Timing intermittent demand with time-varying order-up-to levels

Prak, Dennis; Rogetzer, Patricia

Published in:
European Journal of Operational Research

DOI:
10.1016/j.ejor.2022.03.019

Publication date:
2022

European Journal of Operational Research 303 (2022) 1126–1136

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing, Transportation and Logistics

Timing intermittent demand with time-varying order-up-to levels

Dennis Prak[a,b,*], Patricia Rogetzer[b]

[a] Department of Operations, University of Groningen, PO Box 800, 9700 AV Groningen, the Netherlands
[b] Department Industrial Engineering and Business Information Systems, University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands

ARTICLE INFO

Article history:
Received 22 March 2021
Accepted 9 March 2022
Available online 13 March 2022

Keywords:
Inventory
Intermittent demand
Demand interval
Order-up-to levels

ABSTRACT

Current intermittent demand inventory control models assume that the demand interval is memoryless: the probability of observing a positive demand does not depend on the time since the last demand occurred. Contrarily, several forecasting contributions suggest that demand intervals contain more distributional information. We find that the data of the M5 forecasting competition confirms this. Therefore, we propose an inventory control model that explicitly uses the full distributions of the demand sizes and intervals and thereby acknowledges that the probability of a demand occurrence may vary throughout the interval. To exploit this information, we also allow for time-varying order-up-to levels that flexibly adjust inventories according to the dynamic requirements. We derive the long-run average holding costs, non-stockout probability, order fill rate, and volume fill rate. Inspired by an analogy with multi-item inventory control models, we propose a greedy marginal-analysis heuristic to optimize the order-up-to levels, which we benchmark against the optimal solution on theoretical instances. In a simulation study on the M5 competition data we demonstrate this method's improved on-target service performance compared to that of traditional solutions. We furthermore show that target service levels can be achieved at significantly lower costs with time-varying than with fixed order-up-to levels.

1. Introduction

Intermittent demand patterns are characterized by periods with positive demand, alternated with (one, a few, or many) periods without demand. Intermittent demand used to be associated with only a few item types, such as spare parts (Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016). However, due to ever-growing assortments and numbers of product varieties, it is increasingly recognized that in many warehouses and distribution centers the majority of the stored products is sold irregularly (Doszyń, 2019; Nikolopoulos, 2021). Forecasting intermittent demand patterns is a notoriously difficult task and most often approached by means of so-called size-interval forecasting methods (starting with Croston, 1972). These separately forecast the size of a positive demand and the time between two periods with positive demand, also called the demand interval. Demand intervals have received significant attention in forecasting research, as we will further discuss in Section 2.1.

Contrarily, inventory control research for intermittent demand has until now devoted little attention to the demand interval. As we will further elaborate on in Section 2.2, compound Poisson processes are the standard choice for intermittent demand, although the (fixed) probability of a demand occurrence has also sporadically been modeled explicitly by using a compound binomial model for lead-time demand. However, both compound Poisson and compound binomial demand models have the property that the demand interval is assumed to be memoryless, i.e. the "demand occurrence probability" is the same in every period, irrespective of the time since the last demand occurred. This simplifies the mathematical analysis and leads to stationary order-up-to levels throughout the entire demand interval.

We find evidence that this assumption is too restrictive and moreover ignores an opportunity to make use of all features of the demand interval. In the Walmart data set of the M5 forecasting competition, we find that the vast majority of items show patterns in their demand interval that do not confirm the memorylessness assumption and therefore do not fit to compound Poisson or compound binomial demand models. Particularly, we discover that multiple demands occur shortly after each other on a much more frequent basis than such models suggest. We propose a more general demand model that allows for any demand size and demand interval distribution, and thus also for time-varying

* Corresponding author at: Department Industrial Engineering and Business Information Systems, University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands.
E-mail addresses: d.r.j.prak@utwente.nl (D. Prak), p.b.rogetzer@utwente.nl (P. Rogetzer).

https://doi.org/10.1016/j.ejor.2022.03.019

demand occurrence probabilities. We show how to estimate these probabilities from real data and use them to optimize an order-up-to inventory control policy under either a non-stockout probability, order fill rate, or volume fill rate constraint.

In line with the generalized demand interval distribution, our inventory model allows for time-varying order-up-to levels. This has an important advantage over existing, stationary order-up-to policies: inventories can be adjusted in anticipation of higher or lower probabilities of demand occurring in the future. Contrarily, under classical policies, the order-up-to level would always stay the same, even if the demand occurrence probability in the upcoming periods is very different from the past. In modern, responsive supply chains – where real-time demand information can be used to optimize inventories on-line – this leads to large service level improvements.

An interesting analogy exists between our proposed model and multi-item inventory control models: instead of setting order-up-to levels for different items to fulfill an overall objective, we set order-up-to levels for a single item at different periods in the demand interval. Inspired by this analogy, we propose a greedy marginal-analysis heuristic to optimize the order-up-to levels throughout the demand interval. We demonstrate our approach for several theoretical scenarios and thereafter empirically compare its performance to that of existing methods, using all 30,490 training time series of the M5 forecasting competition. Summarizing, our contribution is threefold: (1) we present a generalized demand and inventory model that allows demand occurrence probabilities and order-up-to levels to vary throughout the demand interval and show how to calculate long-run average holding costs and various service measures, (2) we present and benchmark a greedy marginal-analysis solution procedure to determine the order-up-to levels at every period in the demand interval, and (3) we use the M5 competition data set to empirically benchmark our proposed method's service level performance against that of existing methods.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature. Section 3 presents the data set and discusses goodness-of-fit tests of commonly-used demand distributions. Section 4 describes the inventory control setting and introduces the new, general demand model. Section 5 shows the calculation of holding costs and various service measures and describes how time-varying order-up-to levels can be determined. Section 6 demonstrates the policy on various theoretical scenarios and presents a sensitivity analysis. Section 7 compares the policy's on-target service performance to that of existing models on the M5 competition data and analyzes the added value of using time-varying order-up-to levels. Section 8 concludes the paper.

2. Literature review

We review the relevant literature in three main streams. First, we discuss intermittent demand forecasting and size-interval methods, which provide the rationale for our demand model. Then, we cover relevant previous work on intermittent demand inventory control. Finally, we list contributions to the statistics literature that are relevant for our analysis.

2.1. Intermittent demand forecasting

In supply chain optimization in general and demand forecasting in particular, the relevance of storing and utilizing "big data" is acknowledged. When demand time series are studied at higher granularity (e.g. at a daily instead of a weekly level), their levels of intermittency also increase. This makes intermittent demand foresting increasingly relevant, ultimately leading to the fact that the M5 forecasting competition in 2020 focused mainly on intermittent demand (Makridakis, Spiliotis, & Assimakopoulos, 2021).

Intermittent demand forecasting is notoriously difficult, because demand intervals as well as demand sizes have to be studied. The majority of the related literature consists of so-called size-interval methods that separately analyze these two components of the demand time series. Croston (1972) proposed the first such method, twice applying exponential smoothing. Syntetos & Boylan (2005) corrected the bias in Croston's method that Syntetos & Boylan (2001) discovered. Teunter, Syntetos, & Babai (2011) proposed an estimator that is updated in every period rather than only after a demand occurrence, in order to react to possible obsolescence. New methodological contributions to the size-interval forecasting literature are still made on a frequent basis. Prestwich, Tarim, Rossi, & Hnich (2014) and Babai, Dallery, Boubaker, & Kalai (2019), for instance, derived estimators with improved performance under obsolescence.

Point forecasts of the average demand size and interval give only partial information about the distribution of future demand. Willemain, Smart, & Schwarz (2004) argued that real demand intervals are often longer or shorter than would be expected based on only these point forecasts. They proposed a bootstrapping approach to model lead-time demand. Porras & Dekker (2008) proposed an alternative bootstrapping method that samples from overlapping blocks of periods with the length of the lead time, whereas Viswanathan & Zhou (2008) constructed lead-time demand by sampling from the demand intervals and demand sizes separately. Hasni, Aguir, Babai, & Jemai (2019) suggested other variations of bootstrapping methods which achieved higher cost-service efficiency on a large spare part data set. In the traditional size-interval forecasting literature stream, Pennings, Van Dalen, & van der Laan (2017) showed that by conditioning on the time since the last demand occurrence, and thereby actively anticipating the next demand arrival, forecasting accuracy could be improved. This indicates the importance of considering the full demand interval distribution rather than only its point forecast.

