 2\gamma y = \left( \frac{Z - y}{\sigma_X} - \gamma \sigma X \right)^2 - \gamma \sigma X (\gamma \sigma X - 2Z)$$

Plugging this back into (15), the following result is obtained:

$$\int_0^\infty f(S - y) e^{-\gamma y} dy = e^{\frac{1}{2} \gamma \sigma X (\gamma \sigma X - 2Z)} \frac{1}{\sigma_X} \int_0^\infty \varphi \left( \frac{Z - y}{\sigma_X} - \gamma \sigma_X \right) dy$$

$$= e^{\frac{1}{2} \gamma \sigma X (\gamma \sigma X - 2Z) } \Phi(\gamma \sigma X - 2Z) \Phi(Z - \gamma \sigma X) \tag{16}$$

As a result, the expected backlog per cycle is the following:

$$E[B] \quad = \quad E[B | 0] + \mu_Y \left( \Phi(\gamma \sigma X - 2Z) + e^{\frac{1}{2} \gamma \sigma X (\gamma \sigma X - 2Z) } \Phi(Z - \gamma \sigma X) \right) \tag{17}$$

For a normally distributed demand, the first term is as follows:

$$E[B | 0] = \int_S^\infty (x - S) f(x) dx = \sigma_X L(Z), \tag{18}$$
with \( L(Z) \) the standard normal loss function.

Putting all this together, the following final result for \( E[B] \) is obtained:

\[
E[B] = \sigma_X \left[ L(Z) + \frac{1}{2Z_K} \cdot \Phi(Z_K) \left( \Phi(Z) + e^{\frac{1}{2}\gamma\sigma_X\gamma\sigma_X - 2Z} \cdot \Phi(Z - \gamma\sigma_X) \right) \right]
\]  

(19)

The optimal safety factor \( Z^* \) is derived similarly, starting from Equation (7) and the approximation in (8):

\[
\frac{2T \eta}{2\beta + T \eta} = \int_0^\infty \Phi(S - y)g(y)dy
\]

\[
\approx P_0 \cdot \Phi(S^*) + P_0 \gamma \cdot \int_0^\infty \Phi(S - y)e^{-\gamma y}dy
\]

(20)

Using the results of Equations (13) and (16), this gives the following:

\[
\frac{2T \eta}{2\beta + T \eta} = P_0 \cdot \Phi(S^*) + P_0 \gamma \cdot \int_0^\infty f(S - y)e^{-\gamma y}dy
\]

\[
\Rightarrow \frac{2\beta - T \eta}{2\beta + T \eta} = \Phi(Z^*) - P_0 e^{2\gamma\sigma_X\gamma\sigma_X - 2Z^*} \cdot \Phi(Z^* - \gamma\sigma_X)
\]

(21)

3.3. Derivations for a multi-customer route

Above, the derivations were made for a single-customer route. Now let us consider the situation of a route \( r \) with a cycle time of \( T_r \) days that serves a set of customers \( C_r \), each with independently and normally distributed daily demand with mean \( \mu_i \) and standard deviation \( \sigma_i \).

Together, the customers in the route have a cumulative demand per \( T_r \)-day cycle that is normally distributed with mean \( \mu_r = \sum_{i \in C_r} \mu_i \) and standard deviation \( \sigma_r = \sqrt{T_r \sum_{i \in C_r} \sigma_i^2} \). The shortfall depends on this cumulative demand and is again approximated using Equation (8), with parameters \( P_0 = \Phi(Z_K)/\Phi(Z_K) \) and \( \gamma = \frac{2Z_K}{\sigma_r} \), where \( Z_K = \frac{\kappa - \mu_r}{\sigma_r} \).

The following fair-share policy is proposed for determining the replenishment quantities in case of a shortfall. Rather than using ratios based on the order quantities, delivery quantities are decided such that each customer ends up with a certain fraction \( \delta_i \) of the shortfall, or, in other words, such that each customer is replenished up to \( S_i - \delta_i y \) (with \( \sum_{i \in C_r} \delta_i = 1 \)). Every customer then has the same non-zero shortfall probability \( P_0 \), but a different rate \( \gamma_i = \frac{1}{\sigma_i^2} \gamma \).
For customer \( i \), the optimal safety factor \( Z^*_i \) is found from the extension of (21) to the multi-customer situation:

\[
\frac{2\beta - T_r \eta}{2\beta + T_r \eta} = \Phi(Z^*_i) - \mathcal{P}_0 e^{\frac{1}{2} \gamma_i \sigma_i (\gamma_i \sigma_i - 2Z^*_i)} \Phi(Z^*_i - \gamma_i \sigma_i)
\]  

(22)

If the ratios \( \delta_i \) are determined in such a way that all \( \gamma_i \sigma_i \) are the same, then it means that all customers will have the same \( Z^*_i \) and hence the same relative cost balance between inventory holding and backlogging costs. Furthermore, this has a nice benefit from a computational point of view, since the optimal safety factor only needs to be found once per route instead of once for every customer in a route. Since \( \gamma_i \sigma_i = \frac{\sigma_i}{\sigma_r} \gamma \), the ratios \( \delta_i \) should therefore be based on the customer demand standard deviations \( \sigma_i \):

\[
\delta_i = \frac{\sigma_i}{\sum_{j \in C_r} \sigma_j}, \forall i
\]  

