European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Data-driven inventory management under customer substitution

Sebastian Müller [a,*], Jakob Huber [b], Ralph Alexander Bubak [b], Moritz Fleischmann [c], Heiner Stuckenschmidt [b]

[a] Mainz University of Applied Sciences, School of Business, Lucy-Hillebrand-Str. 2, 55128, Mainz, Germany
[b] University of Mannheim, School of Business Informatics and Mathematics, B6, 26, 68159, Mannheim, Germany
[c] University of Mannheim, Business School, Schloss, 68161, Mannheim, Germany

ARTICLE INFO

Keywords:
Inventory
Multi-product
Substitution
Machine learning

ABSTRACT

Retailers selling perishable goods typically offer multiple products in a product category, e.g., fresh food or fashion. Managing the inventories of these products is especially challenging due to frequent stock-outs and resulting substitution effects within the category. Furthermore, the true product-specific demand distributions are usually unknown to the decision maker. Digital technologies have massively expanded the available data and computing power, which can help improve inventory decisions. In this paper, we present a novel solution approach for the multi-product newsvendor problem. Our method is based on modern machine learning techniques that leverage large available datasets,e.g., data on historical sales, weather, store locations, and special days, and are able to take complex substitution effects into account. We evaluate our approach on two real-world datasets of a large German bakery chain and find that our data-driven approach outperforms the benchmark on the first dataset and performs competitively on the second dataset. We then analyze our approach in a controlled environment with synthetic data to pinpoint the factors that determine its performance. While saving computation time, our approach outperforms relevant benchmarks for a wide range of cost parameters and substitution levels when the demand distribution depends on exogenous data; it is competitive otherwise.

1. Introduction

Stock-outs are common in many retail settings. When customers cannot find their preferred product in stock, some choose an alternative product instead (Gruen et al., 2002; van Woensel et al., 2007). This substitution behavior makes stocking decisions interdependent. As a result, inventory optimization for multi-product portfolios is more challenging than for single products. A second major challenge is that demand distributions are usually unknown and have to be estimated. In this paper, we formalize the sketched problem setting as a multi-product newsvendor problem with one-step substitution, exogenous substitution rates, unknown demand distributions, and explanatory variables ("features"). We focus on the single-period case for perishable goods, where unsold inventory has no salvage value and unmet demand is lost, and adopt a profit-maximizing perspective. To solve the problem we propose and analyze a novel Machine Learning (ML)-based approach. We benchmark it against a traditional baseline on both real and synthetic data, identify factors that drive its performance, and derive managerial insights.

Traditional solution approaches for the multi-product newsvendor problem proceed in three steps: (i) data preprocessing, e.g., demand de-censoring, estimation of substitution rates, (ii) demand estimation and (iii) inventory optimization[1] Hübner et al. (2016), Kök and Fisher (2007), Zhang et al. (2021a). We refer to this approach as separated estimation and optimization (SEO). Despite its prevalence in research and practice, the separation of estimation and optimization is, in general, suboptimal. First, choosing a demand model based on traditional forecast-error criteria,e.g., mean squared error, in the estimation step does not necessarily lead to a model that produces good decisions in the optimization step (Bertsimas & Kallus, 2020; den Boer & Sierag, 2021; Prak et al., 2017; Qi et al., 2023a). Second, the conventional estimation step captures the impact of features on the level of the point forecast only, if at all, but ignores their impact on other characteristics of the demand distribution. Third, updated demand distribution estimates require a de novo execution of the optimization step, which is computationally expensive, especially in dynamic retail environments with many products and categories.

In contrast, we present a novel method using artificial neural networks (ANNs) that builds on the idea of integrated estimation and

[1] The step of data preprocessing can be omitted if uncesored demand data and substitution rates are available. This is rarely the case in practice.

* Corresponding author.
E-mail addresses: sebastian.mueller@hs-mainz.de (S. Müller), jakob.huber@uni-mannheim.de (J. Huber), alexander.bubak@uni-mannheim.de (R.A. Bubak), moritz.fleischmann@uni-mannheim.de (M. Fleischmann), heiner.stuckenschmidt@uni-mannheim.de (H. Stuckenschmidt).

https://doi.org/10.1016/j.ejor.2025.12.034
Received 6 May 2025; Accepted 17 December 2025
Available online 25 December 2025

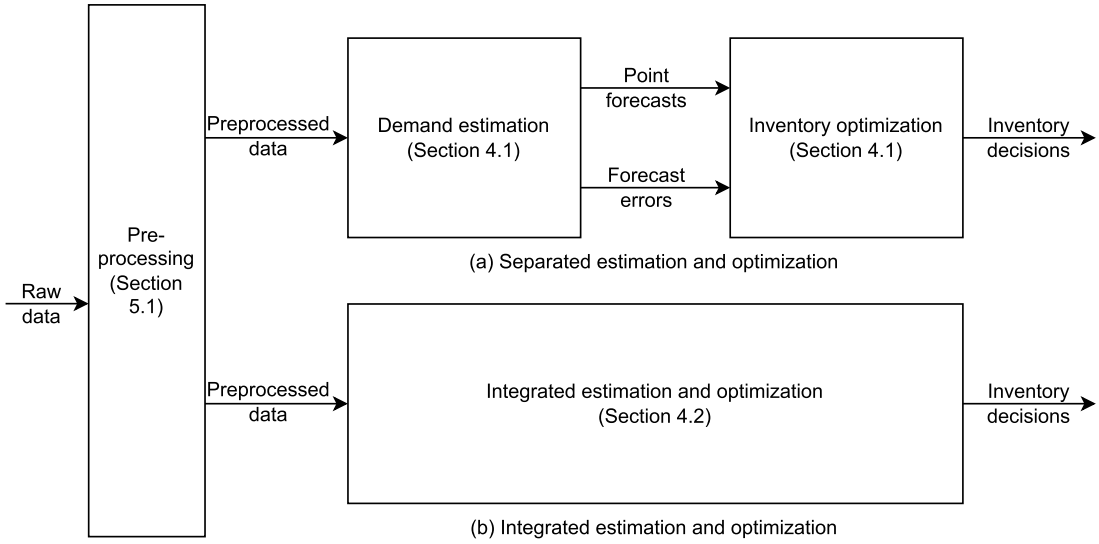


Fig. 1. Separate vs. Integrated Estimation and Optimization (adapted from Huber et al. (2019)).

optimization (IEO). As in the SEO approach, preprocessing of raw sales data is required. However, once preprocessing is completed, our approach integrates estimation and optimization. The model uses ANNs to leverage available data and prescribe ordering decisions for all products, taking into account substitution effects. After being trained, our model yields ordering decisions instantaneously, without re-solving the notoriously difficult multi-product newsvendor problem in each period. The difference between the separated and the integrated approach are shown in Fig. 1.