In line with the popularity of machine learning for general forecasting purposes, especially neural networks have also been applied for predicting intermittent demand. Whereas Gutierrez, Solis, & Mukhopadhyay (2008), Mukhopadhyay, Solis, & Gutierrez (2012), and Lolli, Gamberini, Regattieri, Balugani, Gatos, & Gucci (2017) found that neural networks achieve higher accuracy than classical methods, Kourentzes (2013) and Babai, Tsadiras, & Papadopoulos (2020) found mixed results. Jiang, Huang, & Liu (2021) achieved good performance and computation speed with an approach based on support vector machines. The best-performing entries to the M5 competition were combinations of various machine learning methods, especially neural networks and gradient boosting methods (Makridakis, Spiliotis, & Assimakopoulos, 2022). The good performance of these machine learning approaches confirms that demand time series contain more information than is captured by classical demand models. This was also recognized by Türkmen, Januschowski, Wang, & Cemgil (2021), who distinguished two typical demand interval patterns: "aging" – where it is unlikely that demands occur shortly after each other, but the probability of observing a new demand increases when the interval progresses – and the opposite, "clustering" – where it is highly likely that demands occur shortly after each other and the probability of observing a new demand decreases when the interval progresses.

2.2. Intermittent demand inventory control

If demand is intermittent, then (compound) Poisson models are the default choice in (theoretical and applied) literature. For example, Axsäter (2015) discussed the calculation of policy parameters under various service measures for – next to the normal and

gamma – the compound Poisson class of demand. Empirical work largely follows this selection by fitting normal, gamma, and compound Poisson distributions to intermittent demand (e.g. Snyder, Ord, & Beaumont, 2012; Syntetos, Babai, & Gardner, 2015; Teunter & Duncan, 2009; Turrini & Meissner, 2019). Syntetos, Babai, & Altay (2012), Lengu, Syntetos, & Babai (2014), and Turrini & Meissner (2019), amongst others, found that Poisson arrivals combined with geometric demand sizes perform well empirically.

Other authors have expressed the intermittent nature of demand by modeling period demand as a compound Bernoulli process: a demand occurs in a period with a fixed probability, and if it occurs, then it follows some probability distribution. Dunsmuir & Snyder (1989), Janssen, Heuts, & de Kok (1998), and Strijbosch, Heuts, & Van der Schoot (2000) used this approach to estimate the first two moments of lead-time demand, which is subsequently modeled by a gamma or mixed-Erlang distribution. In line with compound Bernoulli demand per period, Teunter, Syntetos, & Babai (2010) explicitly modeled lead-time demand with a compound binomial distribution to calculate order-up-to levels.

Larsen & Thorstenson (2008, 2014) showed how the order fill rate and volume fill rate can be calculated if demand is modeled as a general compound renewal process. However, they, and the authors mentioned before who studied specific demand classes, assumed that lead-time demand follows a fixed distribution that cannot vary depending on the period in which the lead time starts. This implies that the demand interval must be memoryless. Various authors in the forecasting literature (as discussed in Section 2.1) have suggested that this assumption may be overly simplifying and our empirical analysis of the M5 data (see Section 3) confirms this suggestion. In that spirit, we present an inventory model that does allow demand occurrence probabilities to vary throughout the demand interval and, accordingly, dynamically adjusts inventory levels. Our approach to calculating holding costs and service measures extends the modeling logic of Larsen & Thorstenson (2014) to this more general class of demand processes.

Our model shows an analogy with multi-item inventory control. Instead of setting order-up-to levels for different items, we set order-up-to levels at different periods in a single item's demand interval. Multi-item inventory models are notorious for being computationally demanding, as an exhaustive search over all possible inventory levels for all items quickly becomes infeasible. Various authors (such as Bijvank, Koole, & Vis, 2010; Graves, 1982; Prak, Saccani, Syntetos, Teunter, & Visintin, 2017; Teunter, 2006) therefore proposed and applied greedy marginal-analysis heuristics that iteratively select the item for which an increased inventory level yields the largest service benefit relative to the additional holding costs. Acknowledging this analogy, we will propose a similar heuristic to iteratively select the period in which to increase the inventory level until the desired service level is attained.

2.3. Statistics

Relaxing the assumption of a constant demand occurrence probability allows to fine-tune the inventory policy to the temporal state of the system. The concept of time-varying demand occurrence probabilities is known in the statistics literature as dependent Bernoulli trials. Modeling the joint distribution of multiple such trials is generally computationally intractable (Emrich & Piedmonte, 1991). The earliest statistics solution to the added complexity was a Poisson approximation (Chen, 1975), which interestingly coincides with the mainstream approach in the inventory control literature to model intermittent demand arrivals. Van der Geest (2005) used binary trees to model the number of demand occurrences during multiple subsequent periods, such as a lead time, leading to a binomial-like distribution. In our analysis we use a similar logic to calculate the joint distribution of the number of demand occurrences during the lead time and the state of the system after the lead time, which we subsequently combine with the demand size distribution to find a full specification of lead-time demand.

3. Data and goodness-of-fit of memoryless demand interval models

To motivate the need for a generalized demand model, we analyze the training data of the M5 forecasting competition (International Institute of Forecasters, 2022) and analyze the goodness-of-fit of standard demand models. The M5 forecasting competition is the most recent in a series of five forecasting competitions organized by the Makridakis Open Forecasting Center, this time focusing on intermittent demand time series and attracting 7092 participants (Makridakis et al., 2022). The training data set consists of 3049 different items in ten different Walmart stores located in three different US states, thereby comprising 30,490 time series in total. Given the important role of the M competitions in the development of new forecasting methods (Petropoulos & Makridakis, 2020), we believe that this data set is an important benchmark case not only for forecasting, but also for intermittent demand inventory control. In this section we briefly recap the data characteristics and then discuss how we assess the goodness-of-fit of memoryless demand processes.

We consider the 30,490 item-level time series, which contain daily sales of products in the categories food, household, and hobby. Our focus is – different from the M5 competition itself – explicitly not on forecasting or explaining item-/category level demand patterns. Contrarily, we aim to measure the goodness-of-fit of a class of demand processes on all item-level time series. Some time series show very long periods during which the item was not sold at all. We interpret this as an indication that the item was not in the assortment for at least a large share of these periods, for instance because it is a seasonal product, it was newly introduced, or it was taken out of the assortment. To avoid such non-selling periods being incorrectly considered as very long demand intervals, we pre-process the data by deleting for every item periods of 30 or more consecutive days without demand.

Table 1 shows the remaining time series lengths, as well as descriptive statistics on the mean, standard deviation, and coefficient of variation (CV) of the demand sizes and intervals, across all items. Demand sizes and intervals vary heavily in absolute numbers, but also in their CV. We conclude that this data set represents a broad spectrum of intermittent demand types and is therefore a representative test set for benchmarking intermittent demand models.

To assess the fit of classical demand models on the time series of this data set, we observe the following: under either the compound Poisson or compound binomial demand model, the occurrence of a positive demand in any period is Bernoulli distributed with some probability p. Under the compound Poisson model it holds that $p = 1 - \exp(-\lambda)$. The demand interval length then follows a geometric distribution with parameter p. After having estimated p for each item as the reciprocal of the mean demand interval, we compare the theoretical distribution of the demand interval to the actually observed demand intervals. We apply four different tests, using the R package XNomial (Engels, 2015): 1) the chi-squared test with p-values computed according to the asymptotic distribution, 2) the chi-squared test with p-values computed via Monte Carlo simulation, 3) the likelihood ratio test, 4) the multinomial probability of the observed outcomes. Whereas the chi-squared test is most widely applied, Engels (2009) argues that the likelihood ratio test is more reliable.

Table 2 contains the percentages of the 30,490 item-level time series for which – based on the observed demand intervals – the

Table 1
Descriptive statistics.

	Length (days)	Positive demand sizes			Demand intervals		
		Mean	Std. Dev.	CV	Mean	Std. Dev.	CV
Minimum	14	1.00	0.00	0.00	1.00	0.04	0.04
25% quantile	872	1.31	0.64	0.48	1.53	1.35	0.80
Mean	1247	2.47	1.62	0.59	3.04	2.80	0.90
75% quantile	1677	2.40	1.71	0.68	3.88	3.81	1.01
Maximum	1913	161.20	98.20	10.35	14.34	10.59	2.10

Table 2
Rejection of "fixed demand occurrence probability" hypothesis.

