European Journal of Operational Research 278 (2019) 904–915

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

A data-driven newsvendor problem: From data to decision

Jakob Huber[a,*], Sebastian Müller[b], Moritz Fleischmann[b], Heiner Stuckenschmidt[a]

[a] Data and Web Science Group, University of Mannheim, B6 26, Mannheim 68159, Germany
[b] Business School, University of Mannheim, Schloss, Mannheim 68131, Germany

ARTICLE INFO

Article history:
Received 4 December 2017
Accepted 25 April 2019
Available online 3 May 2019

Keywords:
Inventory
Newsvendor
Retail
Machine learning
Quantile regression

ABSTRACT

Retailers that offer perishable items are required to make ordering decisions for hundreds of products on a daily basis. This task is non-trivial because the risk of ordering too much or too little is associated with overstocking costs and unsatisfied customers. The well-known newsvendor model captures the essence of this trade-off. Traditionally, this newsvendor problem is solved based on a demand distribution assumption. However, in reality, the true demand distribution is hardly ever known to the decision maker. Instead, large datasets are available that enable the use of empirical distributions. In this paper, we investigate how to exploit this data for making better decisions. We identify three levels on which data can generate value, and we assess their potential. To this end, we present data-driven solution methods based on Machine Learning and Quantile Regression that do not require the assumption of a specific demand distribution. We provide an empirical evaluation of these methods with point-of-sales data for a large German bakery chain. We find that Machine Learning approaches substantially outperform traditional methods if the dataset is large enough. We also find that the benefit of improved forecasting dominates other potential benefits of data-driven solution methods.

1. Introduction

Demand uncertainty is a major challenge in supply chain management practice and research. An important remedy for demand risk is the deployment of safety stock. In order to set appropriate stock levels, many inventory models assume a specific demand distribution (Silver, Pyke, & Thomas, 2017). These problems are then solved in a two-step procedure. First, the parameters of a given demand distribution are estimated, and second, an optimization problem based on this distribution is solved. Despite the theoretical insights generated, the distribution assumption is problematic in real-world applications, as the actual demand distribution and its parameters are not known to the decision maker in reality and may even change over time (Scarf, 1958).

The growing availability of large datasets ("Big Data") may help overcome this issue and improve the performance of inventory models in real-world situations. Data that are indicative of future demand provide an opportunity to make better-informed decisions. These data include external information that is available through the Internet and data from internal IT systems. While this potential is widely recognized (see e.g. Bertsimas & Kallus, 2018), it is unclear how to best exploit it. Extant literature is rather fragmented in that regard and proposes multiple alternative directions. Our paper intends to contribute to a more wholistic understanding of the potential of data-driven inventory management. To this end, we distinguish three levels on which data can be used to revise the traditional decision process (see Fig. 1). We discuss how these levels are interrelated, and we quantify their respective impact in a real-life application.

The first level on which data can be exploited is demand estimation. The available data may contain information about future demand that can be extracted by suitable forecasting methods. These methods use historical demand data and other feature data (e.g. weekdays, prices, weather, and product ratings) to estimate future demand. The output of these models is a demand estimate together with historical forecast errors. If additional information can be extracted, the reduced demand risk results in more accurate decisions. Machine Learning (ML) has attracted a great deal of attention in the past decade. ML methods are able to process large datasets and have been successfully applied to numerous forecasting problems (Barrow & Kourentzes, 2018; Carbonneau, Laframboise, & Vahidov, 2008; Crone, Hibon, & Nikolopoulos, 2011; Thomassey & Fiordaliso, 2006).

On the second level, the inventory decision is optimized based on the demand forecast and the historical forecast errors. To this end, it is necessary to incorporate the remaining uncertainty associated with the forecast. Traditionally, uncertainty is modeled

* Corresponding author.
E-mail addresses: jakob@informatik.uni-mannheim.de (J. Huber), s.mueller@bwl.uni-mannheim.de (S. Müller), mfleischmann@bwl.uni-mannheim.de (M. Fleischmann), heiner@informatik.uni-mannheim.de (H. Stuckenschmidt).

https://doi.org/10.1016/j.ejor.2019.04.043
0377-2217/

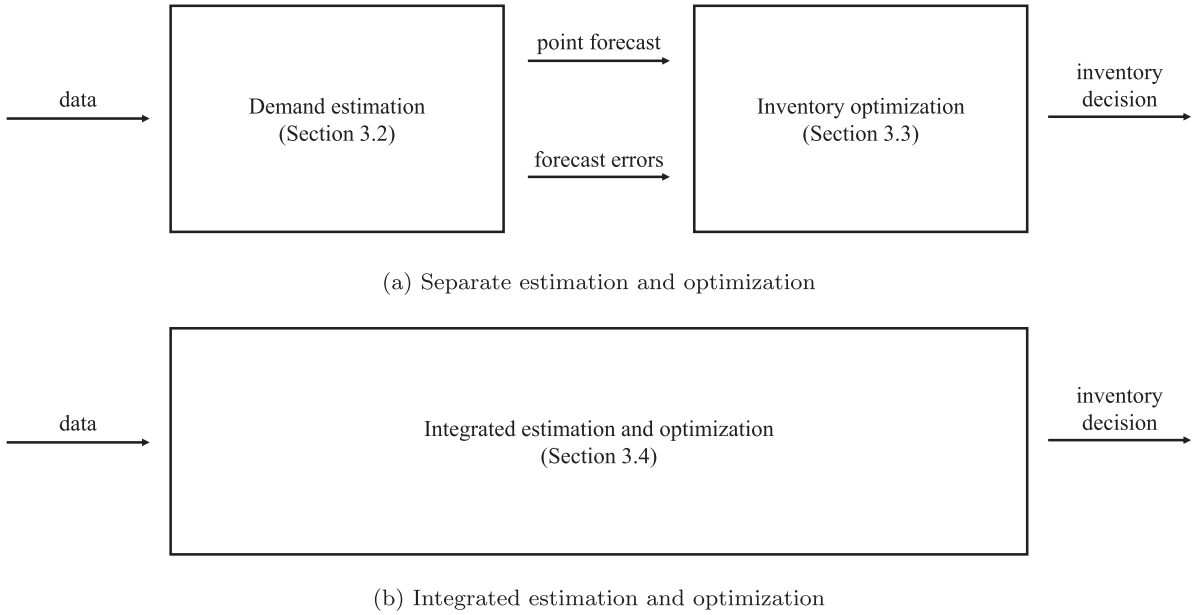


Fig. 1. The three levels of data-driven inventory management.

through a demand distribution assumption (Silver et al., 2017). We call this approach *model-based* since it explicitly models a demand distribution. However, this assumption might be misspecified and leads to suboptimal inventory policies (Ban & Rudin, 2018). Instead of speculating about a parametric demand distribution, the assumption can be replaced by empirical data that are now available on large scale. This approach is called Sample Average Approximation (SAA) (Kleywegt, Shapiro, & Homem-de Mello, 2002; Shapiro, 2003) and we call it *data-driven* as it does not rely on a distribution assumption.

On the third level, demand estimation and optimization are integrated into a single model that directly predicts the optimal decision from historical demand data and feature data, as depicted in Fig. 1(b) (Ban & Rudin, 2018; Bertsimas & Kallus, 2018; Beutel & Minner, 2012; Sachs & Minner, 2014). This approach is also *data-driven*, as it does not require the assumption of a demand distribution and works directly with data.

From the existing literature, it is not yet clear whether and under which circumstances data-driven approaches are preferred to model-based approaches. Furthermore, the question of the conditions under which separate or integrated estimation and optimization is superior remains open.

To shed light on these questions, we focus on the newsvendor problem as the basic inventory problem with stochastic demand. We empirically analyze the effects of data-driven approaches on overall costs on the three levels. Moreover, we develop novel data-driven solution methods that combine modern ML approaches with optimization and empirically compare them to well-established methods.

In our approaches, we integrate Artifical Neural Networks (ANNs) and Decision Trees (DTs) into an optimization model. Most previous work on integrated estimation and optimization assumed the inventory decision to be linear in the explanatory features (Ban & Rudin, 2018; Beutel & Minner, 2012; Sachs & Minner, 2014). This assumption poses many restrictions on the underlying functional relationships. We extend this literature by integrating multiple alternative ML methods and optimization in order to avoid these strong assumptions and incorporate unknown seasonality, breaks, thresholds, and other non-linear relationships. Recently, Oroojlooyjadid, Snyder, and Takác (2016) and Zhang and Gao (2017) also used ANNs in this context.

We evaluate our solution approaches with real-world data from a large bakery chain in Germany. The company produces and sells a variety of baked goods. It operates a central production facility and over 150 retail stores. Every evening, each store must order products that are delivered the next morning. Reordering during the day is not possible. Most of the goods have a shelf life of only one day. Thus, leftover product at the end of the day is wasted, while stock-outs lead to lost sales and unsatisfied customers.

From an optimization perspective, the problem can be represented by a newsvendor model, and the available point-of-sales data can be used to calculate forecasts. We apply our data-driven methods to the problem and compare their performance to the performance of well-established approaches. To summarize, our key contributions include the following:

- We identify and conceptualize three levels of data-driven approaches in inventory management.
- We investigate the impact of the three levels on overall performance in a newsvendor problem.
- We present novel data-driven solution approaches to the newsvendor problem based on Machine Learning.
- We compare our method to well-established approaches on the three levels and show that data-driven methods outperform their model-based counterparts on our real-world dataset in most cases.