Similar approaches exist and perform well for related problems, such as the single-product case (Ban & Rudin, 2019; Beutel & Minner, 2012; Huber et al., 2019; Liu et al., 2022; Liyanage & Shanthikumar, 2005). The latter is comparatively simple, as the optimal solution is known in closed form and the objective function is convex. We extend the IEO approach to the more complex multi-product setting, where these circumstances (closed-form solution and convex objective) no longer apply. To the best of our knowledge, we are the first to propose integrated non-linear decision rules for the multi-period newsvendor.

We test our approach in a real-life setting using two datasets from a large German bakery chain. We find that IEO outperforms the traditional approach for one dataset and is competitive for the other, with less computational effort. To generalize and explain our observed results, we perform experiments in a controlled environment using synthetic data. We synthesize our findings by linking the benefit of our IEO approach over a traditional SEO to a distance metric between probability distributions.

To summarize, our main contributions in this paper are as follows:

- We present a novel data-driven solution approach for the multi-product newsvendor problem with substitution.
- We document the viability of our method for both synthetic and real-world data.
- We identify factors that drive the performance of our integrated approach and capture them in a metric that is applicable beyond the field of inventory management.
- We use real-world data to derive managerial insights for retailers, concerning the impact of data availability and substitution on inventory management decisions.

The remainder of this paper is structured as follows. In the next section, we summarize the related literature. Section 3 contains the formal problem description of the multi-product newsvendor problem with unknown demand distributions. In Section 4, we introduce our solution approach. In Sections 5 and 6, we report and discuss the results of the empirical evaluation on real-world and synthetic data, respectively. Finally, we summarize our findings and outline opportunities for further research in Section 7.

2. Related literature

In this section, we review several streams of literature related to our work, namely ML literature on demand forecasting, OR literature on the multi-product newsvendor problem with substitution and on integrated estimation and optimization, and empirical studies on substitution rates.

A large part of ML is concerned with the prediction of a certain quantity,e.g., demand, given a large dataset of features. For an overview of ML methods, see Hastie et al. (2008). We restrict our review to references that use ML for demand forecasting. Carbonneau et al. (2008) use several ML methods to forecast demand in a supply chain. They find that recurrent neural networks and support vector machines are the most accurate forecasting techniques for their dataset. Barrow et al. (2010) study the performance of different neural network setups in a time-series forecasting task. Crone et al. (2011) report the results of the NN3 forecasting competition with a focus on ANNs. The results suggest that ANNs perform competitively with traditional statistical approaches but do not outperform them. Barrow and Kourentzes (2018) use ANNs to predict call center arrivals from time-series data with a focus on outlying arrival periods. In general, the above ML literature is concerned with prediction alone and does not explicitly consider the impact on decision making. In our approach, we use ANNs to get from preprocessed data to decisions, for a multi-product newsvendor problem.

Many studies in the OR literature investigate the structural properties of the multi-product newsvendor problem under customer substitution and related solution algorithms. Kök et al. (2015) provide a broad review of the topic. Two main modeling approaches for substitution can be distinguished: a) consumer choice models, which consider consumers to be rational utility maximizers and derive consumer behavior from a utility model, e.g., a multinomial logit model (Farahat & Lee, 2018; Musalem et al., 2010; van Ryzin & Mahajan, 1999; Topaloglu, 2013; Vulcano et al., 2010, 2012), and b) so-called exogenous demand models, which represent substitution through explicit substitution rates. We follow the latter approach, which is prevalent in the inventory management literature (Kök & Fisher, 2007) and more suitable for our intention since it avoids the nonlinearities and estimation complexities of choice models.

For the two-product case, Parlar and Goyal (1984) show that the multi-product newsvendor objective function is concave under mild

conditions, which allows for tractable analysis and yields necessary optimality conditions. However, when extending the problem to more than two products, Netessine and Rudi (2003) demonstrate that the objective function is generally no longer concave. More recently, Schlapp and Fleischmann (2018) extend the model further by incorporating capacity restrictions in addition to substitution. Collectively, these studies illustrate that while early results provided structural insights in special cases, the general multi-product substitution problem remains analytically challenging and has motivated our data-driven approach.

Another stream of extant research on the multi-product newsvendor problem focuses on efficient optimization algorithms. Zhang et al. (2021a) and Zhang et al. (2021b) develop mixed-integer linear programs that are able to solve problems of realistic sizes for many applications. For very large problems, they provide approximation algorithms. Closely related to our work is Kök and Fisher (2007), who describe a step-by-step approach from the estimation of substitution rates and demand distributions to the final inventory decision. They develop a heuristic and apply it to the setting of a large supermarket chain. The approach results in a large profit increase relative to current practice. Hübner et al. (2016) build on the work of Kök and Fisher (2007) and develop an exact solution procedure as well as further heuristics.