	Significance	
	95%	99%
Chi-squared with asymptotic p-values	73%	69%
Chi-squared with Monte Carlo p-values	69%	56%
Likelihood ratio	66%	58%
Multinomial probability	70%	61%
Average	70%	61%

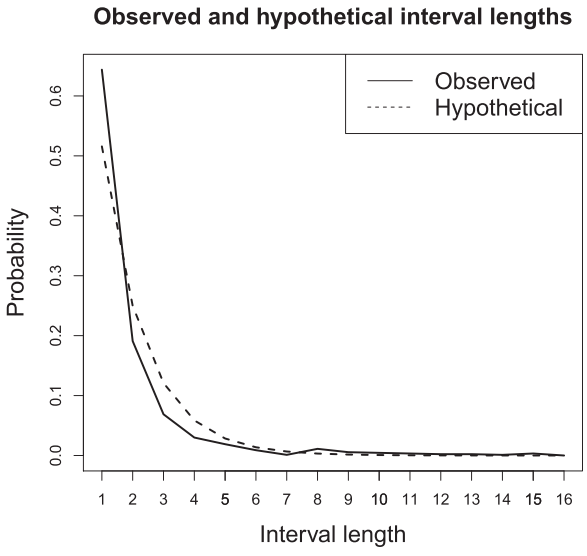


Fig. 1. Observed and hypothetical demand interval lengths for an exemplary item.

hypothesis of a fixed demand occurrence probability is rejected. Although the results vary slightly between the tests, the hypothesis is rejected for the majority of the time series. Taking the average over all tests, we find that for 70% (61%) of the items the demand occurrence probability does indeed vary throughout the demand interval with 95% (99%) certainty.

Figure 1 zooms in on one exemplary item and compares the hypothetical demand interval lengths under the memorylessness assumption with the actually observed demand interval lengths. The probability of an interval of length 1 (i.e. two subsequent periods with positive demand) is underestimated, whereas the probabilities of interval lengths 2 until 7 are overestimated. This corresponds to the "clustering" demand interval pattern in the classification of Tcfflrkmen et al. (2021). A similar pattern can be observed for many items. For 93% of the time series the probability of having positive demand in two subsequent periods is underestimated by a memoryless demand model. Motivated by this empirical evidence, we present our generalized demand and inventory model in the next section.

4. Model

We consider the standard base-stock inventory setting: a continuous-review, discrete-time inventory model with periods $t = 1, 2, \ldots$, a single location, and a single item which can be ordered or discarded without fixed costs. Orders arrive after a lead time of $L = 0, 1, 2, \ldots$ periods. Inventory holding costs $h > 0$ are incurred per unit and per period, and either a non-stockout probability, order fill rate, or volume fill rate should be satisfied. The order of events in a period is as follows: first, a new order is placed and outstanding orders (if any) arrive; second, holding costs are incurred; third, a possible demand occurs; fourth, service is evaluated. Table 3 gives an overview of the notation that is used throughout this paper. In the remainder of this section we describe our demand model and service measures.

4.1. Demand model

We assume that demands in different periods are independent. To model demand per period, we separate the demand occurrence probability and the size of a period demand given that it is positive. However, rather than assuming a constant demand occurrence probability, we allow this probability to vary throughout the demand interval. Let τ be the number of periods since the last period with positive demand. The probability that a demand occurs in the present period is p_τ. Hence, τ can be viewed as the state of a Markov Chain (MC). With probability p_τ the MC resets to state 1, and with probability $1 - p_\tau$ it moves forward to state $\tau + 1$. Without loss of generality we assume the existence of a maximum value τ' for τ for which $p_{\tau'} = 1$, and for all $\tau < \tau'$, $p_\tau < 1$. So, the MC is irreducible.

The long-run probabilities p^s_τ that the MC is in state τ can be found by solving

$$p^s_\tau = (1 - p_{\tau-1}) p^s_{\tau-1} \qquad \text{for } \tau = 2, 3, \ldots, \tau', \tag{1}$$

$$\sum_{\tau=1}^{\tau'} p^s_\tau = 1. \tag{2}$$

This gives (defining $p_0 = 0$ for completeness)

$$p^s_\tau = \frac{\prod_{i<\tau}(1 - p_i)}{\sum_{i=1}^{\tau'} \prod_{j<i}(1 - p_j)}. \tag{3}$$

Note that in the special case that $p_\tau = p$ for all τ, the probabilities p^s_τ reduce to those of a geometric distribution and we obtain the Bernoulli demand occurrence model of Teunter et al. (2010). Furthermore, under the assumption of (compound) Poisson demand with arrival rate λ per period, the demand occurrence probability $p_\tau = 1 - \exp(-\lambda)$ is also constant throughout the demand interval.

We denote total demand in a period, given state τ, by D_τ. Given that demand in a period is positive, its value D^+ has a distribution with probability mass function (pmf) f and cumulative distribution function (cdf) F. We assume that demand is discrete as this is most common in real life and also the case in our data set.

Table 3
Notation.

t	Period
L	Lead time
h	Holding cost per unit and per period
τ	Number of periods since the last demand, state of the Markov Chain
p_τ	Probability of a demand occurrence in state τ
τ'	Maximum number of periods between two demand occurrences, $p_{\tau'} = 1$
p^s_τ	Long-run probability that the Markov Chain is in state τ
D_τ	Stochastic demand in a period with state τ, $D_\tau \geq 0$
D^+	Stochastic demand in a period, conditional on a positive demand occurrence, $D^+ > 0$
f, F	Probability mass function (pmf) and cumulative distribution function (cdf) of D^+
f_n	n-fold convolution of f, pmf of the sum of n i.i.d. positive period demands
$f^N_{L,\tau}(n, \tau_p)$	Joint probability that, given lead time L and current state τ, n demands occurred during the lead time and the state before the lead time was τ_p
$f^D_{L,\tau}(d, \tau_p)$	Joint probability that, given lead time L and current state τ, the total demand during the lead time was d and the state before the lead time was τ_p
P^T	Probability of the sole feasible binary trajectory in Algorithm 1
P^T_i	Probability of feasible binary trajectory i in Algorithm 1
$D^{min}_{\tau,L}$	Minimum demand that must occur in L periods starting in state τ
e_i	ith unit vector
S_τ	Order-up-to level set in state τ
X	Net inventory after replenishment in some period
$\alpha_\tau(S_1, \ldots, S_{\tau'})$	Achieved non-stockout probability in state τ with order-up-to levels $S_1, \ldots, S_{\tau'}$
$\alpha(S_1, \ldots, S_{\tau'})$	Achieved (overall) non-stockout probability with order-up-to levels $S_1, \ldots, S_{\tau'}$
$\beta^o_\tau(S_1, \ldots, S_{\tau'})$	Achieved order fill rate in state τ with order-up-to levels $S_1, \ldots, S_{\tau'}$
$\beta^o(S_1, \ldots, S_{\tau'})$	Achieved (overall) order fill rate with order-up-to levels $S_1, \ldots, S_{\tau'}$
$\text{EFD}_\tau(S_1, \ldots, S_{\tau'})$	Expected fulfilled demand in state τ with order-up-to levels $S_1, \ldots, S_{\tau'}$
$\beta^v(S_1, \ldots, S_{\tau'})$	Achieved (overall) volume fill rate with order-up-to levels $S_1, \ldots, S_{\tau'}$
$H(S_1, \ldots, S_{\tau'})$	Expected overall holding costs with order-up-to levels $S_1, \ldots, S_{\tau'}$

4.2. Lead-time demand

The demand distribution is non-stationary throughout the order interval, because of the varying demand occurrence probabilities. Whereas stationary models allow to calculate the inventory level directly from the (single) order-up-to level and the stationary distribution of lead-time demand, we allow for order-up-to levels that vary per state and thus also need to account for the interplay between state transitions and lead-time demand. Specifically, our analysis in Section 5 requires the joint probability distribution of total demand over L periods and the state in which the MC was L periods ago, given its current state. In the special case where $p_\tau = p$ for all τ, the number of demand occurrences follows a binomial distribution with parameters L and p, and there is only one state. However, in the general case, such a compact formulation does not exist.

Denote by $f^N_{L,\tau}$ the joint pmf of the number of demand occurrences during the lead time and the state of the MC before the lead time, given state τ after the lead time. In the case $L = 0$, lead-time demand is obviously zero and the state remains τ. If $L > 0$, we suggest Algorithm 1 to find $f^N_{L,\tau}(n, \tau_p)$. Its logic is as follows: first, all feasible trajectories i of length L that lead from state τ_p to state τ, given n demand occurrences, are listed. If $n = 0$, then a feasible trajectory only exists if $\tau_p = \tau - L$. If $n > 0$, then feasible trajectories are those that (i) have at most $\tau' - \tau_p$ leading zeros (otherwise the trajectory cannot have started in τ_p), (ii) have exactly $\tau - 1$ trailing zeros (otherwise the trajectory does not end in τ), (iii) contain no sequence of at least τ' zeros (a demand must occur in state τ' at the latest), (iv) have exactly n ones. The probability of each feasible trajectory is found by multiplying the corresponding probabilities of (no) demand occurrences throughout the lead time trajectory. The sum of the probabilities of all feasible trajectories is multiplied by $p^s_{\tau_p}$ (as the system should be in state τ_p before the lead time) and divided by p^s_τ (as we condition on being in state τ after the lead time).