The remainder of this paper is organized as follows. In the next section, we provide an overview of related literature. In Section 3, we describe the problem and introduce the methodology, including the data-driven ML approaches. Section 4 contains an introduction to the reference models, an empirical evaluation, and a discussion of the results. In Section 5, we summarize our findings and outline opportunities for further research.

2. Related literature

Most inventory management textbooks assume that the relevant demand distribution and its parameters are exogenously given and known (Silver et al., 2017). For a review of newsvendor-type problems, see Qin, Wang, Vakharia, Chen, and Seref (2011). In this section, we review the literature on inventory problems in which the demand distribution is unknown. More specifically, we focus

on Robust Optimization, Sample Average Approximation (SAA), and Quantile Regression (QR).

One approach that needs only partial information on demand distributions is robust optimization (Ben-Tal, Ghaoui, & Nemirovski, 2009). Scarf (1958) studies a single period problem in which only the mean and the standard deviation of the demand distribution are known. He then optimized for the maximum minimum (max–min) profit for all distributions with this property. Gallego and Moon (1993) further analyzed and extended it to a setting where reordering is possible. Bertsimas and Thiele (2006) and Perakis and Roels (2008) provide more insights into the structure of robust inventory problems. The main drawback of robust optimization is its limitation to settings with very risk-averse decision makers. For most real-world applications, robust optimization is overly conservative. For our analysis, we focus on methods that minimize expected costs instead of the max–min objective.

A data-driven method with a wider range of applications is Sample Average Approximation (SAA) (Kleywegt et al., 2002; Shapiro, 2003). Here, the demand distribution assumptions are replaced by empirical data. Levi, Roundy, and Shmoys (2007) analyze the SAA solution of a newsvendor model and its multi-period extensions. The authors calculate bounds on the number of observations that are needed to achieve similar results compared to the case with full knowledge of the true demand distribution. These bounds are independent of the actual demand distribution. More recently, Levi, Perakis, and Uichanco (2015) showed that the established bound is overly conservative and does not match the accuracy of SAA obtained in simulation studies. Therefore, they develop a tighter bound that is distribution specific. In this paper, we provide empirical support for the good performance of SAA and compare the results of diverse methods.

Instead of using sequential estimation and optimization, integrating both steps into a single optimization model has been suggested (Bertsimas & Kallus, 2018). Beutel and Minner (2012) incorporate a linear regression function for demand into their newsvendor model. The authors test their approach on simulated data and actual retail data. The model was later extended to situations with censored demand observations (Sachs & Minner, 2014). Ban and Rudin (2018) propose an algorithm that is equivalent to the one in Beutel and Minner (2012), in addition to a kernel optimization method. Furthermore, the authors show several properties of the algorithm and test it with empirical data in a newsvendor-type nurse staffing problem. Oroojlooyjadid et al. (2016) and Zhang and Gao (2017) integrate a neural network into a newsvendor model and compare it to several other approaches from the literature. However, they do not distinguish the effects of estimation, optimization, and integrated estimation and optimization. A drawback of extant research on integrated estimation and optimization is that non-linear relationships between inventory decision and feature data remain understudied. By using ML instead of a linear decision rule, our approaches can detect a priori unknown non-linear relationships between the optimal decision and the input features. Furthermore, we disentangle the effects of the three different levels of data usage highlighted in Fig. 1.

It is well known that the optimal solution to the standard newsvendor model corresponds with a certain quantile of the demand distribution (Silver et al., 2017). Estimating a certain quantile of a distribution is known as Quantile Regression (QR) in the statistics and ML literature (Koenker, 2005). A very general approach to QR is presented by Takeuchi, Le, Sears, and Smola (2006). The authors derive a quadratic programming problem and provide bounds and convergence statements of the estimator. Taylor (2000) use an ANN for QR in order to estimate conditional densities of financial returns. Similarly, Cannon (2011) describes an implementation of ANNs for QR and gives recommendations on solution approaches with gradient algorithms. More related to our application, Taylor (2007) applies QR to forecast daily supermarket sales. The proposed method can be interpreted as an adaption of exponential smoothing to QR. In the empirical evaluation, the author tests three implementations of the method: one with no regressors, one with a linear trend term, and one with sinusoidal terms to account for seasonality. None of the papers on QR we found uses QR to evaluate the costs of an inventory decision. For our solution approach, we build on the existing literature on QR by integrating ML methods into the optimization model and and evaluate the resulting costs of the newsvendor decision.

The challenge of incorporating demand uncertainty in inventory models without demand distribution assumptions is most recently also discussed by Trapero, Cardós, and Kourentzes (2018). They argue that the typical assumption of normal i.i.d. forecast errors should be questioned and suggest using a non-parametric kernel density approach for short lead times. Prak and Teunter (2018) propose a framework for incorporating demand uncertainty in inventory models that mitigates the parameter estimation uncertainty.

To summarize, we empirically evaluate the impact of data-driven approaches on the three levels (1) estimation, (2) optimization, and (3) integrated estimation and optimization. To this end, we extend the literature by proposing novel data-driven approaches to the newsvendor problem that are based on ML and build on the existing knowledge on QR in order to leverage existing big data and computation power for inventory optimization. We also illustrate the connection between QR and integrated estimation and optimization in the newsvendor context. Finally, we empirically compare the data-driven methods to their model-based counterparts and other well-established approaches.

3. Methodology

3.1. Problem description

We consider a classical newsvendor problem with an unknown demand distribution: a company sells perishable products over a finite selling season with uncertain demand. The company must choose the number of products to order prior to the selling season. If the order is too high and not all products can be sold, the company bears a cost of c_o for each unit of overage. If the order is too low and more units could have been sold, the company bears costs of c_u for each unit of underage. Thus, the objective is to minimize the total expected costs according to

$$\min_{q\geq 0} \mathbb{E}\left[c_u(D-q)^+ + c_o(q-D)^+\right], \tag{1}$$

where q is the order quantity and D is the random demand. The well-known optimal solution to this problem is to choose as the order quantity the quantile of the cumulative demand distribution function F that satisfies

$$q^* = \inf\left\{p : F(p) \geq \frac{c_u}{c_u + c_o}\right\}, \tag{2}$$

where $\frac{c_u}{c_u+c_o}$ is the optimal service level. The service level represents the probability of satisfying demand in a given period.

The problem that we address is that in most real-world cases, the actual demand distribution F is unknown. However, historical data $S_n = \{(d_1, \mathbf{x_1}), \ldots, (d_n, \mathbf{x_n})\}$ are available, where d_i is the demand and $\mathbf{x_i}$ is a vector of covariates or features (e.g. weekday, historical demand, and price) in period i. These data can be leveraged in different ways to reduce demand risk.

In the following sections, we present approaches that use the data on the three levels introduced in Section 1. First, we introduce forecasting models based on ML that we use throughout our analysis. Next, we describe a data-driven optimization approach that leverages the empirical distribution of forecast errors. Finally, we present novel data-driven models that integrate ML and the optimization model.

3.2. Demand estimation

If the underlying structure of the demand data is unknown, it is reasonable to consider very general forecasting models. ML methods have been applied to numerous forecasting tasks. Compared to traditional forecasting methods, ML is able to "learn" non-linear relationships between inputs and outputs. The most widely and successfully used methods are Artificial Neural Networks (ANNs) and Gradient Boosted Decision Trees (DTs).

ANNs are data-driven models that can approximate any continuous function (Hornik, 1991), making them suitable for forecasting if enough data are available and it is difficult to specify the underlying data generation process. An overview of time series forecasting with ANNs is provided by Zhang, Patuwo, and Hu (1998). The multilayer perceptron with a single hidden layer is commonly used for time series forecasting (Zhang et al., 1998):

$$\hat{y}(\mathbf{x}) = o\left(\mathbf{W^{(2)}} a\left(\mathbf{W^{(1)}}\mathbf{x} + \mathbf{b^{(1)}}\right) + \mathbf{b^{(2)}}\right) \tag{3}$$

Eq. (3) specifies a fully-connected feed-forward ANN. All input nodes $\mathbf{x}$ are connected to the nodes in the hidden layer, which is represented by the weight matrix $\mathbf{W^{(1)}}$. The activated output of the hidden layer is connected to the output layer by $\mathbf{W^{(2)}}$. The vectors $\mathbf{b^{(1)}}$, $\mathbf{b^{(2)}}$ describe the bias for each node. The functions $a(\cdot)$ and $o(\cdot)$ are the activation functions of the hidden layer and output layer, respectively.

Decision trees (DTs) are simple binary trees that map an input to the corresponding leaf node. Since the introduction of Classification and Regression Trees (CART) several approaches have been developed that combine multiple DTs for one prediction (e.g. Random Forrest Breiman, 2001). Gradient boosted DTs are tree ensemble models, that use K additive functions to predict the output $\hat{y}$ (Friedman, 2001):

$$\hat{y}(\mathbf{x}) = \sum_{k=1}^{K} f_k(\mathbf{x}), \tag{4}$$

where each function f_k represents a decision tree that maps the input $\mathbf{x}$ to the corresponding leaf in the tree.

3.3. Optimization

Recall that the true demand distribution F is unknown to the decision maker. In the following sections, we present two different ways to deal with this problem: traditional model-based optimization and data-driven optimization based on SAA. Both approaches use the point forecast and the historical estimation errors as inputs to determine an inventory decision.