The extant work on the multi-product newsvendor problem relies on SEO. Besbes et al. (2010) and den Boer and Sierag (2021) discuss the general problem of selecting decision models based on goodness-of-fit criteria, as opposed to decision-oriented criteria. Bertsimas and Kallus (2020) present a general concept of SEO with different ML methods, such as k-nearest-neighbor regression, Kernel methods, and Random Forests. In this paper, we integrate demand estimation and inventory optimization into a single optimization problem. Integrated approaches have witnessed a marked increase in popularity in recent times. Qi et al. (2023a) and Elmachtoub et al. (2023) develop frameworks to analyze the properties of IEO and its benefits over SEO statistically. The authors acknowledge that a computational study is a separate yet important aspect, which we address in our work. Our results, particularly those presented in Section 6, support their findings. Chen et al. (2022) propose a semi-integrated approach for settings in which the objective function is not explicitly known. Concurrently, both Qi et al. (2023b) and Liu et al. (2024) solve a multi-period inventory management problem using an IEO approach by initially generating ex-post optimal replenishment decisions from historical data and subsequently training a deep learning model on the labeled dataset. Closer to our work, numerous researchers investigate integrated approaches for the single-product newsvendor (Ban & Rudin, 2019; Beutel & Minner, 2012; Huber et al., 2019; Liu et al., 2022; Liyanage & Shanthikumar, 2005; Oroojlooyjadid et al., 2020). We emphasize that the single-product case is much simpler, as the solution is known in closed form, the objective function is convex and the empirical risk minimization approach reduces to a simple case of quantile regression. None of these characteristics transfer to the multi-product case. Sachs (2015) approaches the two-product case with linear decision rules. Olivares-Nadal (2024) builds upon Zhang et al. (2021a) and a previous version of this article (Müller et al., 2020) to design constant, linear and feature map-induced non-linear decision rules for the multi-product newsvendor. Especially the non-linear approach relies on computationally expensive quadratic programs, which gives rise to concerns regarding scalability with regard to both the size of the training data and the product assortment. To the best of our knowledge, there is no research on the multi-product newsvendor leveraging ANNs.

To estimate exogenous substitution rates, Anupindi et al. (1998) propose a maximum likelihood estimation (MLE) that works with inventory transaction data. They test their approach on data from vending machines. Kök and Fisher (2007) generalize this approach to dynamic choice processes. Fisher and Vaidyanathan (2014) also use MLE. Wan et al. (2018) compare the accuracy of a nested logit customer choice model to exogenous substitution rates in a multi-store environment. For the estimation of the substitution rates in our empirical evaluation, we adapt the methodology of Anupindi et al. (1998).

Several empirical studies measure substitution rates for different product categories. For ground coffee, orange juice, peanut butter, tomato sauce, and toothpaste, Emmelhainz et al. (1991) found that 65% to 83% of the customers substitute in response to a stock-out. Campo et al. (2000) find substitution rates of 44% and 51% for cereals and margarine, respectively. The most extensive study by Gruen et al. (2002) found that substitution rates vary significantly by category and are approximately 45% on average. Most related to our research is the work of van Woensel et al. (2007). The authors investigate consumer responses to stock-outs of bakery bread and find that approximately 82% of the customers are willing to substitute for another product if their first choice is unavailable. From these empirical studies, we conclude that substitution rates are high across different product categories and thus that the multi-product nature of the problem should be taken into account when making inventory decisions.

To summarize, the literature on ML and OR is still relatively disjointed. The ML literature in the retail context is mostly concerned with estimating demand distributions and does not consider optimal decision making. The OR literature on the multi-product newsvendor problem focuses on structural properties and efficient optimization algorithms with strong distributional assumptions. To avoid these strong assumptions, we propose a method that integrates demand estimation and inventory optimization. Our approach leverages large datasets, is able to reflect the inherently challenging substitution effects, and prescribes feature-dependent inventory decisions without the need for solving an optimization problem in each period.

3. Problem formulation

We study the multi-product newsvendor problem with stockout-based substitution and unknown demand distributions. As a starting point, we recap the prevalent model formulation in the inventory management literature (Kök et al., 2015; Netessine & Rudi, 2003).

Consider a retailer selling n partially substitutable products with uncertain demand D_i for product i over a finite selling season. The retailer must choose the order quantity q_i of product i before the selling season and seeks to maximize the expected total profit Π over all products. The unit sales price of product i is p_i, and the unit cost is c_i. Unsold units of product i that are left over at the end of the season have a unit salvage value of s_i. Naturally, $p_i > c_i > s_i \geq 0$.

To model substitution, we assume that a fraction $\alpha_{ji} \in [0,1]$ of customers who cannot find their preferred product j in stock,i.e., when $D_j > q_j$, will substitute for product i, where $\sum_{i \neq j} \alpha_{ji} \leq 1$. This substitution behavior results in an inflation of the *initial demand* D_i for product i. The *substitution demand* for product i is $D_i^s = D_i + \sum_{j \neq i} \alpha_{ji}(D_j - q_j)^+$. This approach only considers one-step substitution: If customers attempt to substitute product j with i and product i is also out of stock, their demand is lost.

Thus, the retailer's objective is to maximize the expected total profit according to

$$\max_{q_i \geq 0} \Pi = \mathbb{E}\left[\sum_i \left[u_i D_i^s - u_i(D_i^s - q_i)^+ - o_i(q_i - D_i^s)^+\right]\right], \tag{1}$$

$$= \mathbb{E}\left[\sum_i \left[u_i q_i + u_i(D_i^s - q_i)^+ - u_i(q_i - D_i^s)^+ - u_i(D_i^s - q_i)^+ - o_i(q_i - D_i^s)^+\right]\right], \tag{2}$$

$$= \sum_i \left(u_i q_i - (u_i + o_i)\mathbb{E}\left[\left(q_i - D_i^s\right)^+\right]\right), \tag{3}$$

where $u_i = p_i - c_i$ and $o_i = c_i - s_i$ are the underage and overage costs of product i, respectively.

The major difficulty in solving problem (1) is that D_i^s depends on the order quantities of the other products $j \neq i$. Netessine and Rudi (2003) show that the objective function with more than two products is not necessarily concave or quasiconcave, which makes the problem particularly hard to solve.