We still have to transform $f^N_{L,\tau}$ into $f^D_{L,\tau}$, the joint pmf of lead-time demand and the previous state. To that end, we recall that all positive period demands are i.i.d. with pmf f, so that the probability distribution of the sum of n positive period demands, with pmf denoted by f_n, is the n-fold convolution of f. Defining $f_0(0) = 1$, we can derive

$$f^D_{L,\tau}(d, \tau_p) = \sum_{n=0}^{L} f^N_{L,\tau}(n, \tau_p) f_n(d).$$

In Section 5 we show how to calculate various service measures and long-run average holding costs under this demand and inventory model, and discuss how the (state-dependent) order-up-to levels can be optimized.

Algorithm 1 Joint pmf of the number of demand occurrences and the state before the lead time.

Require: $\tau, \tau_p \in \{1, 2, \ldots, \tau'\}$, $L \in \mathbb{N}^+$, $n \in \{0, \ldots, L\}$
if $n = 0$ **then**
  if $\tau_p = \tau - L$ **then** $P^T = \prod_{t=\tau_p}^{\tau-1}(1 - p_t)$, the probability of the sole feasible trajectory
  else $P^T = 0$
  end if
else
  List all binary numbers of length L
  Discard all elements that have more than $\tau' - \tau_p$ leading zeros, not exactly $\tau - 1$ trailing zeros, any sequence of at least τ' zeros, or not exactly n ones
  for all binary numbers i **do** $p^T_i = 1$, the probability of trajectory i from τ_p to τ
    Set $\tau_n = \tau_p$
    for all positions n **do**
      if $n = 1$ **then** $p^T_i = p^T_i p_{\tau_n}$, $\tau_n = 1$
      else $p^T_i = p^T_i (1 - p_{\tau_n})$, $\tau_n = \tau_n + 1$
      end if
    end for
  end for
end if
return $f^N_{L,\tau}(n, \tau_p) = \frac{p^s_{\tau_p}}{p^s_\tau} \sum_i p^T_i$

5. Inventory policy analysis

In this section we first show how to calculate the non-stockout probability, order fill rate, volume fill rate, and long-run average holding costs for a given inventory policy. Thereafter, we discuss how to optimize the time-varying order-up-to levels. In all following calculations we assume the existence of a state-dependent order-up-to level S_τ, so that if the last positive demand was observed τ periods ago, the inventory position is raised to S_τ.

5.1. Non-stockout probability

We define the non-stockout probability α as the long-run probability that net inventory is non-negative at the end of an arbitrary period. Consider an arbitrary period where the MC is in state τ, in which the net inventory after the replenishment is X. The probability of completely fulfilling demand from on-hand stock in that period is then $(1-p_\tau)+p_\tau F(X)$. Using the order-up-to levels, the joint distribution of lead-time demand and the state before the lead time, we can derive that the non-stockout probability in state τ equals:

$$\alpha_\tau(S_1,\dots,S_{\tau'}) = \sum_{\tau_p=1}^{\tau'}\sum_{d=0}^{S_{\tau_p}}\left[(1-p_\tau)+p_\tau F(S_{\tau_p}-d)\right] f^D_{L,\tau}(d,\tau_p).$$

The overall non-stockout probability α is then found by taking the expectation over possible states τ:

$$\alpha(S_1,\dots,S_{\tau'}) = \sum_{\tau=1}^{\tau'} \alpha_\tau(S_1,\dots,S_{\tau'})\, p^s_\tau. \tag{4}$$

5.2. Order fill rate

The second service measure that we consider is the order fill rate β^o. It is defined as the long-run probability that an arbitrary (positive) demand can be fulfilled completely from on-hand stock. Let the current state be τ, let the net inventory level after replenishment be X, and let a positive demand occur in this state. This demand is fulfilled completely from on-hand stock with probability $F(X)$. The analysis now proceeds in analogy to Section 5.1. The achieved order fill rate in state τ is

$$\beta^o_\tau(S_1,\dots,S_{\tau'}) = \sum_{\tau_p=1}^{\tau'}\sum_{d=0}^{S_{\tau_p}} F(S_{\tau_p}-d) f^D_{L,\tau}(d,\tau_p).$$

The overall order fill rate is

$$\beta^o(S_1,\dots,S_{\tau'}) = \sum_{\tau=1}^{\tau'} \beta^o_\tau(S_1,\dots,S_{\tau'})\, p^s_\tau. \tag{5}$$

5.3. Volume fill rate

As a third service measure we consider the volume fill rate β^v, which is defined as the long-run fraction of a positive period demand that can be satisfied from on-hand stock. Again, let the period in which a demand occurs have state τ, and let the net inventory after the replenishment be X. Denote a (stochastic) positive demand size by D^+. The expected fulfilled demand in that period is

$$\mathrm{E}[\min(X, D^+)] = \sum_{i=1}^{X} i f(i) + X(1-F(X)).$$

Using again $f^D_{L,\tau}$ and the order-up-to levels S_τ, we find the expected fulfilled demand (EFD) in state τ:

$$\mathrm{EFD}_\tau(S_1,\dots,S_{\tau'}) = \sum_{\tau_p=1}^{\tau'}\sum_{d=0}^{S_{\tau_p}} f^D_{L,\tau}(d,\tau_p)\left[\sum_{i=1}^{S_{\tau_p}-d} i f(i) + (S_{\tau_p}-d)(1-F(S_{\tau_p}-d))\right].$$

Since the expected demand is $\mathrm{E}(D^+) = \sum_{i=1}^{\infty} i f(i)$, the overall volume fill rate is

$$\beta^v(S_1,\dots,S_{\tau'}) = \sum_{\tau=1}^{\tau'} \frac{\mathrm{EFD}_\tau(S_1,\dots,S_{\tau'})\, p^s_\tau}{\sum_{i=1}^{\infty} i f(i)}. \tag{6}$$

5.4. Holding costs

The objective is to minimize long-run average holding costs, which we derive in analogy to the service level calculations. As holding costs are incurred after the replenishment, but before the demand arrival, we evaluate the inventory level after L periods:

$$H(S_1,\dots,S_{\tau'}) = h\sum_{\tau=1}^{\tau'}\sum_{\tau_p=1}^{\tau'}\sum_{d=0}^{S_\tau} p^s_\tau(S_{\tau_p}-d) f^D_{L,\tau}(d,\tau_p). \tag{7}$$

Eq. (7) combines the steady-state probabilities of the τ' possible states of the MC with the expected inventory level after a lead time that ended in any of these states. As h is a constant that only scales the holding costs, we can set $h=1$ without loss of generality.

5.5. Inventory policy

After having computed the service measures and the holding costs for a given state-dependent order strategy, we can now optimize the order-up-to levels to achieve the service requirements with minimum holding costs. The complete inventory problem can be formulated as the following nonlinear integer program:

$$\begin{aligned} \text{Minimize} \quad & H(S_1,\dots,S_{\tau'}) \\ \text{subject to} \quad & \alpha(S_1,\dots,S_{\tau'}) \geq \alpha^* \text{ or } \beta^o(S_1,\dots,S_{\tau'}) \\ & \geq \beta^{o*} \text{ or } \beta^v(S_1,\dots,S_{\tau'}) \geq \beta^{v*} \\ & S_1,\dots,S_{\tau'} \text{ integer,} \end{aligned}$$

where any of the three service measures can be selected. Performing an exhaustive search to solve this program with τ' decision variables is only viable for relatively small instances with low values of τ' and/or low maximum period demand values. Acknowledging the existing analogy with the multi-item inventory control literature (see the discussion in Section 2.2), we suggest a greedy marginal-analysis heuristic solution procedure to optimize the time-varying order-up-to levels.

The solution procedure starts with order-up-to levels of 0 for any period τ, so that no holding costs are incurred. It then finds the period for which increasing the order-up-to level by 1 leads to the largest service level increase relative to the holding cost increase. In the case that no service level increase is achieved because of the minimum number of demands that occur during the lead time, the order-up-to level is increased with this minimum demand, so that an improvement is made in every iteration. Once the service level is at least equal to the threshold, the process reverses. In a similar greedy way order-up-to levels are decreased, every time selecting the period with the largest holding cost saving relative to service level loss. The last found solution that still satisfies the service level threshold is the solution of the heuristic. Algorithm 2 describes the procedure, where e_i is the ith unit vector.