3.3.1. Model-based optimization

The model-based approach assumes a certain forecast error distribution $\hat{F}$ (e.g. normal distribution) whose parameters $\boldsymbol{\theta}$ (e.g. mean and standard deviation) are estimated based on historical forecast errors. The order quantity is then optimized by evaluating the function at the service level quantile and adding it to the forecast:

$$q(\mathbf{x}) = \hat{y}(\mathbf{x}) + \inf\left\{p : \hat{F}(p, \hat{\theta}) \geq \frac{c_u}{c_u + c_o}\right\}, \tag{5}$$

where $\hat{y}(\mathbf{x})$ is the mean forecast, given that the features $\mathbf{x}$, and $\hat{\theta}$ are the parameters of the error distribution estimated from the resulting forecast errors. In our evaluation, we adopt normally distributed errors for the model-based approaches.

Of course, this approach yields the optimal decision if the distribution assumption is true. However, in reality, the distribution is unknown and may even change over time. The observed forecast errors depend on the model chosen to produce the forecast. A misspecified model leads to errors that are not distributed as assumed. If the demand distribution is misspecified, highly distorted decisions may result. Ban and Rudin (2018) show this for the example of a normal distribution assumption where the actual demand is exponentially distributed.

3.3.2. Data-driven optimization with Sample Average Approximation

A data-driven method to optimize the inventory decision is SAA. Here, the error distribution $\hat{F}$ is determined by the empirical forecast errors $\epsilon_1, \ldots, \epsilon_n$. A distribution assumption is not needed. Thus,

$$\hat{F}(p) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\epsilon_i \leq p). \tag{6}$$

To optimize the order quantity, the service level quantile of the empirical distribution is selected and added to the point forecast. Thus, the resulting order quantity given the features $\mathbf{x}$ is

$$q(\mathbf{x}) = \hat{y}(\mathbf{x}) + \inf\left\{p : \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\epsilon_i \leq p) \geq \frac{c_u}{c_u + c_o}\right\}. \tag{7}$$

The performance of the optimization highly depends on the quality of the forecast, the number of available data points, and the target service level. Levi et al. (2015, 2007) provide worst-case bounds for a given number of observations. An important and intuitive result is that if the optimal service level is close to 0 or 1, i.e., extreme quantiles need to be estimated, the required sample size is much higher than for service levels close to 0.5, as extreme observations are rare.

3.4. Integrated estimation and optimization with quantile regression

Instead of sequentially forecasting demand and optimizing inventory levels, one can also directly optimize the order quantity by integrating the forecasting model into the optimization problem. The optimal order quantity q of the standard newsvendor model (1) is then a function of the feature data $\mathbf{x}$. Instead of first estimating the mean demand and the error distribution and then solving the newsvendor problem, we can now directly estimate the optimal order quantity from the feature data. Beutel and Minner (2012) and Ban and Rudin (2018) formulate this problem as a linear program. This implies that the optimal order quantity is a linear function of the features. We extend these approaches by incorporating ML and thus also allowing for non-linear relationships:

$$\min_{\boldsymbol{\Phi}} \frac{1}{n}\sum_{i=1}^{n}\left[c_u(d_i - q_i(\boldsymbol{\Phi}, \mathbf{x_i}))^+ + c_o(q_i(\boldsymbol{\Phi}, \mathbf{x_i}) - d_i)^+\right], \tag{8}$$

where $q_i(\boldsymbol{\Phi}, \mathbf{x_i})$ is the output of the ML method in period i with parameters $\boldsymbol{\Phi}$ (e.g. weight matrix of an ANN) and input variables $\mathbf{x_i}$.

By introducing dummy variables u_i and o_i for the underage and overage in period i, the problem can be reformulated as a non-linear program:

$$\min_{\boldsymbol{\Phi}} \frac{1}{n}\sum_{i=1}^{n}(c_u u_i + c_o o_i) \tag{9}$$

subject to:

$$u_i \geq d_i - q_i(\boldsymbol{\Phi}, \mathbf{x_i}) \qquad \forall i = \{1, \ldots, n\}, \tag{10}$$

$$o_i \geq q_i(\boldsymbol{\Phi}, \mathbf{x_i}) - d_i \qquad \forall i = \{1, \ldots, n\}, \tag{11}$$

$$u_i, o_i \geq 0 \qquad \forall i = \{1, \ldots, n\}. \tag{12}$$

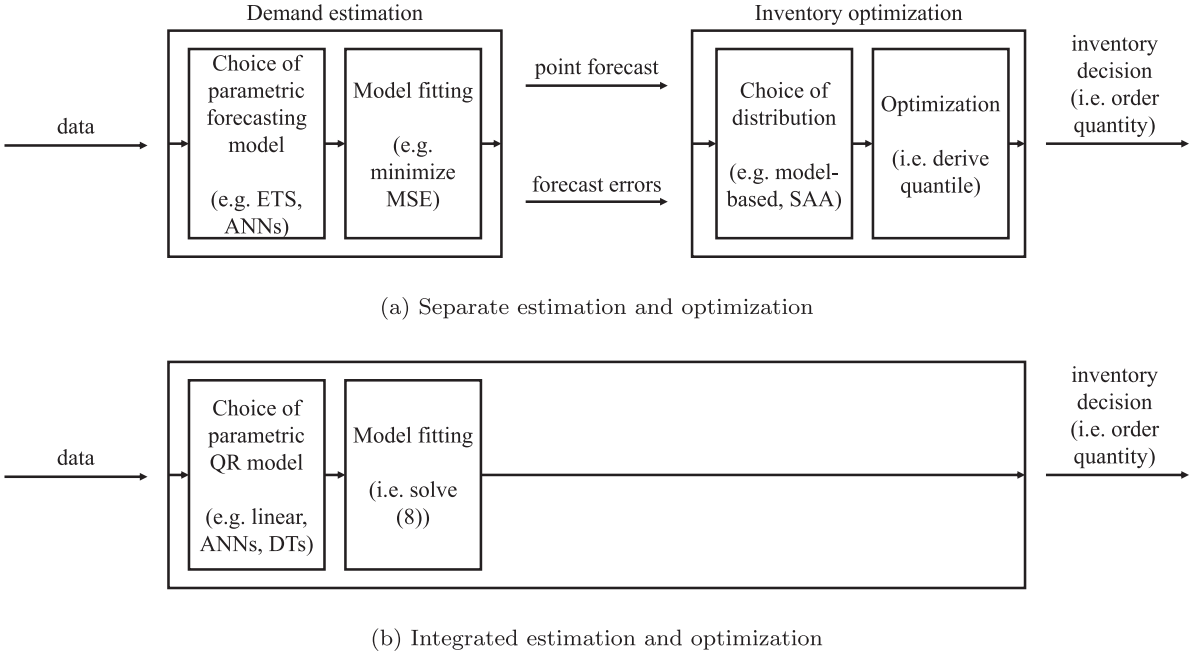


Fig. 2. Relating our methodology the three levels of data-driven inventory management.

The objective function (9) minimizes the empirical underage and overage costs, while the constraints (10)–(12) ensure that deviations of the estimate from the actual demand are correctly assigned to underages and overages. By solving the problem for the empirical data $S_n = \{(d_1, \mathbf{x_1}), \ldots, (d_n, \mathbf{x_n})\}$, we obtain parameters $\mathbf{\Phi}^*$ for the ML method that minimize the empirical costs with respect to these data. Once the model has been trained, the resulting order quantity for period p is the quantile forecast with $q_p(\mathbf{\Phi}^*, \mathbf{x}_p)$.

Bertsimas and Kallus (2018) and Ban and Rudin (2018) showed that integrating forecasting in the optimization model is equivalent to the more general QR problem in Takeuchi et al. (2006). For a better understanding, we elaborate on this relation in more detail. The basic idea of QR is to estimate the unobservable quantile by modifying the loss function of a standard regression model. Minimizing the sum of squared errors $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ yields the mean, while minimizing the sum of absolute errors $\sum_{i=1}^{n}|y_i - \hat{y}_i|$ yields the median. By weighting the underages with the quantile $\tau \in (0, 1)$ and overages with $(1-\tau)$, thus $\sum_{i=1}^{n}\tau(y_i - \hat{y}_i)^+ + (1-\tau)(\hat{y}_i - y_i)^+$, we obtain an estimate for the quantile (Koenker, 2005). The optimal solution of the newsvendor model is the quantile $\tau = \frac{c_u}{c_u+c_o}$ of the demand distribution; thus, $(1-\tau) = \frac{c_o}{c_u+c_o}$. Inserting these values of τ and $(1-\tau)$ into the objective function of the quantile regression yields the optimization problem (9).

The main advantage of QR over the model-based approach and SAA is its ability to model conditional quantiles under heteroscedasticity and for unknown error distributions. However, the performance of the approach depends crucially on the underlying model q. On the one hand, if q is too simplistic (e.g. linear), the model might not be able to capture the structure in the training data. On the other hand, if q is too complex, there is a risk of overfitting the model.

3.5. Summarizing the three levels of data-driven inventory management

We conclude this chapter by linking our methodology explained in Sections 3.2–3.4 to our framework of data-driven inventory management introduced in Fig. 1. To this end, Fig. 2 positions each piece of our methodology in the framework.

On the first level (demand estimation), we choose a parametric forecasting model (e.g. ETS or ANN). For the ML models, this includes the selection and optimization of hyper-parameters (e.g. number of layers of ANNs). We then use the data to fit the model by optimizing its parameters in order to minimize a certain objective function (i.e. MSE). The outputs of the first level of data-driven inventory management are a point demand forecast and the resulting empirical error distribution.