In contrast to the traditional version of the problem, we assume that the probability distributions of the initial demand D_i for each product i are not known to the decision maker a priori. Instead, historical data $S_T = \{(\mathbf{d}_1, \mathbf{x}_1), \dots, (\mathbf{d}_T, \mathbf{x}_T)\}$ are available, where $\mathbf{d}_t = [d_{1,t}, \dots, d_{n,t}]$ is a vector of historical demand realizations for products $1, \dots, n$ in period t, and $\mathbf{x}_t = [x_{1,t}, \dots, x_{m,t}]$ is a vector of m covariates or *features* (e.g., store location, opening hours, and weather data) in period t. We acknowledge that historical data frequently reports sales instead of demand and only carries implicit information on the substitution rates α_{ji}. These issues are addressed during data preprocessing as presented in Section 5.1. In particular, we suggest to decensor sales to arrive at demand values. Therefore, in the following, we can safely assume all models can be trained on demand data.

The traditional approach to solving this problem would be first to estimate a model for the distribution of D_i for every product i from these data and then to optimize the actual inventory problem (1). Each time the distribution estimate changes,e.g., due to seasonality or the influence of other features, the inventory problem needs to be resolved. We propose an alternative solution approach that integrates the estimation into the optimization problem and is solved only once. We elaborate on our integrated method and the traditional separated approach in the following section and empirically compare them in Sections 5 and 6.

4. Solution approaches

4.1. "Traditional" SEO benchmark

Most of the literature on the multi-product newsvendor problem neglect the fact that the demand distributions are unknown to the decision maker (Kök et al., 2015; Netessine & Rudi, 2003; Schlapp & Fleischmann, 2018). Extant papers that do address both demand estimation and inventory optimization take an SEO approach.

We consider two alternative SEO approaches as benchmarks in our numerical experiments. The approaches differ in the way they estimate the demand. The first approach uses only historical demand data whereas the second approach has access to additional feature data. Specifically, our first SEO approach incorporates exponential smoothing (ETS; Hyndman and Athanasopoulos (2014)), which is a popular family of time series methods that have been shown to perform well in a single product setting, even compared to more complex ML methods that leverage additional feature data (Huber et al., 2019). In our implementation, we use the `ets()` function from the `forecast` package (Hyndman & Khandakar, 2008) for the statistical software R (Team, 2017) to fit an ETS model of type ANA, also known as Holt-Winters' additive method. This ETS variant, which employs additive errors, no trend and an additive seasonality, has been identified as the best fit to our data (see Section 5.1).

For our second SEO approach, we implement an artificial neural network (ANN) to estimate the demand. We consider this approach for two reasons: First, we can use it to analyze the contribution of demand feature information by comparing it to the ETS-based SEO method. Second, it allows an impartial comparison of SEO and IEO methods, as both ANN-SEO and IEO use the same data. We elaborate on ANNs in more detail in Section 4.2.

The demand prediction model – either ETS or ANN – provides point forecasts. Given a demand prediction $\hat{d}_i$ for the upcoming period for each product i, the next step is to solve the original multi-product newsvendor problem (1). This problem is known to be notoriously hard due to the nonconcave objective function (Netessine & Rudi, 2003). Therefore, few efficient solution algorithms are available that provide near-optimal solutions in reasonable time for real-world problem sizes (Hübner et al., 2016; Zhang et al., 2021a). In our SEO benchmarks, we use the mixed-integer linear program (MILP) formulation proposed by Zhang et al. (2021a). We summarize the formulation below and refer to Zhang et al. (2021a) for more details.

$$\max_{q_i \geq 0} \sum_i \left(u_i q_i - (u_i + o_i) \frac{1}{T} \sum_t y_{i,t} \right) \tag{4}$$

subject to:

$$y_{i,t} \geq q_i - \hat{d}_{i,t} - \sum_{j \neq i} \alpha_{ji} v_{j,t} \qquad \forall i, t \tag{5}$$

$$v_{i,t} \leq \hat{d}_{i,t} - q_i + M_i z_{i,t} \qquad \forall i, t \tag{6}$$

$$v_{i,t} \geq \hat{d}_{i,t} - q_i - M_i z_{i,t} \qquad \forall i, t \tag{7}$$

$$v_{i,t} \leq \hat{d}_{i,t}(1 - z_{i,t}) \qquad \forall i, t \tag{8}$$

$$v_{i,t}, y_{i,t} \geq 0 \qquad \forall i, t \tag{9}$$

$$z_{i,t} \in \{0, 1\} \qquad \forall i, t \tag{10}$$

The MILP formulation replaces the expectation of the original problem by an empirical average (compare Eqs. (1) and (4)). For the latter, we require T demand samples per product i, $\left(\hat{d}_{i,t}\right)_{t=1,\dots,T}$, which are constructed by predicting each product's demand for the upcoming period and adding the past T prediction errors. Given these demand samples, the variables $y_{i,t} = \left(q_i - \hat{d}_{i,t} - \sum_{j \neq i} \alpha_{ji} (\hat{d}_{j,t} - q_j)^+ \right)^+$ and $v_{i,t} = (\hat{d}_{i,t} - q_i)^+$ represent overages and underages, respectively. Constraints (5) to (9) set the overage and underage variables appropriately. To linearize the $(\cdot)^+$ functions, the formulation uses binary variables $z_{i,t}$, where $z_{i,t} = 1$ if $q_i \geq \hat{d}_{i,t}$ (i.e., $v_{i,t} = 0$) and $z_{i,t} = 0$ if $q_i < \hat{d}_{i,t}$ (i.e., $v_{i,t} = \hat{d}_{i,t} - q_i$). M_i is an upper bound of the order quantity q_i of product i.