6. Demonstration and sensitivity analysis

This section serves two purposes. Firstly, we aim to find insights into the optimal inventory policy by studying it for some typical

Algorithm 2 Heuristic solution procedure.

Require: service level target γ^*, achieved service function $\gamma(S_1,\ldots,S_{\tau'}) \equiv \alpha$ or β^o or β^v

for all τ **do** $D^{min}_{\tau,L} = \lfloor \frac{L+\tau-\tau'-1}{\tau'} \rfloor + 1$
end for
Set $S \equiv (S_1,\ldots,S_{\tau'}) = (0,0,\ldots,0)$
while $\gamma(S) < \gamma^*$ **do**
  for all τ **do**
    if $S_\tau = 0$ **then** $S^{new} = S + e_\tau D^{min}_{\tau,L}$
    else $S^{new} = S + e_\tau$
    end if
    $\mathrm{Incr}(\tau) = (\gamma(S^{new}) - \gamma(S))/(H(S^{new}) - H(S))$
  end for
  $\tau^* = \mathrm{argmax}(Incr)$
  if $S_{\tau^*} = 0$ **then** $S = S + e_{\tau^*} D^{min}_{\tau,L}$
  else $S = S + e_{\tau^*}$
  end if
end while
while $\gamma(S) > \gamma^*$ **do**
  for all τ **do**
    $S^{new} = S - e_\tau$
    $\mathrm{Decr}(\tau) = (H(S) - H(S^{new}))/(\gamma(S) - \gamma(S^{new}))$
  end for
  $\tau^* = \mathrm{argmax}(Decr)$
  $S = S - e_{\tau^*}$
end while
$S = S + e_{\tau^*}$
return S

Table 4
Scenario 1: "aging", $L = 1$.

Service	Order-Up-To Levels		Cost Diff.
	Optimal	Heuristic	
$\alpha = 0.80$	2,5,5,3,5	2,5,5,4,1	0.83%
$\alpha = 0.95$	5,5,5,5,5	5,5,5,5,5,5	**0%**
$\alpha = 0.99$	6,8,10,9,10	6,9,9,8,9	0.55%
$\beta^o = 0.80$	5,5,6,6,7	5,5,6,6,7	**0%**
$\beta^o = 0.95$	6,8,10,9,10	7,8,9,8,8	1.54%
$\beta^o = 0.99$	8,10,10,10,10	9,9,10,10,9	0.80%
$\beta^v = 0.80$	4,5,5,7,6	4,5,6,5,5	0.64%
$\beta^v = 0.95$	6,7,8,8,7	6,7,8,8,7	**0%**
$\beta^v = 0.99$	8,9,9,9,8	8,9,9,9,8	**0%**

Table 5
Scenario 2: "aging", $L = 5$.

Service	Order-Up-To Levels		Cost Diff.
	Optimal	Heuristic	
$\alpha = 0.80$	9,9,10,10,8	9,9,10,10,8	**0%**
$\alpha = 0.95$	12,12,13,13,15	12,13,12,13,11	1.05%
$\alpha = 0.99$	14,15,16,17,17	15,15,15,15,15	0.26%
$\beta^o = 0.80$	11,11,11,13,13	11,11,12,11,11	0.13%
$\beta^o = 0.95$	14,14,14,16,14	14,14,15,14,13	0.28%
$\beta^o = 0.99$	16,17,18,18,16	16,17,18,18,16	**0%**
$\beta^v = 0.80$	10,10,11,13,12	10,11,11,10,8	0.73%
$\beta^v = 0.95$	13,13,14,15,17	13,14,14,13,11	1.01%
$\beta^v = 0.99$	16,16,16,18,17	16,16,17,16,16	0.22%

Table 6
Scenario 3: "clustering", $L = 1$.

Service	Order-Up-To Levels		Cost Diff.
	Optimal	Heuristic	
$\alpha = 0.80$	7,5,0,5,6	7,5,0,5,6	**0%**
$\alpha = 0.95$	9,7,6,5,8	9,7,6,5,8	**0%**
$\alpha = 0.99$	10,9,8,5,9	10,9,8,5,9	**0%**
$\beta^o = 0.80$	7,6,5,5,8	7,6,5,5,8	**0%**
$\beta^o = 0.95$	9,8,6,5,9	9,8,6,5,9	**0%**
$\beta^o = 0.99$	10,9,9,5,9	10,9,8,6,10	0.06%
$\beta^v = 0.80$	6,6,5,4,8	7,4,1,4,7	6.5%
$\beta^v = 0.95$	8,7,7,5,9	8,8,6,5,7	0.57%
$\beta^v = 0.99$	9,9,9,8,10	9,9,9,8,10	**0%**

Table 7
Scenario 4: "clustering", $L = 5$.

Service	Order-Up-To Levels		Cost Diff.
	Optimal	Heuristic	
$\alpha = 0.80$	17,17,17,17,20	18,15,13,15,16	0.65%
$\alpha = 0.95$	21,20,19,19,22	21,20,19,19,22	**0%**
$\alpha = 0.99$	24,23,21,21,23	24,23,21,21,23	**0%**
$\beta^o = 0.80$	18,17,15,16,18	18,17,15,16,18	**0%**
$\beta^o = 0.95$	21,22,21,20,24	22,20,18,18,19	0.54%
$\beta^o = 0.99$	24,24,22,22,24	24,24,22,22,24	**0%**
$\beta^v = 0.80$	17,17,15,16,19	18,14,13,15,17	1.88%
$\beta^v = 0.95$	21,19,18,18,21	21,19,18,18,21	**0%**
$\beta^v = 0.99$	23,24,23,22,24	24,22,21,20,21	1.38%

parameter settings. The optimal policy is found by full enumeration of all possibilities in sufficiently small problem instances. Secondly, we evaluate the heuristic optimization procedure by comparing its solution and corresponding holding costs with the optimum. We study two typical demand interval patterns that correspond to the classification by Türkmen et al. (2021): "aging" (with demand occurrence probabilities $p_1 = 0.2$, $p_2 = 0.4$, $p_3 = 0.6$, $p_4 = 0.8$, and $p_5 = 1$) and "clustering" (with demand occurrence probabilities $p_1 = 0.8$, $p_2 = 0.6$, $p_3 = 0.4$, $p_4 = 0.2$, and $p_5 = 1$). For each pattern we consider a scenario with a short lead time ($L = 1$) and a long lead time ($L = 5$). Within each scenario, we define three service level targets (80%, 95%, and 99%) for each of the three different service measures that we study (α, β^o, and β^v). In all scenarios, the demand sizes are uniformly distributed on the integers $1,\ldots,5$. Tables 4–7 showcase for every scenario and service level target the order-up-to levels according to the optimal solution, the heuristic solution, and the percentage cost difference between both solutions.

In the "aging" scenarios (Tables 4 and 5), many (but not all) optimal policy patterns have order-up-to levels that either monotonically increase during the interval, or first increase and then decrease towards the end of the interval. In the case with $L = 1$ (see Table 4), order-up-to levels are lower than in the case with $L = 5$ (see Table 5), as in the latter case a longer lead time needs to be covered. Similarly, higher service targets require higher order-up-to levels. Low order-up-to levels directly after the last demand are explained by the fact that the probability of a new demand occurrence is lowest at that time, and hence also a lead time that starts directly after a previous demand will likely have fewer demand occurrences than a lead time that starts later in the interval. However, after the first demand in the lead time has occurred, the MC resets. Therefore, for a longer lead time (such as in the case with $L = 5$) the effect of its starting point on the total lead-time demand is smaller and consequently, the variation of the order-up-to levels (relative to their overall sizes) is lower.

For explaining the further solution patterns, we have to consider three other effects: firstly, not every period contributes equally to the achieved service and realized costs as not every period is equally likely to occur. Secondly, as is common in inventory problems, costs increase supralinearly in the service target. This makes it beneficial to achieve slightly higher service in periods where it is cheaper to achieve and lower service in periods where it is more costly. Thirdly, for discrete demand a target service level typically cannot be achieved exactly, so that one solution may outperform another mainly because of a smaller "service surplus." The sum of all these effects may be ambiguous. A good example of this

policy structure ambiguity is the optimal solution in scenario 2 for $\alpha = 0.80$. By construction, a demand must occur in the fifth period and a lead time that starts in that period will therefore certainly include that demand. Yet, it is optimal to set the order-up-to level in period 5 slightly lower than in all other periods. Contrarily, in the same scenario, but for $\alpha = 0.95$, the order-up-to level is set higher in period 5 than in all other periods. Considering all "aging" scenarios, the heuristic finds the optimal solution in 6 out of 18 cases. In 9 other cases the cost difference is less than 1%, whereas the maximum error is 1.54%. 17 out of 18 cases have time-varying order-up-to levels in optimum.