(23)

Thus, all customers in route \( r \) should have base stock \( s_i = \mu_i + Z^*_r \sigma_i \), with the optimal safety factor \( Z^*_r \) such that it satisfies:

\[
\frac{2\beta - T_r \eta}{2\beta + T_r \eta} = \Phi(Z^*_r) - \mathcal{P}_0 e^{\frac{1}{2} \gamma_K \sigma_r (\gamma_K \sigma_r - 2Z^*_r)} \Phi(Z^*_r - 2Z'_{K_r})
\]  

(24)

in which \( Z'_{K_r} = \frac{\sum_{i \in C_r} \sigma_i}{\sigma_r} Z_K \).

4. Illustrative example

To obtain some insights on how the shortfall affects base-stock levels in cyclic inventory routing, we consider the three-customer example shown in Figure 1 that was also used by Raa and Aouam [22]. The figure shows the locations of customers and the depot, along with the travel costs between these four nodes.

This example has the following data. Daily demand at the customers is normally distributed with means 24, 18, and 36 units per day, respectively, and with a coefficient of variation of \( CV = 0.3 \). There is a single vehicle with a capacity \( \kappa \) of 400 units. The holding cost \( \eta \) is \( €0.22 \) per unit per day at each customer and the backlogging cost \( \beta \) is \( €10 \) per unit at each customer. The fixed costs of vehicle dispatching and customer replenishments are zero, i.e., \( \varphi_0 = \varphi_i = 0 \). One can easily verify that the best sequence to serve the three customers together in a single route is \([1, 2, 3] \) and results in a route cost \( F_r = €185 \).

After designing the route, its cycle time and total cost rate have to be determined. Since the cumulative average demand of the three customers is 78 units per day and the vehicle capacity is 400 units, the route cycle time \( T_r \) is limited to \( T_r^{max} = 5 \) days. For a cycle time of 5 days, the
critical percentile is $\frac{2\beta-T\eta}{2\beta+T\eta} = 89.57\%$. With uncapacitated replenishments, this corresponds to a safety factor $Z_r^\infty = 1.258$ and a loss function value $L(Z_r^\infty) = 0.0497$. The base-stock levels at the customers are then $S_1 = 120 + 20.25$, $S_2 = 90 + 15.19$, and $S_3 = 180 + 30.38$ units. Further, assuming customer daily demands are all independent, the cumulative demand per cycle in the route is $\mu_r = 390$ units on average, with a standard deviation of $\sigma_r = 31.4$ units. This is quite close to the vehicle capacity $\kappa = 400$ ($Z_K = 0.3181$ and $\gamma = 0.02024$), resulting in a shortfall probability $P_0 = 0.60$ and an expected shortfall $\mu_Y = 29.67$ units.

If we want to spread the shortfall across all three customers such that they maintain the same cost balance, i.e., with $\delta_i$ based on $\sigma_i$, then we obtain $\delta_1 = 0.3077$, $\delta_2 = 0.2308$ and $\delta_3 = 0.4615$.

Using Equation (24) with $Z_r = 1.258$, we find an actual CSL of only 73.51%, which is way below the critical percentile of 89.57%. To reach the critical percentile, $Z_r$ has to be increased to 2.191. This results in base-stock levels at the customers of $S_1 = 120 + 35.27$, $S_2 = 90 + 26.46$, and $S_3 = 180 + 52.91$ units and expected backlogs per 5-day cycle of 1.580, 1.185, and 2.369 units, respectively.

The resulting (approximate) total cost rate for a cycle time $T_r = 5$ days is then:

$$CR_r(5 \text{ days}) = \frac{1}{5}185 + 0.22 \cdot (390/2 + 114.64 - 29.67 + 5.13/2) + \frac{10}{5} \cdot (5.13) = €109.43 \text{ per day.}$$

If, instead of spreading the shortfall across all three customers, we let only customer 3 deal with the shortfall, then customers 1 and 2 stick to the safety factor $Z_r^\infty = 1.258$ and have base stock levels $S_1 = 120 + 20.25$ and $S_2 = 90 + 15.19$ units. To balance inventory holding and backlogging costs for customer 3 while absorbing the shortfall, however, its safety factor has to be increased
much more, to 3.826, giving a base stock level $S_3 = 180 + 92.40$. The expected backlogs per 5-day cycle are then 0.802, 0.601, and 5.153 units, respectively.

The resulting (approximate) total cost rate for a cycle time $T_r = 5$ days in this case is:

$$CR_r(5 \text{ days}) = \frac{1}{5} \cdot 185 + 0.22 \cdot (390/2 + 127.84 - 29.67 + 6.56/2) + \frac{10}{5} \cdot (6.56) = \mathsf{\varepsilon}115.33 \text{ per day.}$$

Not spreading the shortfall across the customers thus results in both higher overall safety stock levels (114.64 → 127.84) and higher overall backlogs (5.13 → 6.56).