On the second level (inventory optimization), we operationalize a model-based approach by fitting a normal distribution and distinguish it from a data-driven (SAA) approach. We then optimize by selecting a certain quantile of the respective demand distribution. This gives us the resulting order quantity.

On the third level (integrated estimation and optimization), we choose a parametric QR model (e.g. ANNs) and fit its parameters by solving problem (8) instead of minimizing the MSE.

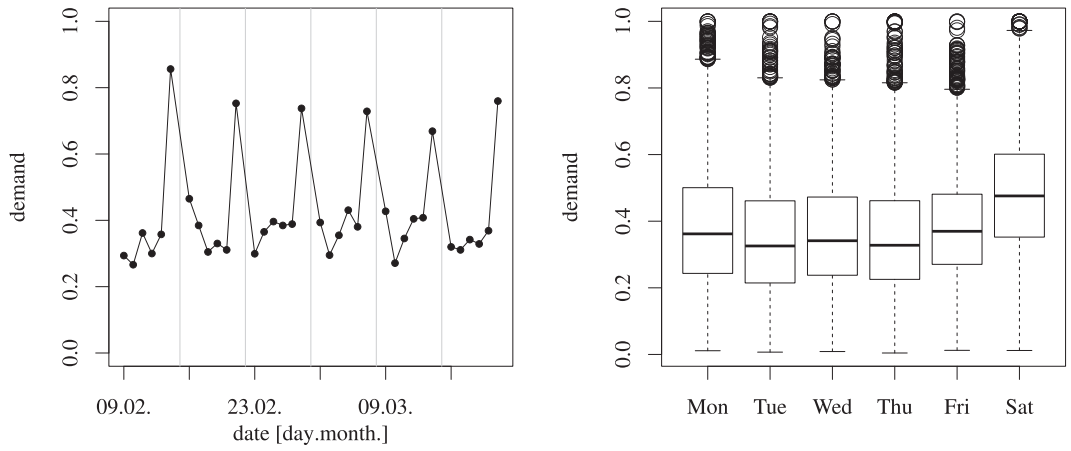
From the existing literature, it is not yet clear how the choices on each of the three levels affect performance. In the following, we investigate this question empirically.

4. Empirical evaluation

Our empirical evaluation aims to assess the impact of data-driven approaches for the three levels – (1) demand estimation, (2) optimization, and (3) integrated estimation and optimization – on average costs for the newsvendor problem. To this end, we evaluate the performance of the methods with respect to costs by using a real-world dataset to compare it to various standard approaches.

4.1. Data

We evaluate the proposed approaches using daily demand data of a German bakery chain. The observed sales are not necessarily equal to demand, as stock-outs occur and lead to censored demand information (Conrad, 1976). In order to estimate the daily demand in the case of a stock-out, we leverage intra-day sales patterns of point-of-sales data (Lau & Lau, 1996). In particular, for each product and weekday, we determine the average demand proportion of each hour in relation to the total demand on days on which the



(a) Exemplary demand time series of six weeks

(b) Boxplot of demand per weekday

Fig. 3. The demand shows a strong weekly seasonality. The demand levels for working days (Mon–Fri) are comparable, while the demand level on the weekend (Sat) is noticeably higher.

Table 1
Features used in the machine learning methods.

Data source	Features
Master Data	store class, product category, opening times (day, hours/duration)
Transactional Data	lagged sales, rolling median of sales, binary promotional information
Calendar	day of year, month, day of month, weekday, public holiday, day type, bridge day, nonworking day, indicators for each special day, school holidays
Weather	temperature (minimum, mean, maximum) and cloud cover of target day
Location	general location (city, suburb, town); in proximity to the store: shops (bakeries, butcher, grocery, kiosk, fast-food, car repair), amenities (worship, medical doctors, hospitals), leisure (playground, sport facility, park), education (kindergarden, school, university)

Table 2
Training and test periods for different sample sizes.

Sample	1.0	0.8	0.6	0.4	0.2	0.1
Train length (days)	378	300	228	150	78	36
Test length (days)	150	150	150	150	150	150

product was not sold out. This process allows us to interpolate the sales when a stock-out occurs and obtain an estimate for historical demand. The approach is feasible because we have access to point-of-sales data and information on the overage for each product per day. Fig. 3 shows the strong weekly seasonality of demand for (a) a representative product and (b) a box plot that confirms this pattern for all time series. While median demand on Tuesdays and Thursdays is the lowest, it is slightly higher on Mondays, Wednesdays and Fridays. The median demand on Saturday is higher than it is for all other days. The standard deviation of demand does not vary strongly across the weekdays.

The dataset comprises eleven stock-keeping units, namely, six breads and five buns, for five stores over a period of 88 weeks, where each store is open from Monday to Saturday. This configuration amounts to 55 ordering decisions per day. Additionally, we enrich the dataset with external explanatory features related to calendar, weather, and location of the store (see Table 1). We split the dataset into a training set containing up to 63 weeks and a test set containing the remaining 25 weeks (see Table 2). We perform a rolling 1-step-ahead prediction evaluation on the test set in order to assess the performance of the methods. We fit the models and distribution parameters every 10 days on a rolling training dataset with constant size. Due to computational constraints, we fit the parameters of the ANNs every 50 days only. To evaluate the effect of the amount of available data, we use different sample sizes for the training set. The full training set (sample size 1.0) covers 63 weeks, while the smallest training set (sample size 0.1) contains only 6 weeks (see Table 2).

While traditional time series methods such as exponential smoothing or ARIMA are able to process only a single times series at a time, a major advantage of the ML methods is their ability to deal with a large number and variety of features. In order to leverage this advantage, we do not only train them with a single time series per product but alternatively also across products and stores. In the latter case, we also include the features listed in Table 1.

4.2. Experimental design

In our experiment, we evaluate the impact of different (1) estimation, (2) optimization, and (3) integrated estimation and optimization approaches on the costs of the newsvendor model. We start by assessing the impact of forecast performance. In addition to the ANNs and DTs introduced in the previous section, we evaluate six different reference forecasting methods, which we outline in the next section. For each forecasting method, we measure the forecast accuracy (Section 4.4) and then investigate its impact on costs (Section 4.5.1). Second, we compare the model-based optimization assuming a normal distribution (*Norm*) with the data-driven optimization using *SAA*. To this end, we calculate the average costs for different target service levels (Section 4.5.2). Third, we assess the performance of the integrated estimation and optimization approach with QR and compare it to the separate approaches (Section 4.5.3). Fourth, we evaluate the sensitivity to the sample size in order to assess the value of a large training set (Section 4.5.5). Overall, the database of the evaluation results comprises more than 9.1 million entries, i.e., close to 0.6 million point forecasts and approximately 8.6 million order quantities. We employ the Wilcoxon signed-rank test to test the statistical significance of our results at the 5% significance level.

4.3. Reference methods and ML setup

In order to evaluate the ML approaches, we compare them to well-established forecasting methods. With the exception of the first approach (*Median*), we rely on methods that are explicitly able to model seasonal time series because the demand for baked goods exhibits a strong weekly seasonality (see Fig. 3).

4.3.1. Reference methods

Median and Seasonal-Median. The first benchmark forecast is the median of the entire training set (*Median*); it does not consider seasonality. Nonetheless, we include it in our comparison in order to evaluate the benefit of seasonal demand models. Its seasonal variant estimates the median by weekday (*S-Median*).

Seasonal-Naïve. A popular benchmark method for forecasting is the Naïve method and its seasonal variant (*S-Naïve*). The forecast is set to the last observed value from the same part of the season: $\hat{y}_{t+h} = y_{t+h-m}$. Hence, we need to specify only the frequency of the seasonality m, which we set to 6 for the considered time series.

Seasonal moving average. The seasonal moving average method (*S-MA*) sets the forecast to an average of the last observations from the same part of the season: $\hat{y}_{t+h} = \frac{1}{k}\sum_{i=1}^{k} y_{t+h-mk}$. Besides setting the frequency of the seasonality m, we must set k, which controls the number of considered values. We determine k in the range from 3 to 12 based on the last 20% of the training set for each time series. We choose the value of k that minimizes the sum of squared errors.

Seasonal autoregressive integrated moving average. Autoregressive integrated moving average (ARIMA) and its seasonal variant *S-ARIMA* represent a widely used forecasting method. The autoregressive part of ARIMA represents a linear combination of past values, while the moving average part is a linear combination of past forecast errors. The time series must be stationary, which can be achieved by differencing. We employ the method `auto.arima()` function from the `forecast` package (Hyndman & Khandakar, 2008) for the statistical software R (R Core Team, 2017) in order to identify the most suitable model per time series. The `auto.arima()` function selects a suitable model using a step-wise approach that traverses the space of possible models in an efficient way until the best model is found.

Exponential smoothing. Exponential smoothing methods calculate the forecast by computing a weighted average of past observations. The weights decay as the observations get older. Hyndman, Koehler, Snyder, and Grose (2002) and Hyndman, Koehler, Ord, and Snyder (2008) propose innovation space models that generalize exponential smoothing methods (*ETS*). These models include a family of 30 models that cover different types of errors, seasonal effects and trends (none, additive, multiplicative). We use the `ets()` function from the `forecast` package (Hyndman & Khandakar, 2008) for the statistical software R (R Core Team, 2017).

4.3.2. ML setup

In this subsection, we introduce methods that take multiple time series and additional features (see Table 1) into account. For these methods, we also evaluate the integrated estimation and optimization approach introduced in Section 3.4.