4.2. IEO approach

We compare the above benchmark approaches to our newly developed IEO approach that uses ML to determine feature-dependent order quantities directly from the data. To this end, we express the order quantities $q_{i,t}$ as a function of the feature vector $\mathbf{x}_t$ and parameters $\mathbf{W}$ of an ML algorithm, e.g., the weight matrix of an ANN. We use the objective function of the multi-product newsvendor problem as the loss function of the ML model as follows:

$$\max_{\mathbf{W}} \frac{1}{n} \sum_t \sum_i \left(u_i q_{i,t}(\mathbf{W}, \mathbf{x}_t) - (u_i + o_i) \left(q_{i,t}(\mathbf{W}, \mathbf{x}_t) - d^s_{i,t} \right)^+ \right), \tag{11}$$

where $d^s_{i,t} = d_{i,t} + \sum_{j \neq i} \alpha_{ji} (d_{j,t} - q_{j,t}(\mathbf{W}, \mathbf{x}_t))^+$.

Given the historical dataset of demand and features $S_T = \{(\mathbf{d}_1, \mathbf{x}_1), \dots, (\mathbf{d}_T, \mathbf{x}_T)\}$, we can train the network, i.e., optimize the parameters $\mathbf{W}$ in (11). Note that in contrast to the SEO approach, the decision variables of the optimization problem are no longer the inventory decisions but rather the parameters of a non-linear regression model. These parameters are estimated by maximizing the expected profit objective (Eq. (11)), so that the resulting mapping $\mathbf{q}(\mathbf{W}, \mathbf{x})$ prescribes profit-maximizing order quantities for new data samples. In this sense, the ANN does not serve as a descriptive regression model, but as a prescriptive descision rule learned from historical data. This approach parallels existing work on the single-product newsvendor, in that there, too, regression models are trained on the problem-specific objective function (cf. Eq. (8) in Huber et al. (2019)). The difference is that the single-product loss function is so well-behaved that the optimization problem reduces to a simple case of quantile regression. In the multi-product case, the substituted demand for one product depends on the order decisions for the others and the objective function is not necessarily (quasi-)concave, so it is yet to be analyzed if a regression model can solve this task satisfactorily.

While (11) holds for a generic regression model, we choose an ANN in our approach because it is able to approximate any continuous function (Hornik, 1991), which makes it suitable for forecasting if enough data are available and it is difficult to specify the underlying data generation process. We rely on feed-forward neural networks, i.e., multilayer perceptrons. The latter have been shown to work well in the

Table 1
Hyperparameter spaces for ANN tuning.

Hyperparameter	`hyperopt` search space	Range ([lower bound, upper bound])
Learning rate	`quniform(low=0.0001, high=0.01, q=0.00001)`	[0.0001, 0.01]
Patience	`quniform(low=1, high=50, q=1)`	[0, 50]
Batch size	`qloguniform(low=2.3, high=4.85, q=1)`	[10, 128]
Num. of hidden layers	`choice([1, 2, 3])`	[1, 3]
Num. of hidden nodes	`qloguniform(low=2.3, high=4.85, q=1)`	[10, 128]

single-product case (Huber et al., 2019; Oroojlooyjadid et al., 2020). In a feed-forward neural network with L hidden layers ($L \geq 1$), the output $\mathbf{h}^{(k)}(\mathbf{x})$ of each layer k is passed to the next layer ($1 \leq k \leq L+1$):

$$\mathbf{h}^{(k+1)}(\mathbf{W}, \mathbf{x}_t) = \sigma^{(k+1)}(\mathbf{b}^{(k+1)} + \mathbf{W}^{(k+1)}\mathbf{h}^{(k)}(\mathbf{x})) \tag{12}$$

The output of the input layer is defined as $\mathbf{h}^0(\mathbf{x}) = \mathbf{x}$, while the output of the last layer represents the prescribed order quantities, i.e.,$\mathbf{q}(\mathbf{W}, \mathbf{x}) = \mathbf{h}^{(L+1)}(\mathbf{W}, \mathbf{x})$. The output of each layer is connected with a fully connected weight matrix $\mathbf{W}^{(k)}$ to the next layer. The input of a layer is adjusted with the biases $\mathbf{b}^{(k)}$ of each neuron before it passes an activation function $\sigma^{(k)}$.

The performance of an ANN depends on its architecture and the hyperparameter settings. We implement our ANNs in `PyTorch` and optimize these hyperparameters (learning rate, patience for early stopping, batch size, number of hidden nodes, and number of hidden layers) with `hyperopt`'s tree of parzen estimators (Bergstra et al., 2011, 2013) in combination with ten-fold cross-validation of the training set. We define the hyperparameter search spaces as reported in Table 1. We use rectified linear units (RELUs) as activation functions at the input layer and the hidden layers and a linear activation function at the output layer (Zhang et al., 1998). Moreover, we train and employ an ensemble of six ANNs with the *median* ensemble operator, as this approach is robust to the initial weights (Barrow et al., 2010; Kourentzes et al., 2014). The models selected for the ensemble are those with the lowest mean validation loss. To optimize the weights of an ANN, we use the stochastic gradient-based algorithm ADAM proposed by Kingma and Ba (2015) in combination with early stopping to avoid overfitting.

5. Empirical evaluation

5.1. Data description and preparation

We test the validity of our approach in a real-world setting using data from a large German bakery chain. In this section, we describe the datasets and the data preparation process. The datasets comprise the hourly sales data of the six most frequently sold stock-keeping units, both from the product categories buns and breads, for nine stores over a period of 987 days. In addition, the datasets contain an hourly binary stockout indicator. In approximately 24% and 28% of sales hours, a stockout occurred for buns and breads, respectively, allowing for the estimation of substitution rates (see below). Figs. 2–5 give a detailed overview of our sales data.