If the route cycle time is set to be $T_r = 4$ days, the critical percentile increases to 91.57% and the cumulative demand per cycle in the route $\mu_r$ is only 312 units on average, with a standard deviation of $\sigma_r = 28.1$ units. This gives $Z_K = 3.13$, a shortfall probability of only 0.09% and hence a very limited expected shortfall $\mu_Y = 0.0039$ units. Since the expected shortfall is so small, its effect on cost rates and the requirement adjustment of the safety factor is negligible. The optimal safety factor is $Z_r = 1.377$ with a loss function value $L(Z_r) = 0.0386$. The resulting base-stock levels are $S_1 = 96 + 19.83$, $S_2 = 72 + 14.87$, and $S_3 = 144 + 29.74$ units. The expected shortages per cycle are 0.556, 0.417, and 0.834 units, respectively.

This gives a total cost rate of the route for a cycle time of 4 days that turns out to be lower than for a cycle time of 5 days:

$$CR_r(4 \text{ days}) = \frac{1}{4} \cdot 185 + 0.22 \cdot (312/2 + 64.45 - 0.0039 + 1.806/2) + \frac{10}{4} \cdot (1.806) = \mathsf{\varepsilon}99.46 \text{ per day.}$$

5. Solution approach

To solve the cyclic inventory routing problem with capacitated replenishments under a fixed-partition policy, each customer must be allocated to a single route, and each route must be optimized in terms of its cycle time and the base-stock levels of the customers in the route. The solution procedure adopted here builds on a state-of-the-art metaheuristic framework that was developed in earlier work [23, 22].

This section first explains the construction and improvement heuristics for the route design. Then, the procedure to optimize base-stock levels in a given route is presented, and finally, it is shown how the metaheuristic framework uses these building blocks.

5.1. Route design

A savings-based construction heuristic is adopted to generate initial solutions. In the savings-based heuristic, each customer starts in a separate route and routes are merged in descending order of
savings until no more feasible or cost-reducing merges are possible. When merging two routes, it is not merely the savings in driving time and cost that matters, but the entire cost rate of the route needs to be determined, which includes adjusting the route cycle time and the customer base-stock levels.

Once an initial solution is built, a local-search-based improvement phase is applied. Four classical local-search operators from the vehicle routing literature are adopted, namely, relocate, exchange, 2-opt, and 2-opt*. Again, the evaluation of any change to the solution requires reoptimizing route cycle times and customer base-stock levels. Each of these operators individually is applied exhaustively in a best-accept manner. The local-search phase iterates across the four operators as long as improvements are found.

5.2. Procedure for finding $Z^*_r$

Apart from partitioning the customers into routes and determining the sequence of customers in a route to minimize $F_r$ (and $D_r$), we also need to determine the base-stock levels for the customers in each route.

If we take a given $Z_r$-value and plug it into the right-hand-side of Equation (24), we find the cycle service level (CSL) that corresponds to this $Z_r$-value. Due to the shortfall, this CSL will be lower than if the vehicle capacity would be unlimited. Indeed, it can be seen that the first term of that r.h.s. is exactly the CSL without shortfall, so the second term is the correction to take shortfall into account.

Using Equation (24), the value for $Z^*_r$ can be determined using an iterative search algorithm. The $Z_r$-value is initialized with the uncapacitated $Z^\infty_r = \Phi^{-1}\left(\frac{2\beta - T_r \eta}{\sqrt{2\beta^2 + T_r \eta^2}}\right)$, which can be found in a standard normal table.

Since this initial $Z_r$ leads to a CSL that is too low, $Z_r$ must be increased. This is done as follows: we find the actual CSL corresponding to $Z_r$ from the right-hand side of Equation (24). For this actual CSL, we can find in the standard normal table to which $Z'_r = \Phi^{-1}(\text{CSL})$ this corresponds. This $Z'_r$ is of course smaller than $Z^\infty_r$ and the difference between $Z^\infty_r$ and $Z'_r$ is the step size we take to increase our $Z_r$.

E.g., suppose we aim for a critical percentile of 90%, corresponding to $Z^\infty_r = 1.28$. If the right-hand side of Equation (24) gives us an actual CSL of only 80% for $Z_r = 1.28$, then we know $Z_r$ must be increased. This 80% CSL corresponds to $Z'_r = 0.84$, which is 0.44 less than $Z^\infty_r$. This 0.44 is the amount with which we increase our $Z_r$ so it becomes 1.72 in the next iteration. If the right-hand side of Equation (24) next tells us that $Z_r = 1.72$ achieves an actual CSL of 87.5%, we have to
increase $Z_r$ a bit more. The 87.5% CSL corresponds to $Z_r' = 1.15$, which is getting closer to $Z_r^\infty$, but still 0.13 less, so we increase $Z_r$ to 1.85 for the next iteration, and so on, until $Z_r$ converges.