In the "clustering" scenarios (Tables 6 and 7) we observe the opposite: order-up-to levels are relatively high early in the interval, as demands are likely to occur shortly after each other. Order-up-to levels then decrease, but typically increase again towards the end of the interval. The latter is due to the "end effect": all scenarios must have $p_5 = 1$ in order to keep the intervals short enough to find the optimal solution by full enumeration. Similar to the aging scenarios, order-up-to levels are higher for longer lead times and/or higher service levels, whereas they are more variable for short lead times. For example, in the case $L = 1$ and $\alpha = 0.80$ it is optimal to keep no inventory at all in period 3, whereas in earlier and later periods order-up-to levels of 5, 6, and 7 units should be set. Considering all "clustering" scenarios, the heuristic finds the optimal solution in 11 out of 18 cases. In 4 other cases the cost difference is less than 1%. However, one scenario shows a cost difference of 6.5%. All cases have time-varying order-up-to levels in optimum.

Summarizing, we find that 35 out of 36 cases have time-varying order-up-to levels in optimum, which confirms the usefulness of a general policy that allows for these. The variation is largest if the lead time is short compared to the expected demand interval. This showcases that if the supply chain is agile, so that the inventory policy can quickly respond to the actual state of the system, the benefit of actually using the state information of the system is largest. The heuristic finds the optimal solution in 17 out of 36 cases and is within 1% cost difference in 30 out of 36 cases. The average performance loss is 0.55%. In the next section we examine the performance gain that can be achieved with this model and the heuristic solution procedure over fixed order-up-to-level policies with classical distributional assumptions.

7. Empirical results

In this section we analyze the inventory performance of the model and heuristic solution procedure on the data set described in Section 3. First, we compare – in a simulation experiment on the entire data set and for several given target service levels – its achieved service with that of a standard base-stock system with two commonly-used demand models. Then, we zoom in on one specific item of the data set and show the advantage of using time-varying order-up-to levels, measured by the inventory costs required to achieve a given target service level. We restrict attention to the non-stockout probability service level (α) for brevity, remarking that the procedures work completely analogously for the order fill rate and volume fill rate.

7.1. On-target service performance on the full data set

We consider as benchmarks the normal lead-time demand model, which is (sometimes implicitly) used in many applied papers, and the Poisson-geometric as most popular compound Poisson demand process. For every item of the data set described in Section 3, we fit the respective distributions to the demand time series by estimating the parameters as follows: for the normal distribution, we use the sample mean and standard deviation. For the Poisson-geometric distribution, we use the Method-of-Moments estimators given in Axsäter (2015). For the newly presented model, we use the empirical probabilities for the demand size and demand interval distribution. The latter probabilities p_τ^s are transformed into the demand occurrence probabilities p_τ by inverting the recurrent relationship of Eq. (1):

$$p_1 = p_1^s$$

$$p_\tau = \frac{p_\tau^s}{\prod_{i<\tau}(1-p_i)} \qquad \text{for } \tau = 2, \ldots, \tau'.$$

Subsequently, we calculate for all items the order-up-to levels to satisfy a given target service level under the fitted distributions. For the normal and Poisson-geometric benchmarks the order-up-to level is fixed throughout the entire horizon, whereas the newly proposed method will set time-varying order-up-to levels $S_1,\ldots,S_{\tau'}$ throughout the demand interval, calculated with the heuristic solution procedure. We select lead times of 1 and 5 days and consider various non-stockout probability targets between 50% and 99%. Then, we find the achieved service in a simulation experiment by applying the policy in every period and evaluating the inventory levels after subtracting the observed lead-time demand.

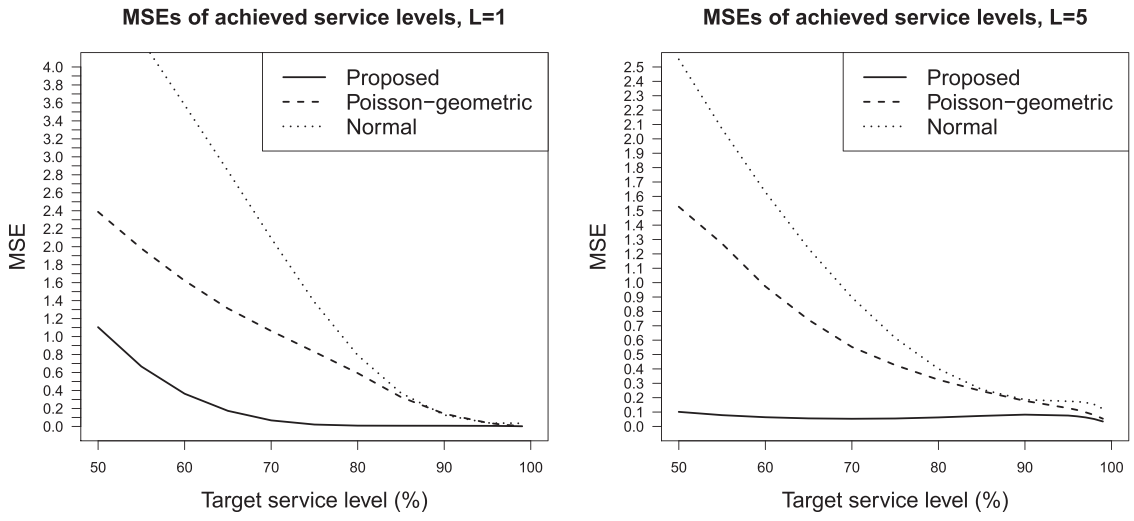
We calculate the Mean Squared Errors (MSEs, across all items) of the achieved service levels from their targets. This symmetric error measure avoids that an underachievement for one item is offset by an overachievement for another item and therefore provides a fair judgment of overall performance (see e.g. Prak, Teunter, Babai, Boylan, & Syntetos, 2021). Fig. 2 shows the results for the entire data set. Fig. 2a presents the results for $L = 1$, Fig. 2b corresponds to the case $L = 5$, and Figs. 2c and d zoom in on service levels between 95% and 99%. The first observation is that the proposed model achieves closer to the target over the entire range of service levels and for both the short lead time of 1 day and the long lead time of 5 days. This indicates that it is indeed beneficial for inventory performance to explicitly model the demand interval and size distributions rather than assuming a standard arrival process or lead-time demand model.

The normal lead-time demand model performs worst for almost all settings, except for service levels between 90% and 95%, where it slightly outperforms the Poisson-geometric model. All MSEs naturally decrease when the target service level increases, as higher achieved service levels are less sensitive to inventory differences. Therefore, the MSEs also converge to each other as the service level increases. Best visible for $L = 5$ in Fig. 2b, the normal and proposed model exhibit an S-shape, indicating that performance slightly deteriorates for higher service levels. For the proposed model, this can likely be attributed to the heuristic solution procedure which deviates stronger from the optimum for higher service levels. For the normal model it may be a result of the misfitting distribution shape.

All MSEs are significantly larger for $L = 5$ than for $L = 1$, as the period which the order-up-to level has to account for is also larger in that case. The largest improvement by the proposed method is achieved for $L = 1$, because a shorter lead time implies a more responsive system and thus a larger potential for time-varying order-up-to levels. Indeed, as $L \to \infty$, the effect of timing the intermittent demand diminishes and the optimal solution of the proposed model converges to a stationary order-up-to level. An important secondary finding is therefore that, to make optimal use of this highly responsive model, the inventory system should be designed in an agile way, so that items can be quickly reordered or replaced to where they are expected to be demanded in the near future.