Linear regression. The linear regression model uses lagged demand data (lags: 1, 2, ..., 6, 12, 18) which are linearly scaled between 0 and 0.75 as input. The weekly seasonality is modeled through binary variables. When all time series across stores and products and the extended feature set are used for the prediction, further variables are introduced. In order to avoid overfitting, we include a regularization term in the objective function. The integrated linear approach is equivalent to the models in Beutel and Minner (2012) and Ban and Rudin (2018).

ANNs. We apply *ANNs* as described in Section 3.2. Several hyper-parameters (learning rate, batch size, number of hidden nodes, activation function of hidden layer) are optimized by a random search (Bergstra & Bengio, 2012) in combination with cross-validation on the training set. As activation function for the output layer we use a linear function, which is reasonable for regression with ANNs (Zhang et al., 1998).

In order to encode deterministic seasonality, we use trigonometric functions as features, as proposed by Crone and Kourentzes (2009). This is a parsimonious approach which requires only two additional input variables. Additionally, the approach is non-parametric, as no seasonal indices need to be estimated. The two variables are $x_{i,1}$ and $x_{i,2}$ in period i, with m representing the frequency of the seasonality:

$$x_{i,1} = \sin(2\pi i/m) \tag{13}$$

$$x_{i,2} = \cos(2\pi i/m) \tag{14}$$

The input consists of lagged demand information (lags: 1, 2, ..., 6, 12, 18), which is linearly scaled between 0 and 0.75, as this is similar to what other seasonal methods consider. When all time series across products and stores are considered, we enrich the dataset with further explanatory features (see Table 1).

The performance of an ANN depends on its initial weights, which are randomly set. Therefore, we employ an ensemble of ANNs with the *median* ensemble operator, as this approach is robust to the initial weights and provides reliable results (Barrow, Crone, & Kourentzes, 2010; Kourentzes, Barrow, & Crone, 2014). Another crucial aspect is the training of ANNs. We use the stochastic gradient-based algorithm ADAM proposed by Kingma and Ba (2015) to optimize the weights of the ANN. We also employ early stopping to avoid overfitting and train an ensemble of 50 *ANNs* in order to obtain more reliable and accurate results (Barrow et al., 2010; Kourentzes et al., 2014).

DTs. The DT approach is a tree-based ensemble model as described in Section 3.2. We use Microsoft's LightGBM implementation (Ke et al., 2017). Similar to the ANNs, several hyper-parameters (learning rate, number of leaves, minimum amount of data in one leaf, maximum number of bins, maximum depth of tree) are selected based on a random search within the training data (Bergstra & Bengio, 2012). The number of trees is controlled by early stopping, which also reduces the risk of overfitting. We consider the same features as in the other ML methods.

4.4. Point forecast analysis

The relevant performance measure of the newsvendor model is overall costs (overage and underage). Before evaluating the impact of the different estimation and optimization approaches on cost in Section 4.5, we separately measure the accuracy of the point forecasts in order to relate it to overall costs in the subsequent analysis.

For each forecasting method introduced in the previous section, we compute a set of common accuracy measures, including the Mean Percentage Error (MPE), Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE) (Hyndman & Koehler, 2006), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Relative Absolute Error (RAE). We provide more than one measure because each of them has its strengths and weaknesses. For instance, RMSE and MAE are scale-dependent error measures and do not allow for comparisons between time series at different scales, while percentage-based error measures (SMAPE, MAPE) are not always defined and may result in misleading outcomes if demand

Table 3
Forecast performance of the point predictions (sample size: 1.0). The best performance for each metric is underlined. Results that do not differ from the one of the best method at a significance level of 5% for each metric are printed in bold face.

Method	MPE	SMAPE	MAPE	MASE	RMSE	MAE	RAE
Median	−22.34	29.71	39.43	1.01	39.89	15.70	1.72
S-Median	−21.45	24.74	33.73	0.82	28.42	11.99	1.31
S-Naïve	**−11.84**	28.71	34.86	0.92	27.80	12.56	1.37
S-MA	−14.61	23.32	30.15	0.75	22.27	10.14	1.11
ETS	−12.47	22.19	28.47	0.71	21.83	9.66	1.06
S-ARIMA	−14.35	22.88	29.71	0.73	21.40	9.87	1.08
Linear	−18.73	23.75	32.07	0.77	23.43	10.54	1.15
DT-LGBM	−18.80	22.88	31.13	0.73	21.98	9.92	1.08
ANN-MLP	−14.73	22.63	29.59	0.72	21.28	9.75	1.07
Linear (all)	−14.33	22.14	29.18	0.71	21.23	9.63	1.05
DT-LGBM (all)	−13.44	21.51	28.34	**0.68**	**20.06**	**9.15**	**1.00**
ANN-MLP (all)	**−12.62**	**21.42**	**27.87**	**0.68**	**20.09**	**9.16**	**1.00**

is low. Table 3 shows the average forecast accuracy over all time series by method.

Not surprisingly, the worst accuracy is achieved by the *Median* forecast, which is the only method that does not incorporate the weekly seasonality pattern. The results improve noticeably (more than 5 percentage points in MAPE) when the weekly seasonality is considered (*S-Median*). *S-Median* is also more robust against sudden changes in demand and provides more reliable results than *S-Naïve*. *S-MA* outperforms all baseline methods (*Median*, *S-Median*, *S-Naïve*) and its accuracy is even competitive to more sophisticated approaches. It is not as prone to outliers but follows minor level shifts. Overall, *ETS* is the best method compared to models that are trained on a single time series as it captures the main characteristics of the time series by computing the weighted average of past observations. Even the more complex ML approaches cannot improve the forecast. However, when trained across stores and products with additional features, the ML methods further improve significantly. *ANN-MLP* and *DT-LGBM* also outperform ETS. The information contained in the features and supplementary time series has additional explanatory potential that is effectively extracted by all three ML approaches.

We note that the negative MPE throughout all methods indicates that in the test data, there are low-demand events that cannot be foreseen by the models based on historical demand. These low-demand events are more frequent, more extreme, or both during the test period than events of unexpectedly high demand. This observation might be due to the fact that situations with very low demand (e.g. supply disruption, partial shop closing, and construction) are more likely than situations with extremely high demand.

4.5. Inventory performance analysis

The purpose of the newsvendor model is to determine the cost-minimal order quantity by considering demand uncertainty and underage and overage costs. In order to perform a comprehensive analysis of the introduced methods, we calculate the order quantities and compute the resulting average costs for each approach. As underage and overage cost may vary among products and stores, we analyze multiple target service levels. The target service level $c_u/(c_u + c_o)$ is the optimal probability of having no stock-out during the day. In the repeated newsvendor model, this corresponds to the long run fraction of periods in which demand is fully satisfied. By setting the unit price and the sum of underage and overage costs (c_u + c_o) to 1.00 and varying their relative share, we obtain six different target service levels. This process allows us to interpret c_u as the profit margin and c_o as the unit costs (e.g. material and production costs) of an item. In order to compare the different methods, we measure the performance relative to the best method for each target service level. Additionally, we report the realized average service level for each approach. We calculate the realized service level as the relative share of days on which total demand was met. A large deviation of the realized service level from the target service level indicates that a method tends to overestimate or underestimate the optimal order quantity. Note that the reported service level just serves to characterize the solution by relating it to the newsvendor solution. It does not reflect a cost-service trade-off since costs include both overage and underage costs. The results are reported in Table 4.

In the following sections, we analyze the effects of (1) demand estimation, (2) optimization, and (3) integrated estimation and optimization on average costs and observed service levels. Furthermore, we evaluate the sensitivity of the results to the size of the available sample.

4.5.1. The effect of demand estimation

To evaluate the effect of demand estimation on costs, we compare the average cost of the different estimation approaches for each target service level in Table 4. The best approach for each target service level is underlined. We see that the approaches based on the ML forecasts that use data across stores and products and additional features (*all*) provide the lowest average costs for all target service levels. The performance of *ANN-MLP* and *DT-LGBM* is very similar, while methods based on the *Linear* forecast yield higher costs. An interesting result is that *ETS* performs best when training is restricted to single time series. This is particularly noteworthy when considering its computational efficiency compared to the ML methods. Overall, we observe that approaches based on accurate estimation methods achieve significantly lower costs, independent of the optimization approach. Thus, the level of demand estimation has a substantial impact on overall performance.

In order to further substantiate this statement, we conduct a correlation analysis. We compute Spearman's rank correlation coefficient ρ between costs and forecast accuracy (SMAPE and RMSE) for each store-article-service level combination. The results are depicted in Table 5.

The analysis supports the claim that the general ranking of methods with respect to costs is similar to the ranking with respect to forecast accuracy, with a median ρ of 0.8799 for the rank correlation of costs and SMAPE and 0.9406 for the median rank correlation of costs and RMSE. The reason for this observation is that more accurate point predictions lead to more precise demand distribution estimates, which make the succeeding optimization phase less crucial.

We complement the above cost analysis by looking at the realized service levels which provide further insights into the order quantities obtained from the different methods.

Table 4
Inventory performance analysis: Average cost increase relative to the best approach and average service level (SL) for various target service levels (TSLs) and a sample size of 1.0. Methods denoted with *all* are trained on data across all products and stores. The best approach for each target service level is underlined. Results that do not differ from the one of the best method at a significance level of 5% for each service level are printed in bold face.