The closer we get to the desired cycle service level $\frac{2\beta - T_r \eta}{2\beta + T_r \eta}$, the closer $Z_r'$ will also get to $Z_r^\infty$, and hence the smaller the step size with which $Z_r$ is increased. As a result, we know that this procedure will converge. Some preliminary testing showed that it actually converges in fewer iterations (and hence using less computationally expensive evaluations of the right-hand side of Equation (24)) than other line search methods (such as the bisection method). The pseudocode for this iterative search for $Z_r^*$ is shown in Algorithm 1.

**Algorithm 1: Iterative procedure for $Z_r^*$**

<table>
<thead>
<tr>
<th>Initialize: $CSL^* = \frac{2\beta - T_r \eta}{2\beta + T_r \eta}$; $Z_r^\infty = \Phi^{-1}(CSL^*)$; $Z_r = Z_r^\infty$;</th>
</tr>
</thead>
<tbody>
<tr>
<td>do</td>
</tr>
<tr>
<td>$CSL_{app} = \Phi(Z_r) - P_0 e^{2Z_k}(Z_k - Z_r) \Phi(Z_r - 2Z_k)$; $Z_r' = \Phi^{-1}(CSL_{app})$; $Z_r = Z_r + (Z_r^\infty - Z_r')$;</td>
</tr>
<tr>
<td>while $CSL_{app} &lt; CSL^* - \epsilon$;</td>
</tr>
</tbody>
</table>

As explained above, the shortfall is spread across all customers using the $\delta_i$ ratios. These are based on the customer’s standard deviation of demand per cycle, such that all customers in a route obtain the same optimal safety factor, namely $Z_r^*$, and hence the same relative cost balance, leading to the optimal cost trade-off for the entire route. As a result, when evaluating a possible cycle time for a route (which happens all of the time during the construction and local search phases of the route design heuristics), the $Z_r^*$ value only needs to be calculated once.

5.3. Metaheuristic framework

The route design building blocks are embedded in a population-based metaheuristic. Individuals in the population are created using the savings-based construction heuristic, in which a randomized parameter is included to create diverse solutions. Then, the local search phase is applied to each individual. In the local search, the order in which the four operators are applied is randomized every time. Furthermore, within each operator, some GRASP-like randomization is also applied. When an improving move is found, its improvement potential is multiplied with a random value between 1 and 2. As such, the local search operators do not really apply best-accept, but also help find more diverse solutions.

In creating the next generation of individuals, or offspring, pairs of parent individuals are selected. Crossover then happens by keeping common arcs from the parents in the offspring solution, while all
nodes that do not share a common edge in the parents are put in a separate route in the offspring. The offspring solution thus consists of a copy of the common part of the parent solutions, with the remaining customers in separate routes. Then, nodes that do share a common edge in the parents are randomly removed with a 10% probability and reinserted in a separate route as well. This offspring is then optimized using the same savings-based construction and local-search-based improvement phases.

After an offspring is created and has been optimized, it is checked whether a copy of this solution is already present in the population. If this is the case, the offspring solution is subject to a mutation phase, in which a random number of customers (between 25% and 75%) is removed from the solution. After this mutation, the removed customers are reinserted using cheapest insertion and the local-search-based improvement phase is applied.

Since there is already a strong focus on intensification when optimizing the individuals, the crossover and mutation operators are quite disruptive. Furthermore, the population management also takes care of diversification in the search. The population size is limited to 30 individuals. In each generation, 30 offspring solutions are generated. The 5 best parents and 5 best offspring are retained for the next generation, along with the 20 offspring solutions that are most diverse relative to these 10 elite solutions and themselves. Diversity is measured in terms of common arcs in the solutions. When duplicates remain in the population, they are replaced with a newly generated solution. Also, when 100 generations go by without finding a new best solution, all but the current best solution are replaced with new solutions. Overall, the algorithm terminates after 500 generations without finding a new best solution or when the predetermined running time is reached (60 seconds in our experiments).

5.4. Monte-Carlo simulation

In our derivations, the shortfall distribution was approximated by the mass exponential distribution shown in Equation (8). To verify the accuracy of this approximation, Monte-Carlo simulation is performed.

There is no interaction between the different routes in a solution, so each route can be simulated separately for a large number of cycles. Since consecutive cycles are correlated because of the shortfall that spills over from one cycle to the next, values are not sampled every cycle, but once every certain interval. The sampling interval was fixed at 100 cycles [9].

The values being sampled are the average inventory during the cycle, the backlog at the end of the cycle, and, of course, the shortfall after replenishment. Based on these values, a sample of the route cost rate (and its constituent components) can be calculated. When the estimate of the route cost rate based on these samples has a high enough reliability, the simulation finishes.
If we denote the route cost rate based on sample $s$ with $\xi_s$ and there are $N$ samples, then the route cost rate estimate $\bar{\xi}$ is the average of the sample data and the relative error of the estimate $\epsilon$ is based on this average and on the standard deviation of the sample data $\sigma_\xi$:

$$\bar{\xi} = \frac{1}{N} \sum_{s=1}^{N} \xi_s$$

$$\sigma_\xi^2 = \frac{1}{N-1} \sum_{s=1}^{N} (\bar{\xi} - \xi_s)^2$$

$$\epsilon = \frac{1}{\sqrt{N}} \frac{\sigma_\xi}{\bar{\xi}}$$

If the relative error $\epsilon$ reaches a target accuracy, $\epsilon < 0.05\%$ in our experiments, or when one million samples have been reached, then the simulation finishes. To reduce the variance and hence limit the number of samples to achieve high enough accuracy, two antithetic simulation runs (in which the random numbers being generated are highly negatively correlated) are always performed together.