Fig. 2 depicts the strong hourly seasonality in our data. While sales of buns exhibit one peak around 8 a.m. and abruptly decline again after 9 a.m., sales of breads exhibit two peaks – around 10 a.m. and 4 p.m. – and are slightly more evenly distributed in general. These observations hold on average as well as for the individual store-article combinations. This intraday non-stationarity is important for substitution rate estimation (see below). With our solution approaches we predict daily demand and optimize daily order quantities. Fig. 3 depicts daily aggregated sales. Again, we observe a strong seasonality for the product group buns with fairly constant sales throughout the week and noticeable peaks on weekends, in particular on Sundays. Sales for breads are distributed more evenly with a slight increase on Mondays and Saturdays and a noticeable peak on Sundays. These patterns hold for both average sales and individual profiles except on Sundays, where some store-article combinations are noticeably more popular than others. In the category breads, only one article is for sale on Sundays. Our observations suggest calendar-related information to be valuable sales predictors. Fig. 4 plots a rolling mean of the average daily sales over all store-article combinations by product group with. The selected window size of seven days smooths out the weekly seasonality, allowing to detect trending behavior. From the graph we conclude that our data is not subject to noticeable trends. Finally, Fig. 5 plots the standard deviation of the daily demand percentage against its mean. Each marker represents one weekday, the dotted lines are fitted linear regressions and illustrate a slight positive correlation. We can observe that both categories exhibit moderate heteroscedastic errors, i.e., sales variability increases moderately with sales level. We emphasize that these observations – strong weekly seasonality and no trend – have informed our decision to select the ANA variant of the ETS family (cf. Section 4.1). While the daily sales distributions exhibit some heteroscedasticity, we made the decision to model errors in an additive way since our data is not strictly positive, in which case multiplicative ETS models are numerically unstable Hyndman and Athanasopoulos (2014). In addition, since we only use the ETS model for point forecasts, the error model is of relatively low importance Hyndman and Athanasopoulos (2014).

We expand the datasets by adding master data, transactional data, and external explanatory features related to the calendar, weather, and store location (see Table 2). All weather-related variables are forecasts and therefore suitable as predictors.

For our assessment, we further need price and cost parameters, substitution rates, and daily demand data. We report price and cost parameters for both product categories in Tables 3 and 4. Detailed historical price data was not available. Therefore, prices were used only to calculate cost parameters of the newsvendor model, not as model features. Unsold units have no salvage value, which reflects the perishability of the products considered, e.g., bakery items.

We estimate the substitution probabilities similar to Anupindi et al. (1998). Based on the transaction data, the method measures the spillover demand from stocked-out products to available products. From the magnitude of spillover demand in stock-out periods compared to non-stock-out periods, the substitution rates can be estimated. As the assumption of stationary demand during the day does not hold in our case due to a strong intraday sales pattern, we apply the approach to each hour of the day. We obtain estimates for the substitution matrices for each hour of the day and then compute the average over all hours. The results are shown in Tables 5 and 6.

Overall, we estimate that 43% to 75% of customers are willing to buy another product if their first choice is not available. To validate our results, we compare our estimates to earlier empirical work on substitution. van Woensel et al. (2007) found rates of 75% to 82% for bakery products. However, the product portfolio in their study was much larger (208 products); therefore, it is also more likely that a substitution with a similar product can take place and that substitution rates are higher. We note that in the category buns, the substitution rates to product 1 are relatively high. The same is true for product 4 in the category breads. These products stand out, as they have the lowest prices within their respective categories.

A main drawback of point-of-sale data,e.g., our dataset, is that lost demand is unobservable in cases of stock-outs. Frequent stock-outs distort sales data in two ways. First, the sales data of the out-of-stock product are *censored* if customers cannot find their preferred product and

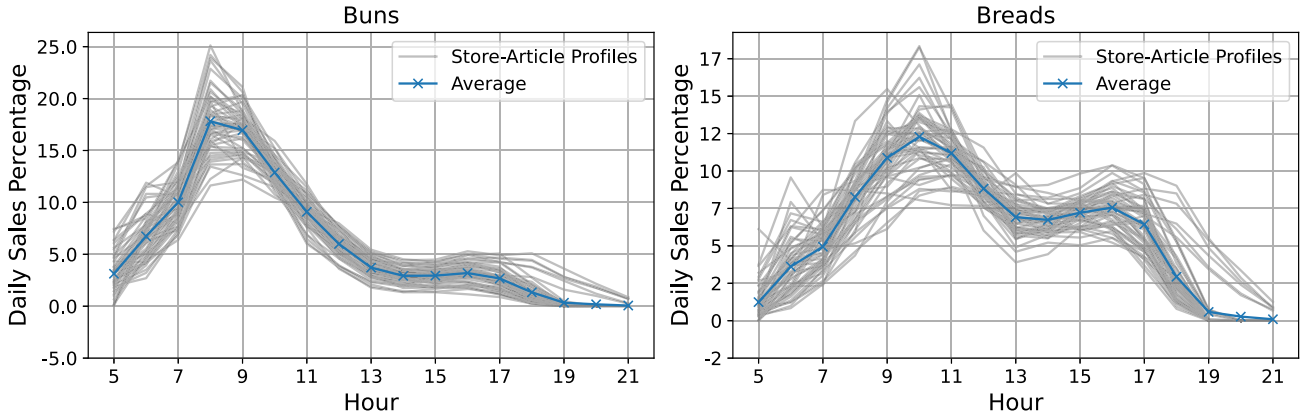


Fig. 2. Daily sales percentages for all opening hours for buns and breads.