	Method		TSL = 0.5		TSL = 0.6		TSL = 0.7		TSL = 0.8		TSL = 0.9		TSL = 0.95	
	Estimation	Optimization	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL
Benchmarks	Median	Norm	72.5	0.61	83.9	0.72	93.2	0.79	99.4	0.86	97.8	0.92	92.0	0.95
		SAA	72.5	0.61	79.4	0.70	87.9	0.79	99.6	0.87	109.5	0.94	101.9	0.98
	S-Median	Norm	31.8	0.64	33.5	0.74	34.7	0.83	34.9	0.89	32.2	0.95	29.6	0.97
		SAA	31.8	0.64	30.3	0.72	30.4	0.80	29.5	0.88	27.4	0.95	31.2	0.98
	S-Naive	Norm	38.0	0.51	37.5	0.63	37.0	0.75	37.3	0.85	37.2	0.93	37.2	0.96
		SAA	38.4	0.51	37.6	0.61	35.6	0.71	34.3	0.81	32.2	0.91	33.3	0.96
	S-MA	Norm	11.5	0.56	13.6	0.68	16.0	0.78	17.6	0.86	18.3	0.94	16.6	0.97
		SAA	10.5	0.52	11.0	0.62	11.4	0.73	11.7	0.82	12.2	0.92	13.9	0.96
	ETS	Norm	6.1	0.53	6.7	0.64	7.0	0.74	7.1	0.83	5.6	0.91	5.7	0.95
		SAA	6.2	0.50	6.5	0.61	6.7	0.71	6.7	0.80	5.6	0.90	5.9	0.95
	S-ARIMA	Norm	8.5	0.55	8.9	0.65	8.8	0.75	8.3	0.84	7.5	0.92	7.2	0.95
		SAA	8.0	0.52	8.1	0.62	8.0	0.71	7.7	0.81	6.5	0.91	7.2	0.95
ML single time series	Linear	Norm	15.8	0.58	17.7	0.69	19.6	0.78	20.8	0.85	20.2	0.93	20.9	0.95
		SAA	15.6	0.56	17.2	0.66	18.9	0.75	20.1	0.84	20.7	0.93	21.8	0.96
		QR	10.6	0.54	10.7	0.64	11.4	0.73	11.2	0.82	11.8	0.91	18.9	0.96
	DT-LGBM	Norm	9.0	0.60	8.6	0.68	8.5	0.76	8.8	0.83	10.2	0.89	**15.2**	0.93
		SAA	7.9	0.57	7.7	0.65	7.8	0.73	8.4	0.81	10.0	0.89	14.4	0.94
		QR	11.1	0.59	10.8	0.68	12.1	0.78	15.3	0.85	20.8	0.93	29.0	0.96
	ANN-MLP	Norm	7.2	0.55	8.4	0.66	9.0	0.75	9.6	0.83	9.4	0.91	10.5	0.95
		SAA	6.6	0.52	7.6	0.63	8.2	0.72	8.6	0.82	8.6	0.91	10.2	0.95
		QR	7.5	0.53	7.9	0.64	8.6	0.73	9.8	0.82	13.0	0.91	18.1	0.95
ML pooled time series + features	Linear (all)	Norm	5.9	0.53	5.5	0.64	5.6	0.75	6.1	0.84	4.9	0.91	4.0	0.95
		SAA	5.4	0.51	5.3	0.62	5.0	0.72	5.3	0.82	5.2	0.91	4.9	0.95
		QR	5.1	0.52	4.5	0.62	5.2	0.72	7.2	0.81	10.0	0.90	12.8	0.95
	DT-LGBM (all)	Norm	**0.6**	0.53	**0.4**	0.62	0.0	0.71	0.1	0.80	0.4	0.87	2.1	0.92
		SAA	0.9	0.51	**0.4**	0.61	**0.0**	0.69	**0.0**	0.79	0.2	0.88	1.7	0.92
		QR	1.6	0.52	1.7	0.61	1.6	0.71	3.1	0.80	6.4	0.90	11.4	0.94
	ANN-MLP (all)	Norm	0.7	0.52	0.7	0.63	0.7	0.73	0.7	0.82	**0.0**	0.90	**0.0**	0.95
		SAA	**0.3**	0.51	**0.2**	0.61	0.3	0.72	0.4	0.81	0.4	0.90	1.5	0.95
		QR	**0.0**	0.50	**0.0**	0.61	0.9	0.72	3.3	0.82	6.8	0.91	11.2	0.95

Table 5 also shows the Spearman Correlations between the absolute service level deviations (i.e. difference between average observed service level and the newsvendor target service level) and costs and forecast errors, respectively.

From Table 4, we can see that all methods overachieve the target service level on average. This matches our observation of Section 4.4, that all forecasting methods overestimate the demand on average, due to events with unexpectedly low demand in the test data.

We further see that the correlation between the absolute service level deviation and costs is relatively low (0.4202). This shows that the ability of a method to achieve a desired service level on average is not a very good indicator for the cost performance of that method. The service level measures only whether or not there was a stock-out and thus indicates the direction of the deviation from the optimal order quantity on average. It does not take into account the order of magnitude of overages and underages. The low correlation between the forecast accuracy measures and the service level deviation confirms this conclusion.

Table 5
Median of Spearman's correlations (± standard deviation) between absolute service level deviation (SL), costs, and forecast accuracy (SMAPE, RMSE).

	Costs	SMAPE	RMSE
Costs	–	0.8799 (±0.1211)	0.9406 (±0.0481)
SL	0.4202 (±0.2879)	0.4253 (±0.2834)	0.3034 (±0.2678)

4.5.2. The effect of optimization

To assess the impact of model-based vs. data-driven optimization on costs, we compare the average cost of *Norm* and *SAA* for each estimation method and target service level. We perform a Shapiro–Wilk test on the residuals of the forecasts of S-ARIMA and ETS and find that for approximately one quarter of the time series the residuals are normally distributed at 95% confidence level. Thus, the normal distribution assumption can be justified, although one cannot expect that all residuals follow the distribution assumption in a real-world data set. We observe that the performance differences between *SAA* and *Norm* are relatively small and the effect of accurate demand estimation clearly outweighs the effect of data-driven optimization. However, for the majority of estimation methods, *SAA* leads to lower costs than *Norm* for target service levels up to 0.9, while the normal distribution assumption can be beneficial for higher service levels.

The good performance of *SAA* and its weaknesses for higher service levels are in line with the theoretical results of Levi et al. (2015). The authors provide a bound on the accuracy of *SAA* for the newsvendor model (Theorem 2 *Improved LRS Bound*) that does not rely on assumptions on the demand distribution. The bound has an exponential rate that is proportional to the sample size and $\min(c_u, c_o)/(c_u + c_o)$. In our case, the bound implies that using *SAA*, in order to obtain the same accuracy for a service level of 0.9 (0.95) as for a service level of 0.8, we would need 1.5 (4) times more data. However, in the bakery industry, such high service levels are not common, and our dataset is sufficient to let *SAA* outperform *Norm* for service levels up to 0.9 for most approaches.

4.5.3. The effect of integrated estimation and optimization

We also employ the QR approach that integrates the demand estimation into the optimization model for the *linear* approach and the ML methods *DT-LGBM* and *ANN-MLP*. In order to focus on

Table 6
The effect of the sample size: average cost increase relative to the best approach (over all sample sizes) and average service level (SL) for the target service level 0.7 and various sample sizes (S). Methods denoted with *all* are trained on data across all products and stores. The best approach for each sample size is underlined. Results that do not differ from the one of the best method at a significance level of 5% for each sample size are printed in bold face.