6. **Computational experiments**

6.1. **Benchmark instances**

For the computational experiments, we reused some of the benchmark instances of Raa [19]. A set of 20 instances is used, with 80 to 120 customers scattered randomly across a single depot, within a radius of 100 km. Customer demand rates are between 1 and 10 units per day and vehicles have a capacity of 100 units. In fact, the set of 20 instances contains two times the same 10 instances, but with a different holding cost $\eta$ (€0.08 vs. €0.80 per unit per day).

In our experiments, the daily demand rate of each customer is assumed to be i.i.d. across days and to be normally distributed with a customer-specific mean $\mu_i$ and standard deviation $\sigma_i$. A customer’s daily demand rate is also assumed to be independent from other customers’ demand rates. A customer served in route $r$ with cycle time $T_r$ thus has a demand per cycle that is normally distributed with an average of $T_r \mu_i$ and a standard deviation of $\sqrt{T_r \sigma_i}$. The cumulative demand per cycle in route $r$ with cycle time $T_r$ is normally distributed with an average of $T_r \sum_{i \in C_r} \mu_i$ and a standard deviation of $\sqrt{T_r \sum_{i \in C_r} \sigma_i^2}$.

The time and cost for loading and dispatching the vehicles are 30 minutes and 20 euro ($t_0 = 0.5h$, $\varphi_0 = €20$), while the time and cost for customer deliveries are 15 minutes and 10 euro ($t_i = 0.25h$, $\varphi_i = €10, \forall i \in C$). Euclidean travel distances are used. Vehicles drive at an average speed of 50 km/h and incur a cost of €1.2 per km. The route duration is limited to $H = 8$ hours.
As in Raa and Aouam [22], two more parameters are introduced for the case of stochastic customer demand:

- Demand variability: the coefficient of variation of daily demand (denoted CV) is assumed to be the same for all customers: $\sigma_i = CV \cdot \mu_i, \forall i \in C$. Different levels of variability, ranging from $CV = 0$ to $CV = 0.3$ are considered, with $CV = 0.3$ the default value.

- Backlogging cost: the backlogging cost $\beta$ is the same for all customers and takes a value of €10 (default) or €50 per unit short.

6.2. Experimental results

Impact of variability

In a first experiment, the impact of variability is studied. All 20 instances are solved for different levels of demand variability, with the coefficient of variation ranging from 0 to 0.3 in steps of 0.1. The backlogging cost $\beta$ has the default value of €10 per unit short.

Table 1: Overall impact of demand variability on the cost balance and vehicle loading rate

<table>
<thead>
<tr>
<th>CV</th>
<th>Total</th>
<th>Distr</th>
<th>Cycle-Inv</th>
<th>Safety-Inv</th>
<th>Backlog</th>
<th>Loading rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1400.20</td>
<td>991.74</td>
<td>408.46</td>
<td>0.00</td>
<td>0.00</td>
<td>96.78%</td>
</tr>
<tr>
<td>0.1</td>
<td>1466.96</td>
<td>1003.98</td>
<td>402.47</td>
<td>39.01</td>
<td>21.50</td>
<td>95.68%</td>
</tr>
<tr>
<td>0.2</td>
<td>1535.06</td>
<td>1019.21</td>
<td>394.54</td>
<td>78.44</td>
<td>42.87</td>
<td>94.06%</td>
</tr>
<tr>
<td>0.3</td>
<td>1603.84</td>
<td>1037.56</td>
<td>383.81</td>
<td>118.72</td>
<td>63.75</td>
<td>92.93%</td>
</tr>
</tbody>
</table>

Table 1 shows the results, averaged out across all 20 instances. With increasing variability, more buffering is required and a different cost balance is obtained. Overall, customers are visited more frequently, so distribution costs increase and cycle inventory costs decrease. At the same time, the safety inventory and backlogging costs increase at the same rate. For these instances, safety inventory costs remain just below double of the backlogging costs across all instances. It can also be seen that the average vehicle loading rate decreases slightly with increasing variability, such that there is a bit more slack in the vehicle capacity and hence lower shortfall. This reduced loading rate also explains the overall increase in delivery frequencies.

Impact of inventory-related cost parameters $\eta$ and $\beta$

The second experiment evaluates the influence of the cost parameters $\eta$ and $\beta$. As mentioned above, the set of 20 instances consists of two times the same ten instances, but with different holding costs: $\eta = €0.08/\text{unit/day}$ and $\eta = €0.80/\text{unit/day}$. For the default level of variability
\((CV = 0.3)\) the results for the instances with low and high holding cost are shown separately in Table 2, both for the default \(\beta = \text{€}10\) per unit short (corresponding to the results of Table 1) and the increased \(\beta = \text{€}50\) per unit short.