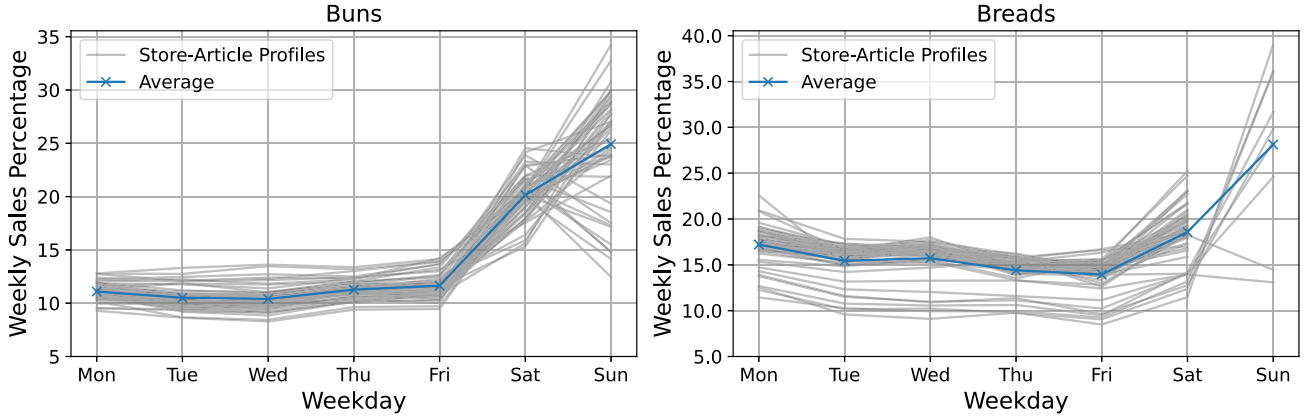


Fig. 3. Weekly sales percentages for all open weekdays for buns and breads.

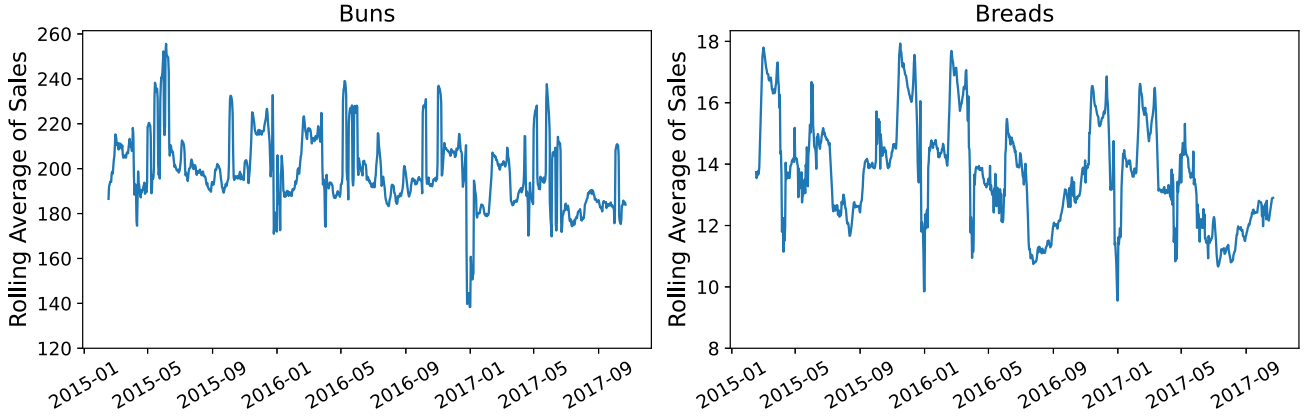


Fig. 4. Rolling mean of daily demand for buns and breads, averaged over all stores and articles in one product category. The rolling window size is seven days.

Table 2
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 of year, day of month, day of week, public holiday, day type, bridge day, non-working day, indicators for each special day, school holidays
Weather	temperature (minimum, mean, and maximum) and cloud cover of target day
Location	general location (city, suburb, and town); in proximity to the store: shops (bakeries, butcher, grocery, kiosk, fast food, and car repair), amenities (worship, medical doctors, and hospitals), leisure (playground, sports facility, and park), education (kindergarten, school, and university)

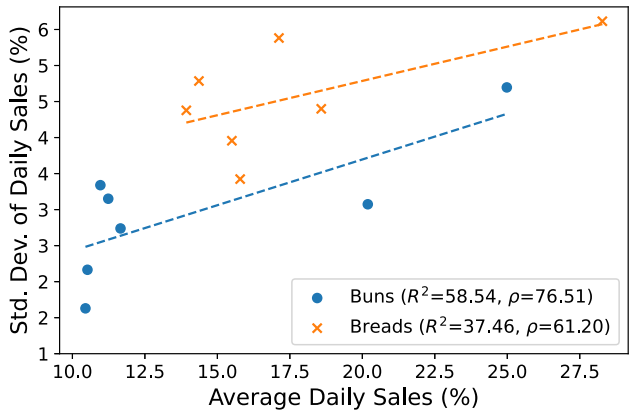


Fig. 5. Standard deviation vs. mean of daily sales percentage. Each marker represents one weekday. The dotted lines are fitted linear regressions. The legend reports the R^2 of the linear regression and the coefficient of correlation ρ for each product category.

Table 3
Price and cost parameters for buns.

Buns	P1	P2	P3	P4	P5	P6
Unit price [EUR]	0.30	0.50	0.60	0.50	0.50	0.50
Unit cost [EUR]	0.06	0.06	0.06	0.06	0.06	0.06

Table 4
Price and cost parameters for breads.

Breads	P1	P2	P3	P4	P5	P6
Unit price [EUR]	1.75	2.45	2.45	1.70	2.85	2.95
Unit cost [EUR]	0.60	0.60	0.60	0.60	0.60	0.60

Table 5
Estimated substitution rates of buns (from row to column).

↗	P1	P2	P3	P4	P5	P6	Total
P1	–	0.11	0.15	0.14	0.14	0.15	0.69
P2	0.22	–	0.08	0.11	0.12	0.12	0.65
P3	0.24	0.07	–	0.07	0.08	0.07	0.53
P4	0.26	0.09	0.07	–	0.12	0.12	0.66
P5	0.19	0.10	0.10	0.12	–	0.14	0.65
P6	0.17	0.13	0.11	0.11	0.13	–	0.65

Table 6
Estimated substitution rates of breads (from row to column).