	Method		$S = 0.1$		$S = 0.2$		$S = 0.4$		$S = 0.6$		$S = 0.8$		$S = 1.0$	
	Estimation	Optimization	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL	Δ Cost (%)	SL
Benchmarks	Median	Norm	79.5	0.72	82.0	0.75	87.1	0.79	92.7	0.80	92.7	0.80	93.2	0.79
		SAA	77.2	0.70	78.2	0.74	82.6	0.78	88.0	0.79	87.7	0.79	87.9	0.79
	S-Median	Norm	13.5	0.70	15.5	0.76	23.2	0.81	32.5	0.84	33.6	0.83	34.7	0.83
		SAA	13.3	0.65	12.1	0.72	18.5	0.79	26.8	0.81	29.1	0.80	30.4	0.80
	S-Naive	Norm	36.7	0.72	36.5	0.73	36.8	0.75	37.5	0.75	37.1	0.75	37.0	0.75
		SAA	40.2	0.68	37.4	0.69	36.2	0.70	35.9	0.71	35.7	0.71	35.6	0.71
	S-MA	Norm	15.8	0.73	14.6	0.75	15.1	0.77	15.9	0.78	15.6	0.78	16.0	0.78
		SAA	19.2	0.67	15.4	0.67	12.1	0.70	11.0	0.72	11.8	0.72	11.4	0.73
	ETS	Norm	**12.1**	0.70	**8.8**	0.73	7.9	0.74	7.7	0.74	6.6	0.74	7.0	0.74
		SAA	14.0	0.66	9.6	0.68	8.1	0.69	7.6	0.70	6.5	0.71	6.7	0.71
	S-ARIMA	Norm	25.8	0.69	13.8	0.72	10.7	0.74	10.3	0.75	9.2	0.75	8.8	0.75
		SAA	29.2	0.64	**14.5**	0.68	10.5	0.68	9.8	0.70	8.6	0.72	8.0	0.71
ML single time series	Linear	Norm	29.8	0.70	18.0	0.71	18.9	0.76	21.1	0.79	20.4	0.78	19.6	0.78
		SAA	30.2	0.68	18.3	0.70	18.1	0.74	20.1	0.76	19.3	0.76	18.9	0.75
		QR	32.1	0.67	17.4	0.68	14.3	0.71	13.2	0.73	12.0	0.73	11.4	0.73
	DT-LGBM	Norm	50.4	0.72	22.6	0.73	13.6	0.77	12.5	0.79	10.4	0.78	8.5	0.76
		SAA	48.7	0.68	22.6	0.68	11.1	0.73	10.2	0.75	8.5	0.75	7.8	0.73
		QR	52.4	0.73	27.3	0.75	17.7	0.78	16.8	0.80	13.6	0.79	12.1	0.78
	ANN-MLP	Norm	33.2	0.69	14.3	0.73	13.9	0.78	12.0	0.78	9.9	0.78	9.0	0.75
		SAA	33.5	0.70	14.9	0.72	12.4	0.74	10.2	0.75	8.4	0.75	8.2	0.72
		QR	33.1	0.67	17.8	0.71	12.5	0.75	10.7	0.75	9.2	0.75	8.6	0.73
ML pooled time series + features	Linear (all)	Norm	**16.0**	0.69	**14.2**	0.70	14.8	0.73	8.8	0.74	8.2	0.75	5.6	0.75
		SAA	**17.2**	0.64	14.5	0.66	13.5	0.69	7.3	0.71	7.6	0.72	5.0	0.72
		QR	**15.3**	0.65	12.7	0.67	13.3	0.69	7.7	0.71	7.1	0.72	5.2	0.72
	DT-LGBM (all)	Norm	**21.1**	0.67	10.4	0.69	**8.0**	0.69	**5.5**	0.71	2.5	0.70	0.0	0.71
		SAA	**20.3**	0.65	10.6	0.66	**7.5**	0.67	**4.7**	0.69	**2.4**	0.68	**0.0**	0.69
		QR	18.5	0.71	10.8	0.69	7.7	0.70	**4.8**	0.71	3.4	0.70	1.6	0.71
	ANN-MLP (all)	Norm	33.5	0.60	12.4	0.69	6.1	0.69	3.6	0.73	0.7	0.71	0.7	0.73
		SAA	36.3	0.58	12.6	0.66	**5.4**	0.67	**2.8**	0.71	**0.3**	0.69	0.3	0.72
		QR	**18.7**	0.63	12.4	0.65	6.2	0.68	**3.1**	0.70	0.3	0.70	0.9	0.72

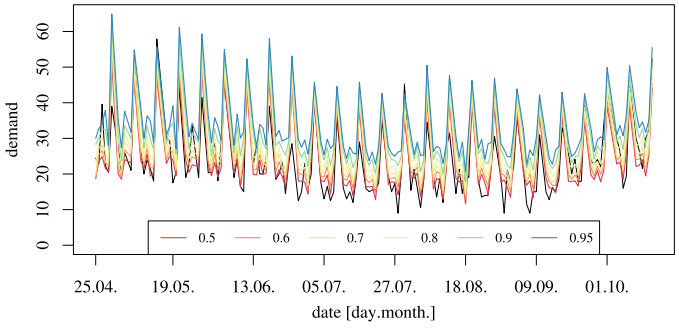


Fig. 4. Forecasts for different service levels using ANN QR.

the effect of integrated estimation and optimization, we compare *QR* to *SAA* for the respective approaches. For *DT-LGBM* and *ANN-MLP* trained on single time series, *QR* performs worse than *SAA*, while *Linear QR* outperforms *SAA*. For high service levels *QR* generally performs relatively poor for all three estimation approaches. When trained on data across stores and products and including features, integration of estimation and optimization improves the performance of *Linear (all)* and *ANN-MLP (all)* for low service levels. However, for high target service levels, *SAA* and *Norm* perform better than *QR* for all estimation approaches.

The theoretical advantage of the *QR* approach is its ability to estimate *conditional* quantiles that depend on the features (see Fig. 4). The observation that for the approaches trained only on single time series, *QR* is not beneficial, might be explained by the fact that too little features are available to leverage the feature-dependency of the quantile. The previous statement is supported by the fact that *Linear (all)* and *DT-LGBM (all)* improve through integration at low service levels as more data are available and feature-dependent variance can be estimated more accurately. However, this theoretical advantage cannot be observed for higher service levels. We suspect that more extensive hyperparameter optimization in combination with alternative scaling of the input data for each individual target service level might improve the performance.

Our results for the single time series case are in line with the outcome of the empirical analysis of Ban and Rudin (2018) who also report that separate estimation and optimization outperforms the linear integrated approach on their relatively small dataset of one year. We observe that this effect gets smaller when the models are trained with pooled time series and features.

4.5.4. The effect of learning across products and external features

Our dataset comprises sales data of several breads and buns across multiple stores. These products are relatively similar to one another and therefore one time series might contain information about the other. Univariate time series models can only consider a single product at time, while ML methods are able to process a large number of inputs. Therefore, we train *linear (all)*, *DT-LGBM (all)*, and *ANN-MLP (all)* across all products and stores. The pooling of training data also makes it possible to enhance the data set with a large number of additional features that cannot be employed if the models are trained per time series.

From Table 4 we observe that indeed all ML methods benefit from the additional data and improve significantly. *DT-LGBM (all)* and *ANN-MLP (all)* perform similarly and outperform all other

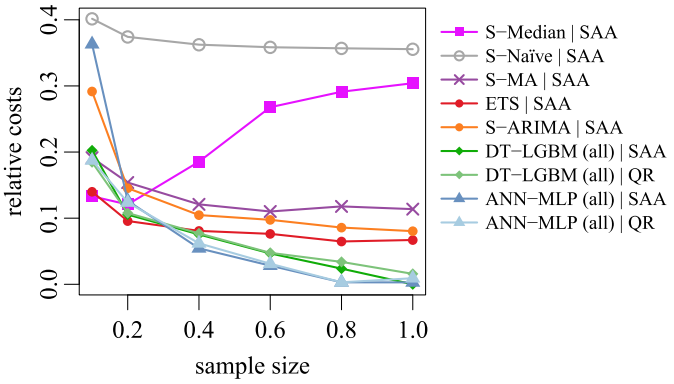


Fig. 5. Effect of the sample size (TSL = 0.7).

methods. We note that a similarity of time series is not specific to our case but can be found in many retail settings.

4.5.5. Sensitivity to sample size

The power of the data-driven approaches lies in their ability to leverage large amounts of available data, which makes them very flexible but may limit their deployability if not enough data is available. In order to determine the dependency of the different approaches on data availability, we vary the size of the training data and compare the results on a fixed test set (see Table 2). The results of this experiment are given in Table 6 and depicted in Fig. 5 for the data-driven approaches. We present only the results for target service level 0.7, noting that the qualitative results also apply to the other service levels.

Based on our results, the methods can be divided into three groups: The first group consists of methods whose performance hardly depends on the sample size. In our case this includes methods based on the *S-Naïve* forecast. The *S-Naïve* approaches simply forecast the demand of the same weekday of the week before. Thus, it does not improve as more data becomes available. The second group consists of methods whose performance diminishes as more training data become available. The approaches with a *Median* (not depicted in Fig. 5, see Table 6) and *S-Median* forecast are part of this group. The costs increase as more training data are available and as more "outdated" data are included. In our real-world case, this observation implies that, for example, demand data from Winter is used to estimate the median forecast for Summer although these data are not representative of this season. The third group consists of methods whose performance improves as more data become available. This group comprises the ML methods proposed in this paper. We also include methods based on *S-ARIMA*, *ETS*, and *linear* forecast in this group. However, the performance of *S-ARIMA* and *ETS* stagnates for sample sizes larger than 0.6. This effect might be due to the fact that we use a little over one year of training data and consequently some months are included twice. It seems that the ML approaches can account for this matter. Thus, in the present application, the purely data-driven approaches benefit most from a large training set.

Comparing the different optimization methods, we find that with a sample size of $S = 0.4$ (150 days) and larger, the data-driven *SAA* method yields lower costs than its model-based counterpart *Norm* for most forecasting methods at a service level of 0.7. This observation implies that a normal distribution assumption is beneficial in our case only if a very limited dataset is available or if the target service level is very high (see Section 4.5.2).

The performance and the ranking of the methods varies depending on the sample size. However, if more data are available, it is possible to employ a method that reduces the costs compared to the best method on the smaller dataset. For sample size 0.1, *ETS Norm* is the best approach, while costs can be reduced by 17.4% using an *DT-LGBM Norm* with a sample size of 1.0.

5. Conclusion

In this study, we propose a framework for how data can be leveraged in inventory problems on three different levels: demand estimation, optimization, and integrated estimation and optimization. We highlight that integrated estimation and optimization in the newsvendor problem is equivalent to the Quantile Regression problem, and we introduce novel data-driven methods for the newsvendor problem based on Machine Learning and Quantile Regression. Moreover, we empirically compare the methods to well-established standard approaches on a real-world dataset. We are specifically interested in the effect of data-driven approaches on the three levels on the overall performance.

The key result of our evaluation is that data-driven approaches outperform their model-based counterparts in most cases. In our evaluation, this finding already holds for a demand history of beyond 25 weeks (i.e. 150 data points). However, overall performance depends heavily on the demand estimation method employed. We found that poor forecasts cannot be compensated for by the choice of the subsequent optimization approach. Thus, the selection of the forecast model is the most crucial decision in the case of separated estimation and optimization.