Table 2: Impact of backlogging and holding cost parameters

<table>
<thead>
<tr>
<th>(\beta)</th>
<th>(\eta)</th>
<th>Total</th>
<th>Distr</th>
<th>Cycle-Inv</th>
<th>Safety-Inv</th>
<th>Backlog</th>
<th>Loading rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.08</td>
<td>1083.46</td>
<td>834.58</td>
<td>172.80</td>
<td>57.28</td>
<td>18.80</td>
<td>92.38</td>
</tr>
<tr>
<td>10</td>
<td>0.80</td>
<td>2124.22</td>
<td>1240.54</td>
<td>594.82</td>
<td>180.15</td>
<td>108.69</td>
<td>93.49</td>
</tr>
<tr>
<td>50</td>
<td>0.08</td>
<td>1110.83</td>
<td>849.47</td>
<td>164.29</td>
<td>82.08</td>
<td>15.05</td>
<td>91.79</td>
</tr>
<tr>
<td>50</td>
<td>0.80</td>
<td>2278.28</td>
<td>1290.22</td>
<td>557.59</td>
<td>348.00</td>
<td>82.47</td>
<td>92.24</td>
</tr>
</tbody>
</table>

Comparing the results for high holding costs to those for low holding costs in Table 2, it can be seen that the solution approach succeeds in finding a completely different cost balance. On the one hand, inventories are reduced by replenishing customers much more frequently, which leads to a strong increase in distribution costs. On the other hand, the CSL at the customers is reduced, so more backlogging is accepted in order to reduce safety stocks.

The impact of an increased backlogging cost on the total cost rate is less pronounced, but the results again confirm that an optimized cost balance is obtained. When backlogging is more expensive, the CSL at the customers and hence the safety stocks are increased to reduce the risk of stocking out. In fact, backlogging is reduced so strongly that it contributes less to the total cost rate (\(\text{€}15.05\) vs. \(\text{€}18.80\) per day and \(\text{€}82.47\) vs. \(\text{€}108.96\) per day for low and high holding costs, respectively), even though the per unit backlogging cost is five times higher. Furthermore, delivery frequencies are slightly higher and a bit more slack is provided in the vehicle loading rates, which limits the risk of shortfall and the resulting increase in backlogging.

**Validation through simulation**

In our derivations, we adopted an approximation of the shortfall distribution. To verify the accuracy of this approximation, we also performed Monte-Carlo simulation on the results reported in Table 2 above.

The cost rates as predicted by the approximate method are compared to the cost rate estimates from the simulation. Remember that simulations are run until the accuracy of the cost rate estimate is within 0.05%.

A first accuracy measure \(\alpha_1\) expresses the deviation of the approximate result (denoted \(\tilde{\xi}\)) from the simulated result (denoted \(\bar{\xi}\)) as a fraction of (the more accurate) simulated result:

\[
\alpha_1 = \frac{\tilde{\xi} - \bar{\xi}}{\bar{\xi}}
\]
The total cost rate consists of multiple components. The cyclic planning approach means that the distribution costs and cycle inventory holding costs are predicted exactly when evaluating solutions. The approximation relates to the shortfall and its impact on safety stock and backlogging. Therefore, a second accuracy measure $\alpha_2$ is also reported that is based on the sum of safety inventory and backlogging costs only.

Table 3: Accuracy of the approximation

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\eta$</th>
<th>Approx</th>
<th>Simul</th>
<th>$\alpha_1$ (%)</th>
<th>$\alpha_2$ (%)</th>
<th>Sol/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.08</td>
<td>1083.46</td>
<td>1084.82</td>
<td>−0.13</td>
<td>−1.76</td>
<td>189</td>
</tr>
<tr>
<td>10</td>
<td>0.80</td>
<td>2124.22</td>
<td>2126.53</td>
<td>−0.11</td>
<td>−0.80</td>
<td>254</td>
</tr>
<tr>
<td>50</td>
<td>0.08</td>
<td>1110.83</td>
<td>1112.80</td>
<td>−0.18</td>
<td>−2.00</td>
<td>190</td>
</tr>
<tr>
<td>50</td>
<td>0.80</td>
<td>2278.28</td>
<td>2281.44</td>
<td>−0.14</td>
<td>−0.73</td>
<td>234</td>
</tr>
</tbody>
</table>

Looking at the $\alpha_1$-values reported in Table 3, the first and main conclusion is that the total cost rates are actually predicted very accurately and are only underestimated very slightly when using the approximation.

Considering the $\alpha_2$-values, it can be seen that the underestimation of the joint safety inventory and backlogging costs by the approximation is somewhat larger. However, since these cost components are relatively small compared to the distribution and cycle inventory costs, the overall underestimation remains acceptably small.

When the holding costs $\eta$ is low and the backlogging cost $\beta$ is high, this underestimation is stronger. This means that the accuracy of the approximation decreases with higher CSL, or, in other words, further down the right tail of the demand distribution.