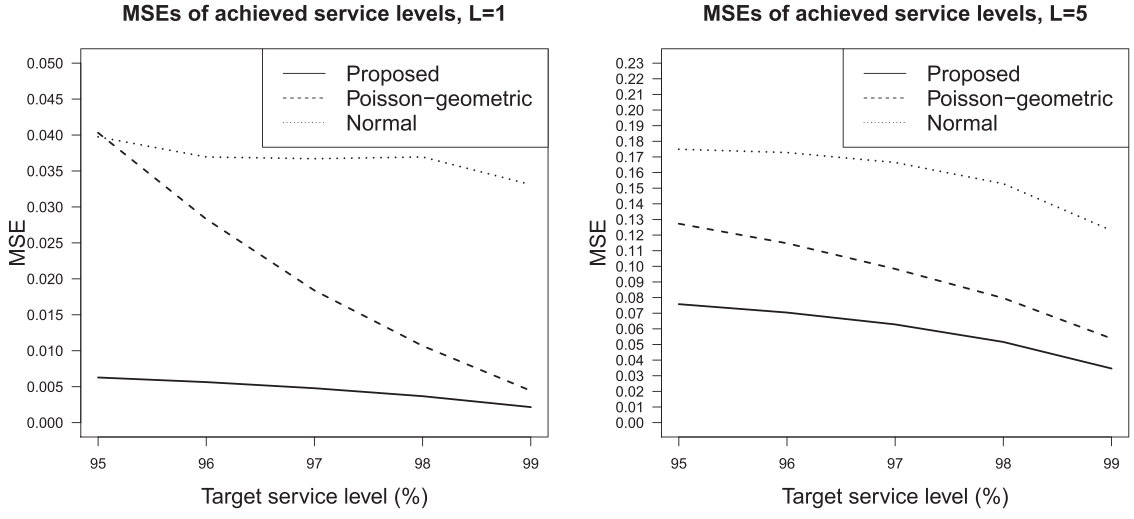
7.2. The added value of time-varying order-up-to levels

Having compared the overall performance of the proposed model and solution procedure with that of traditional demand



(a) Target service levels 50%-99%, $L = 1$

(b) Target service levels 50%-99%, $L = 5$



(c) Target service levels 95%-99%, $L = 1$

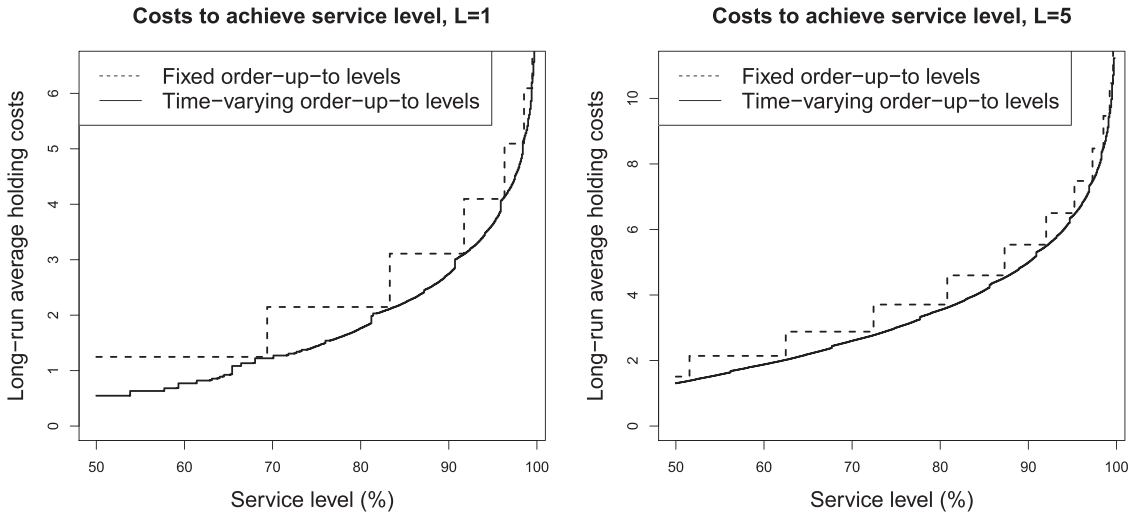
(d) Target service levels 95%-99%, $L = 5$

Fig. 2. Mean Squared Errors of achieved service levels over all items.

models, we now zoom in on the added value of time-varying order-up-to levels in particular. Given our demand model, one can still restrict the set of allowable inventory policies to only those with fixed order-up-to levels. We compare the holding costs required to achieve a given service level when using a fixed order-up-to level with the holding costs required when using time-varying order-up-to levels with the heuristic solution procedure presented in this paper. To do so, we use the same exemplary item that was also discussed in Figure 1 to estimate the demand size and interval distribution, and plot the long-run average holding costs (or equivalently, the average inventory level, as we set $h = 1$ throughout the paper) that are necessary to achieve non-stockout probabilities between 50% and 99%.

Figure 3a shows the results for $L = 1$ and displays an important benefit of using time-varying order-up-to levels, namely the larger number of variables available to "fine-tune" the inventory policy. Increasing a fixed order-up-to level by one unit leads to a large jump in the service level and holding costs. Service levels between these jump points cannot be achieved exactly, and therefore both the service level and the holding costs are overshot. This overshoot – which is omnipresent for any discrete inventory policy – is much smaller with time-varying order-up-to levels, leading to a much smoother curve. In line with the typical pattern of decreasing service level returns to cost investment, the jumps are largest for lower service levels. For example, increasing a fixed order-up-to level from 2 to 3 increases the service level from 69% to 83%, and the holding costs from 1.25 to 2.15 (a 72% cost increase), whereas increasing a fixed order-up-to level from 5 to 6 increases the service level from 96.3% to 98.5%, and the holding costs from 4.10 to 5.09 (a 24% cost increase).

The cost advantage of using time-varying order-up-to levels is largest for service levels immediately after jumps in the fixed order-up-to policy. For example, a service level of 69.4% can be achieved at 43% lower costs with time-varying order-up-to levels than with fixed order-up-to levels. In coherence with the decreasing jump sizes, the magnitude of the difference between both methods also decreases for larger service levels, although it remains substantial. A service level of 98.5% can be achieved at 15% lower costs.



(a) $L = 1$ (b) $L = 5$

Fig. 3. Long-run average holding costs needed to achieve a given non-stockout probability, comparison between a fixed order-up-to level and the heuristic solution procedure with time-varying order-up-to levels, for an exemplary item.

The cost advantage gradually decreases between the jumps. Immediately before a jump in the fixed order-up-to policy, both methods perform very similarly. In most such points the time-varying policy still yields lower or the same costs. Immediately before the jump points at service levels of 96.3% and 98.5%, though, the fixed order-up-to level performs slightly better. This demonstrates that the heuristic solution procedure – although it overall clearly dominates the benchmark policy – is not guaranteed to find the most cost-efficient solution for each service level. To avoid these rare cases, the heuristic could easily be extended with a check whether a fixed order-up-to policy outperforms the best found solution. It should be noted that the points and magnitudes of the jumps are item-specific.

Figure 3b shows the results for $L = 5$. In line with our findings of Section 7.1, we find that with a larger lead time the jumps of both methods become smaller. This can be explained by the fact that for a longer lead time, lead-time demand can take on more values and thus becomes smoother. Also the performance difference between both methods decreases, as the effect of the current position in the demand interval on the lead-time demand distribution becomes smaller. Nevertheless, using time-varying order-up-to levels still leads to significant advantages. In this case, a service level of 51.5% can be achieved at 35% lower costs with time-varying order-up-to levels, whereas a service level of 98.5% can be achieved at 10% lower costs. Also for $L = 5$ the gains decrease gradually between the jumps. We conclude that even though the largest relative gains can be achieved for short lead times and low service levels, significant advantages can be observed for this item over the entire range of service levels and for both lead times.

8. Conclusion

We presented a generalized intermittent demand inventory control model which allows for any discrete distribution of the demand interval and demand size. Our model allows for time-varying order-up-to levels that follow the distributional shape of both the demand size and interval, so that inventories can be adjusted throughout the demand interval, in anticipation of varying future requirements. We showed how to calculate the long-run average holding costs, achieved non-stockout probability, order fill rate, and volume fill rate. We furthermore suggested a greedy marginal-analysis heuristic solution procedure to optimize order-up-to levels under any of these service constraints.

Several authors found that the assumptions underlying standard demand models – such as the compound Poisson and compound binomial – are debatable on real-life data sets. In the M5 competition data set we also found that there are more consecutive periods with positive demands than would be expected based on the classical assumption of time-independent demand occurrence probabilities. An imbalance exists between forecasting developments on the one hand and advances in inventory control on the other hand. Whereas alternative forecasting methods (such as bootstrapping of the lead-time demand distribution) have been presented, current inventory control models with time-independent control parameters cannot fully exploit their predictions. Our model can, and is therefore a tool to manage inventories in an agile way.

Time-varying order-up-to levels provide a two-fold benefit. First, they allow to anticipate with greater accuracy on upcoming changes in the demand for an item. Second, they provide significantly increased flexibility over a single, fixed order-up-to level to minimize the service level overshoot that exists for discrete demand inventory models. Whereas this overshoot is largest for low service levels, we found that significant savings are achieved also for high service levels. Our model's on-target inventory performance dominates that of the Poisson-geometric and normal demand model. The largest gains can be achieved for short lead times, as these imply a more responsive inventory system.

Three main limitations of our study can be identified. First, although the M5 forecasting competition provides an established benchmark data set for a range of forecasting and inventory applications, it should be noted that our empirical results are limited to this data set. Second, applying the proposed model in practice entails estimating (next to the demand size distribution) a number of parameters equal to the maximum observed demand interval length, which may be prohibitive if only a short demand history is available. Finally, although the heuristic solution procedure is motivated by existing literature, yields close-to-optimal results in the small benchmark instances, and substantially outperforms the benchmark methods in the empirical study, it is – like all heuristics – not guaranteed to give the optimal solution.