The empirical evaluation of the Quantile Regression approaches revealed that integrating forecasting and optimization is beneficial only if enough data are available to estimate the conditional quantiles and limited to target service levels smaller than 0.8. When working with single time series, separate estimation and optimization yields superior results. This finding is in line with the empirical anlysis of Ban and Rudin (2018).

More sophisticated estimation methods such as ANNs and Gradient Boosted Decision Trees require more training data in order to produce reliable results. However, these methods are also the only methods that constantly improve as more data becomes available. In our example, the demand history should contain more than six months of training data before employing Machine Learning. If a limited amount of data is available, simple methods such as the seasonal moving average can be suitable alternatives.

The major advantage of ML methods is that they are very flexible with respect to the input and that they are naturally able to process large datasets. The ability of ML methods to leverage similarities of time series across products and stores significantly improved their performance in our case. Additionally, they do not require restrictive assumptions on the demand process. Hence, they can identify patterns that traditional time series methods cannot detect. For instance, they can model multiple seasonalities (e.g. week and year), special days (e.g. public holidays), promotional activities and outliers (Barrow & Kourentzes, 2018). A drawback of these approaches is that they are a black box, which makes it more difficult to justify the resulting predictions. However, when the improvements in forecast accuracy can be easily measured, as in the case of baked goods, the advantage of accurate predictions should outweigh the issue of interpretability.

Data-driven inventory management is an active field of research with a variety of opportunities for future work. Our analysis is based on a particular data set of bakery products. It would be interesting to repeat the analysis on other data sets, including other products. The methodology is applicable to perishable products with repetitive sales (bread, fresh produce, newsprint,...). In other newsvendor situations, little or no historical sales data may be available (fashion, electronics, sport events,...). In that case, forecasting requires other leading indicators than historical sales. It will be interesting to investigate the performance of alternative approaches to derive decisions from data under those circumstances.

We presented a data-driven approach for the single-item newsvendor model. It seems natural to explore the multi-product case as well. Particularly in the bakery domain, it is a common practice to plan safety stocks on the product category level. This step is reasonable because the substitution rates within a category in the case of stock-outs are high for perishable goods (Van Woensel, Van Donselaar, Broekmeulen, & Fransoo, 2007). Thus, it could be possible to leverage hierarchical demand forecasts (Huber, Gossmann, & Stuckenschmidt, 2017) in order to optimize inventory and to make globally optimal decisions. Especially for the multi-product case, joint capacity restrictions and lot sizes should also be considered.

Some bakery products can be sold over multiple days. Thus, expanding the model to a multi-period inventory model is reasonable. It would widen the application of the model to many other grocery products that can be reordered during the selling season. There are several papers that deal with the the multi-period problem with unknown demand distribution (e.g. Godfrey & Powell, 2001; Levi et al., 2007). Given the inherent similarity between reorder point calculations and newsvendor trade-offs, one may expect machine learning approaches to also be beneficial in that context.

In our application, there is no lead time. However, in other problem settings lead time plays an important role. Prak, Teunter, and Syntetos (2017) show that using one-period-ahead forecast errors to optimize inventories leads to insufficient safety stock levels in case of a positive lead time.

In addition to the problem specific extensions, the methodology of the presented approaches may also be adjusted. Other machine learning approaches can be used for integrated forecasting and optimization, e.g., random forest or kernel methods.

Acknowledgments

This research was supported by *OPAL – Operational Analytics GmbH* (http://www.opal-analytics.com).

References

Ban, G.-Y., & Rudin, C. (2018). The big data newsvendor: practical insights from machine learning. *Operations Research, 67*(1), 90–108.
Barrow, D., Crone, S., & Kourentzes, N. (2010). An evaluation of neural network ensembles and model selection for time series prediction. In *Proceedings of the 2010 international joint conference on neural networks (IJCNN)*. Barcelona, Spain
Barrow, D., & Kourentzes, N. (2018). The impact of special days in call arrivals forecasting: a neural network approach to modelling special days. *European Journal of Operational Research, 264*(3), 967–977.
Ben-Tal, A., Ghaoui, L. E., & Nemirovski, A. (2009). *Robust optimization*. New Jersey: Princeton University Press.
Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.
Bertsimas, D., & Kallus, N. (2018). From predictive to prescriptive analytics. *Management Science, forthcoming*.
Bertsimas, D., & Thiele, A. (2006). A robust optimization approach to inventory theory. *Operations Research, 54*(1), 150–168.
Beutel, A. L., & Minner, S. (2012). Safety stock planning under causal demand forecasting. *International Journal of Production Economics, 140*(2), 637–645.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers and Geosciences, 37*(9), 1277–1284.
Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research, 184*(3), 1140–1154.
Conrad, S. A. (1976). Sales data and the estimation of demand. *Operational Research Quarterly, 27*(1), 123–127.
Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting, 27*(3), 635–660.
Crone, S. F., & Kourentzes, N. (2009). Forecasting seasonal time series with multilayer perceptrons-an empirical evaluation of input vector specifications for deterministic seasonality. In *Proceedings of the 2009 international conference on data mining* (pp. 232–238). Las Vegas, USA
Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.
Gallego, G., & Moon, I. (1993). The distribution free newsboy problem: review and extensions. *The Journal of the Operational Research Society, 44*(8), 825–834.
Godfrey, G. A., & Powell, W. B. (2001). An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Science, 47*(8), 1101–1112.
Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks, 4*(2), 251–257.
Huber, J., Gossmann, A., & Stuckenschmidt, H. (2017). Cluster-based hierarchical demand forecasting for perishable goods. *Expert Systems with Applications, 76*, 140–151.
Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software, 26*(3), 1–22.
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting, 22*(4), 679–688.
Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach*. New York: Springer.
Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting, 18*(3), 439–454.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3146–3154). Curran Associates, Inc.
Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *Proceedings of the international conference on learning representations 2015*. San Diego
Kleywegt, A. J., Shapiro, A., & Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization, 12*(2), 479–502.
Koenker, R. (2005). *Quantile regression*. New York: Cambridge University Press.
Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications, 41*(9), 4235–4244.
Lau, H.-S., & Lau, A. H. L. (1996). Estimating the demand distributions of single-period items having frequent stockouts. *European Journal of Operational Research, 92*(2), 254–265.
Levi, R., Perakis, G., & Uichanco, J. (2015). The data-driven newsvendor problem: new bounds and insights. *Operations Research, 63*(6), 1294–1306.
Levi, R., Roundy, R. O., & Shmoys, D. B. (2007). Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research, 32*(4), 821–839.
Oroojlooyjadid, A., Snyder, L. V., & Takác, M. (2016). Applying deep learning to the newsvendor problem. *CoRR*. arXiv: http://arxiv.org/abs/1607.02177
Perakis, G., & Roels, G. (2008). Regret in the newsvendor model with partial information. *Operations Research, 56*(1), 188–203.
Prak, D., & Teunter, R. (2018). A general method for addressing forecasting uncertainty in inventory models. *International Journal of Forecasting, In Press*. doi:10.1016/j.ijforecast.2017.11.004.
Prak, D., Teunter, R., & Syntetos, A. (2017). On the calculation of safety stocks when demand is forecasted. *European Journal of Operational Research, 256*(2), 454–461.
Qin, Y., Wang, R., Vakharia, A. J., Chen, Y., & Seref, M. M. H. (2011). The newsvendor problem: Review and directions for future research. *European Journal of Operational Research, 213*(2), 361–374.
R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Sachs, A. L., & Minner, S. (2014). The data-driven newsvendor with censored demand observations. *International Journal of Production Economics, 149*, 28–36.
Scarf, H. E. (1958). A min-max solution of an inventory problem. In K. J. Arrow, S. Karlin, & H. E. Scarf (Eds.), *Studies in the mathematical theory of inventory and production* (pp. 201–209). Stanford: Stanford University Press.
Shapiro, A. (2003). Monte carlo sampling methods. In A. Ruszczynski, & A. Shapiro (Eds.), *Handbooks in operations research and management science: 10* (pp. 353–425). Boston, USA: Elsevier Science B.V.
Silver, E. A., Pyke, D. F., & Thomas, D. J. (2017). *Inventory and production management in supply chains*. New York: Taylor and Francis.
Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research, 7*, 1231–1264.
Taylor, J. W. (2000). A quantile regression approach to estimating the distribution of multiperiod returns. *The Journal of Forecasting, 19*, 299–311.
Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research, 178*(1), 154–167.
Thomassey, S., & Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. *Decision Support Systems, 42*(1), 408–421.
Trapero, J. R., Cardós, M., & Kourentzes, N. (2018). Empirical safety stock estimation based on kernel and GARCH models. *Omega, In Press*, 1–13. doi:10.1016/j.omega.2018.05.004.
Van Woensel, T., Van Donselaar, K., Broekmeulen, R., & Fransoo, J. (2007). Consumer responses to shelf out-of-stocks of perishable products. *International Journal of Physical Distribution & Logistics Management, 37*(9), 704–718.
Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: the state of the art. *International Journal of Forecasting, 14*(1), 35–62.
Zhang, Y., & Gao, J. (2017). Assessing the performance of deep learning algorithms for newsvendor problem. In *Proceedings of the ICONIP 2017*. In *LNCS: 10634* (pp. 912–921). doi:10.1007/978-3-319-70087-8_93.