The final column of Table 3 displays the number of different solutions (i.e., individuals in the population) that are evaluated per second during the solution process. It can be seen that fewer solutions are evaluated when the holding cost is low. This is also explained by the higher CSL, because the iterative search requires more iterations before converging when the $Z$-values are larger. Hence, evaluating a single solution takes longer.

To achieve higher accuracy throughout the solution approach, a computationally more demanding approach would be required, such as the improved approximation presented by Roundy and Muckstadt [25] or a simulation-based approach as in Betts [9].

6.3. Computational results and insights

To show the value of buffering shortfall with additional safety stock at the customers and taking shortfall into account throughout the route design, the following experiment is set up. Different
versions of the solution method are created that deal with shortfall differently. All versions are then applied to the benchmark instances and Monte-Carlo simulation is performed on all the resulting final solutions.

1. The first version simply ignores the shortfall, based on the naive assumption that shortfall effects are negligible. In this version, the impact of shortfall will inevitably result in more backlogging at the customers.

2. The second version imposes a chance constraint such that the shortfall probability is limited to a small value (5%). In this version, the impact of shortfall is buffered by providing slack capacity in the vehicles. This corresponds to the approach of Aghezzaf [1].

3. The third version deals with shortfall through expediting. Whenever shortfall occurs, an expedited shipment is dispatched, so replenishments at the retailers are guaranteed. A penalty cost for this expedited shipment (multiplied with the shortfall probability) is included. In this version, the impact of shortfall is buffered by providing backup capacity, i.e., vehicles that are available for expedited shipments. This corresponds to the approach of Raa and Aouam [22].

4. The fourth version ignores shortfall while optimizing the route design with the heuristics, but adjusts the base-stock levels of a solution before entering it into the population. This version buffers the impact of shortfall by providing more safety inventory at the customers, but only does so in a post-processing phase, after the route design has been finalized.

5. The fifth version is the ‘full version’, in which shortfall is taken into account at every step along the way. This version thus also buffers the impact of shortfall by providing more backlogging at the customers. Whenever a possible change to a route needs to be evaluated, all base-stock levels are updated as well with the iterative search procedure.

Table 4 shows the results for each of the five versions applied to the 20 benchmark instances. For each of them, the simulation results are compared to the cost rates that the solution method

<table>
<thead>
<tr>
<th>Version</th>
<th>Predicted</th>
<th>Simulated</th>
<th>Dev (%)</th>
<th>Distr</th>
<th>Cycle-Inv</th>
<th>Buffer</th>
<th>Loading rate</th>
<th>Sol/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>1577.78</td>
<td>1880.73</td>
<td>-16.42</td>
<td>1009.36</td>
<td>394.62</td>
<td>476.76</td>
<td>96.49</td>
<td>577</td>
</tr>
<tr>
<td>v2</td>
<td>1659.04</td>
<td>1659.75</td>
<td>-0.04</td>
<td>1096.67</td>
<td>389.37</td>
<td>173.70</td>
<td>84.70</td>
<td>309</td>
</tr>
<tr>
<td>v3</td>
<td>1671.22</td>
<td>1671.21</td>
<td>0.00</td>
<td>1083.19</td>
<td>388.56</td>
<td>199.45</td>
<td>85.30</td>
<td>320</td>
</tr>
<tr>
<td>v4</td>
<td>1613.68</td>
<td>1615.42</td>
<td>-0.11</td>
<td>1027.86</td>
<td>395.42</td>
<td>192.14</td>
<td>92.82</td>
<td>572</td>
</tr>
<tr>
<td>v5</td>
<td>1603.84</td>
<td>1605.68</td>
<td>-0.12</td>
<td>1037.56</td>
<td>383.81</td>
<td>184.30</td>
<td>92.93</td>
<td>221</td>
</tr>
</tbody>
</table>
predicts. Columns 5-7 show the cost rate breakdown into routing costs, cycle inventory holding costs, and buffering costs, which is the sum of safety stock holding costs and backlogging costs (and expediting costs for version 3).

As can be expected, version 1 results in the lowest predicted cost rate, since it ignores the effects of shortfall. Simulations show that the actual cost rates are underestimated by 15% to 45%, due to a major underestimation of the backlogging costs. It is thus obvious that the impact of shortfall is not negligible at all.

Version 2 does not calculate the cost effects of the shortfall, but imposes that there is no more than 5% probability of having shortfall in any route. Because of this low shortfall probability, its impact on the total cost rate is very limited, and it can indeed be observed that the predicted cost rates only underestimate the simulated cost rates very slightly. However, this version leads to solutions that are not necessarily the most cost-effective. The spare capacity results in a low loading rate (shown in column 8 of Table 4), and hence routing costs are higher compared to versions 4 and 5, while the cycle inventory levels are similar. The spare capacity does lead to lower buffering costs, but this saving is smaller than the increase in routing costs.