Future research should proceed on the interface of intermittent demand forecasting and inventory control, especially on the transfer of distributional (lead-time) demand forecasts to inventory decision models. Furthermore, time-varying inventory control parameters are also useful for non-intermittent demand patterns, for example when seasonality or a trend is involved. The current model can in principle handle any (also non-intermittent) demand pattern, but an inventory model with time-varying control parameters can be tailored to any forecasting model that predicts varying demand (levels or distributions) for different periods ahead. A natural extension can be made to a multi-location inventory system, where an item can be relocated between locations in anticipation of diverging future requirements. A final research avenue is the development of interfaces between non-parametric (e.g. machine learning) demand forecasts and the optimization of inventory control parameters.

Acknowledgements

Funding: This research was supported by the Dutch Research Council (NWO) [grant nr. 019.191SG.003].

References

Axsäter, S. (2015). *Inventory control* (3rd ed.). Basel: Springer International Publishing.
Babai, M. Z., Dallery, Y., Boubaker, S., & Kalai, R. (2019). A new method to forecast intermittent demand in the presence of inventory obsolescence. *International Journal of Production Economics, 209*, 30–41.
Babai, M. Z., Tsadiras, A., & Papadopoulos, C. (2020). On the empirical performance of some new neural network methods for forecasting intermittent demand. *IMA Journal of Management Mathematics, 31*(3), 281–305.
Bijvank, M., Koole, G., & Vis, I. F. A. (2010). Optimising a general repair kit problem with a service constraint. *European Journal of Operational Research, 204*(1), 76–85.
Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability, 3*(3), 534–545.
Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly, 23*(3), 289–303.
Doszyń, M. (2019). Intermittent demand forecasting in the enterprise: Empirical verification. *Journal of Forecasting, 38*(5), 459–469.
Dunsmuir, W. T. M., & Snyder, R. N. (1989). Control of inventories with intermittent demand. *European Journal of Operational Research, 40*(1), 16–21.
Emrich, L. J., & Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician, 45*(4), 302–304.
Engels, W. R. (2009). Exact tests for Hardy–Weinberg proportions. *Genetics, 183*(4), 1431–1441.
Engels W.R., XNomial: Exact goodness-of-fit test for multinomial data with fixed probabilities, 2015. R package version 1.0.4. Last accessed 15 January 2022, https://CRAN.R-project.org/package=XNomial.
Graves, S. C. (1982). Note – a multiple-item inventory model with a job completion criterion. *Management Science, 28*(11), 1334–1337.
Gutierrez, R. S., Solis, A. O., & Mukhopadhyay, S. (2008). Lumpy demand forecasting using neural networks. *International Journal of Production Economics, 111*(2), 409–420.
Hasni, M., Aguir, M. S., Babai, M. Z., & Jemai, Z. (2019). On the performance of adjusted bootstrapping methods for intermittent demand forecasting. *International Journal of Production Economics, 216*, 145–153.
International Institute of Forecasters, Time series data – M competition, 2022. Last accessed 15 January 2022, https://forecasters.org/resources/time-series-data.
Janssen, F., Heuts, R., & de Kok, T. (1998). On the (R, s, Q) inventory model when demand is modelled as a compound Bernoulli process. *European Journal of Operational Research, 104*(3), 423–436.
Jiang, P., Huang, Y., & Liu, X. (2021). Intermittent demand forecasting for spare parts in the heavy-duty vehicle industry: A support vector machine model. *International Journal of Production Research, 59*(24), 7423–7440.
Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. *International Journal of Production Economics, 143*(1), 198–206.
Larsen, C., & Thorstenson, A. (2008). A comparison between the order and the volume fill rate for a base-stock inventory control system under a compound renewal demand process. *Journal of the Operational Research Society, 59*(6), 798–804.
Larsen, C., & Thorstenson, A. (2014). The order and volume fill rates in inventory control systems. *International Journal of Production Economics, 147*, 13–19.
Lengu, D., Syntetos, A. A., & Babai, M. Z. (2014). Spare parts management: Linking distributional assumptions to demand classification. *European Journal of Operational Research, 235*(3), 624–635.
Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., & Gucci, S. (2017). Single-hidden layer neural networks for forecasting intermittent demand. *International Journal of Production Economics, 183*, 116–128.
Makridakis S., E. Spiliotis, V. Assimakopoulos, The M5 competition: Background, organization, and implementation, 2021. In press. Last accessed 15 January 2022, doi:10.1016/j.ijforecast.2021.07.007.
Makridakis S., E. Spiliotis, V. Assimakopoulos, The M5 accuracy competition: Results, findings and conclusions, 2022. In press. Last accessed 15 January 2022, doi:10.1016/j.ijforecast.2021.11.013.
Mukhopadhyay, S., Solis, A. O., & Gutierrez, R. S. (2012). The accuracy of non-traditional versus traditional methods of forecasting lumpy demand. *Journal of Forecasting, 31*(8), 721–735.
Nikolopoulos, K. (2021). We need to talk about intermittent demand forecasting. *European Journal of Operational Research, 291*(2), 549–559.
Pennings, C. L. P., Van Dalen, J., & van der Laan, E. A. (2017). Exploiting elapsed time for managing intermittent demand for spare parts. *European Journal of Operational Research, 258*(3), 958–969.
Petropoulos, F., & Makridakis, S. (2020). The M4 competition: Bigger. Stronger. Better. *International Journal of Forecasting, 36*(1), 3–6.
Porras, E., & Dekker, R. (2008). An inventory control system for spare parts at a refinery: An empirical comparison of different re-order point methods. *European Journal of Operational Research, 184*(1), 101–132.
Prak, D., Saccani, N., Syntetos, A., Teunter, R., & Visintin, F. (2017). The repair kit problem with positive replenishment lead times and fixed ordering costs. *European Journal of Operational Research, 261*(3), 893–902.
Prak, D., Teunter, R., Babai, M. Z., Boylan, J. E., & Syntetos, A. (2021). Robust compound Poisson parameter estimation for inventory control. *Omega, 104*, 102481.
Prestwich, S. D., Tarim, S. A., Rossi, R., & Hnich, B. (2014). Forecasting intermittent demand by hyperbolic-exponential smoothing. *International Journal of Forecasting, 30*(4), 928–933.
Snyder, R. D., Ord, J. K., & Beaumont, A. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting, 28*(2), 485–496.
Strijbosch, L. W. G., Heuts, R. M. J., & Van der Schoot, E. H. M. (2000). A combined forecast–inventory control procedure for spare parts. *Journal of the Operational Research Society, 51*(10), 1184–1192.
Syntetos, A. A., Babai, M. Z., & Altay, N. (2012). On the demand distributions of spare parts. *International Journal of Production Research, 50*(8), 2101–2117.
Syntetos, A. A., Babai, M. Z., & Gardner, E. S., Jr. (2015). Forecasting intermittent inventory demands: Simple parametric methods vs. bootstrapping. *Journal of Business Research, 68*(8), 1746–1752.
Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research, 252*(1), 1–26.
Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International Journal of Production Economics, 71*(1–3), 457–466.
Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting, 21*(2), 303–314.
Teunter, R. H. (2006). The multiple-job repair kit problem. *European Journal of Operational Research, 175*(2), 1103–1116.
Teunter, R. H., & Duncan, L. (2009). Forecasting intermittent demand: A comparative study. *Journal of the Operational Research Society, 60*(3), 321–329.
Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2010). Determining order-up-to levels under periodic review for compound binomial (intermittent) demand. *European Journal of Operational Research, 203*(3), 619–624.
Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research, 214*(3), 606–615.
Türkmen, A. C., Januschowski, T., Wang, Y., & Cemgil, A. T. (2021). Forecasting intermittent and sparse time series: A unified probabilistic framework via deep renewal processes. *PLoS ONE, 16*(11), e0259764. https://doi.org/10.1371/journal.pone.0259764.
Turrini, L., & Meissner, J. (2019). Spare parts inventory management: New evidence from distribution fitting. *European Journal of Operational Research, 273*(1), 118–130.
Van der Geest, P. A. G. (2005). The binomial distribution with dependent Bernoulli trials. *Journal of Statistical Computation and Simulation, 75*(2), 141–154.
Viswanathan, S., & Zhou, C. (2008). A new bootstrapping based method for forecasting and safety stock determination for intermittent demand items. *Working paper*. Nanyang Business School, Nanyang Technological University Singapore.
Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting, 20*(3), 375–387.