Version 3 deals with shortfall by providing expedited shipments. This version does not ignore shortfall or rely on approximations to determine cost rates, and therefore the predicted cost rates correspond to the simulated cost rates. However, the results show that expediting as a measure to tackle shortfall is not the most cost-effective approach either. As in version 2, routing costs are also higher compared to versions 4 and 5, while the cycle inventory levels are similar. However, the buffering costs – including expediting – are also higher in this version. This is explained by the fact that an expedited shipment is made even for the slightest shortfall. This incurs an additional delivery cost at one of the customers (€10 in our instances) along with additional driving costs for a vehicle (at €60 per hour). So, even if one would consider using the spare capacity in one route as backup capacity for another route, i.e., if one would dynamically adjust routes to avoid shortfall, it would still be relatively expensive. Furthermore, this dynamic re-routing would sacrifice the consistency of the fixed partition policy and would require advanced lookahead policies for deciding exactly how to redesign the routes, see, e.g., [10, 26]. The fact that expediting is relatively expensive also explains the low loading rates. It is then no surprise that holding some extra safety stock at a cost of only €0.8 or 0.08 per unit per day turns out to be a more cost-efficient shortfall buffer.

Version 4 ignores shortfall during route design and optimization, but evaluates its effect in post-processing. So, solutions that happen to have a limited cost impact of shortfall survive in the population, but the approach does not actively search for the best overall cost trade-off. Compared to avoiding shortfall (versions 2 and 3), this already leads to more cost-effective solutions. The accuracy of these results is in line with the validation results above. In fact, it is even slightly
better, again, because solutions that happen to be impacted less by shortfall tend to survive.

Finally, version 5 considers shortfall throughout the whole solution approach, in every step of the construction and improvement heuristics. These are the same solutions (for $\beta = 10$) as in Table 3. Thus, we can conclude that although it is computationally more demanding to consider shortfall and adjust base-stock levels at all times, it is definitely worthwhile as it eventually leads to more cost-effective solutions. Much better loading rates are achieved than in versions 2 and 3 and hence there are lower routing costs for similar cycle inventory levels. At the same time, the carefully balanced buffering costs remain relatively limited.

The final column of Table 4 illustrates the computational burden by displaying how many different solutions are evaluated per second in the different versions. As can be expected, this number is the highest for version 1, which completely ignores the shortfall. For version 4, it is only slightly lower, because it needs to do postprocessing of the routes only once per solution. Version 3 needs to keep track of which customer in a route should receive the expedited shipments if necessary (i.e., the one nearest to the depot), so it requires a bit more work throughout the route design and hence can evaluate fewer solutions per second. Although version 2 does not include additional computations to deal with shortfall, it produces fewer solutions per second. This may be counterintuitive, but can be explained by the fact that it often encounters promising solutions that have to be discarded if one or more of the routes has a more than 5% shortfall probability. Finally, version 5 can evaluate significantly fewer solutions per second than the other versions. However, as mentioned before, the additional computational effort of updating base-stock levels with the iterative search at every move of the heuristics pays off by leading to improved solutions.

7. Conclusion

This paper studies the cyclic inventory routing problem with stochastic customer demand and illustrates how the base-stock levels at the customers have to be adjusted in order to deal with the shortfall that results from the limited capacity of the vehicles making the replenishments.

An approximation of the shortfall distribution based on its asymptotic behavior is adopted and an iterative search procedure is presented to find the optimal safety stocks that balance expected inventory holding costs and backlogging costs at the customers in the case of normally distributed demand per cycle. This iterative procedure based on the approximate shortfall distribution is computationally far less demanding than simulation-based shortfall modelling. As a result, it can embedded in a powerful local-search-based metaheuristic solution approach for the cyclic inventory routing problem to obtain state-of-the-art solutions.
Computational experiments show that the solution approach indeed manages to find high-quality solutions in which the various routing-related and inventory-related cost components are adequately balanced, under varying circumstances, such as different levels of demand variability and different orders of magnitude of the underlying cost parameters. Furthermore, Monte-Carlo simulation of these solutions shows that the accuracy of the cost rates obtained by using the approximate shortfall distribution is very high.

Because the distributor cannot always replenish up to the base-stock levels and some shortfall remains in some cycles, the customers observe a stochastic service from the distributor, for which they have to provide additional safety stocks. As an alternative, the distributor can provide a guaranteed service to the customers by dispatching expedited shipments in the case of shortfall. In future work, a hybrid could be envisaged where limited shortfall is absorbed by safety stocks at the customers, while expedited shipments or dynamic adjustments to the routes are only made when shortfall is sufficiently large. This will of course make it even more complicated to optimize the routes, their cycle times and the appropriate customer base-stock levels.

Another complication that requires attention in future research is the situation in which the holding and backlogging cost parameters have customer-specific values and/or in which the customers have limited storage capacity. In that case, the cost trade-off at each customer is slightly different and it has to be investigated how shortfall can be cost-efficiently spread across the customers in a route.

A final suggestion for further research is to take the inventory at the distributor side into account. So far, supply at the distributor is assumed unlimited and no corresponding costs are considered. Extending the scope to include ordering/production and inventory costs at the distributor would mean that the different cyclic routes to the customers would have to be synchronized to balance the demands being placed on the distributor’s depot. Furthermore, an interesting aspect to be analyzed is how customer demand variability dissipates through such a two-echelon distribution system.