↗	P1	P2	P3	P4	P5	P6	Total
P1	–	0.14	0.09	0.25	0.12	0.12	0.72
P2	0.19	–	0.09	0.19	0.14	0.14	0.75
P3	0.07	0.09	–	0.07	0.10	0.10	0.43
P4	0.17	0.16	0.09	–	0.20	0.10	0.72
P5	0.18	0.12	0.09	0.24	–	0.12	0.75
P6	0.16	0.13	0.08	0.22	0.15	–	0.74

choose to substitute another product or leave the store without buying anything at all. Second, the sales data of the substitute products are *inflated* by the demand of the out-of-stock products. To address these problems, we *decensor* and *deflate* historical sales data by applying the following procedure to each product category on each day:

The *decensoring* of sales in Step 4 is based on Lau and Lau (1996). In particular, we calculate the average hourly share of demand for product $\tilde{P}$ in relation to the total demand on days on which the product was not sold out. Based on this averaged intraday demand pattern, we extrapolate the sales data when a stock-out occurs to estimate the uncensored demand. To *deflate* the demand of the products in L, we subtract the sales in each hour of stock-out of product $\tilde{P}$ that are due to substitution demand Algorithm 1. Table 7 shows the average share of the daily demand of each product within each category. While this proportion is relatively homogeneous for breads (8.5% to 23.8%), product 1 dominates buns with a share of 64.3%.

Define the set of all products in the category as $L = \{P_1, \ldots, P_n\}$;
while *at least one product in L goes out of stock* **do**
  find product $\tilde{P}$ that goes out of stock first;
  delete $\tilde{P}$ from L;
  decensor the demand of product $\tilde{P}$ based on Lau and Lau (1996);
  deflate the demand of all products in L based on substitution rates;

Table 7
Average share of the demand of each product within each category.

Category	P1	P2	P3	P4	P5	P6
Buns	0.643	0.107	0.070	0.068	0.057	0.055
Breads	0.220	0.139	0.205	0.238	0.113	0.085

We split the dataset into a training set of 110 weeks worth of daily data and a test set containing the remaining 31 weeks, and we perform a rolling 1-day-ahead prediction evaluation on the test set to assess the performance of the methods. Using a sliding window approach, we fit the models and determine the error distributions every 14 days on a rolling training dataset of constant size (110 weeks).

5.2. Results

In this section, we report the results of our integrated approach relative to the ex post optimal decision, i.e., relative to the optimal decision of the deterministic problem after demand has been realized. To put these results into perspective, we compare them to a traditional approach that does not use additional feature data. As discussed in 4.1, we use exponential smoothing in the estimation step and the MILP in the optimization step. We solve the MILP to near optimality (optimality gap $\leq 0.01\%$) with Gurobi 8.1.

Fig. 6 depicts the average profit of the integrated and the separated approaches, relative to the ex post maximum profit. The ex post optimal decision is obtained from the deterministic variant of the MILP (4) - (10) by setting $T = 1$ and $\hat{d}_{i,1} = d_i$, where d_i is the realized demand of product i.

From Fig. 6, we observe that the considered SEO approach, SEO-ETS, achieves 95% and 88% of the ex post optimal profit for buns and breads, respectively. Our IEO method outperforms the SEO benchmark by 2 percentage points for buns and matches it for breads. We note that the exponential smoothing benchmark already performs rather well. The reasons are twofold. First, the ETS method that constitutes the demand estimation model of the SEO approach achieves a very high forecast accuracy; i.e., it captures the mean demand quite well, especially in the category breads. This confirms earlier results (Huber et al., 2019) and is discussed in more detail in Sections 5.2.1 and 5.2.2. The main reason is that calendar information, especially the day of the week, is a very important feature in our setting due to a prominent weekly seasonal demand pattern. This information is also available to the exponential smoothing method. The second component of a successful approach is an accurate estimation of the forecast error distribution, i.e., the distribution around the mean demand. We know from earlier experiments on this dataset that the error distribution can be approximated quite well by a normal distribution with constant variance (Huber et al., 2019),

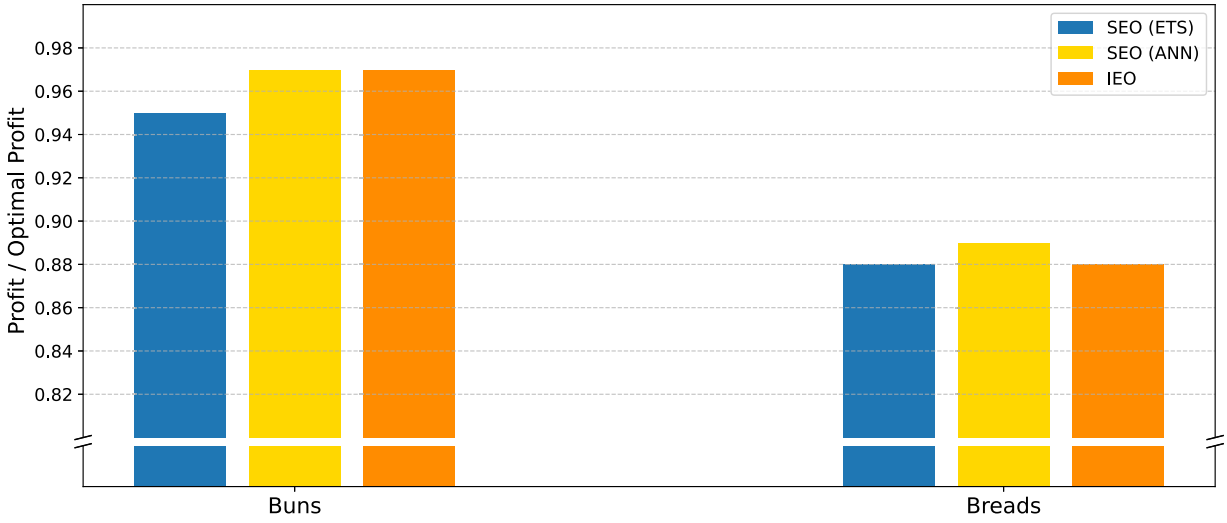


Fig. 6. Average profit relative to the ex post maximum profit.

which coincides with the assumption of the ETS method. We verify these arguments in Section 6.

In Fig. 7, we report the corresponding average order quantities of our IEO approach and the SEO benchmark as well as the ex post optimal decisions. For buns and breads, it is optimal not to order product 1 and product 4, respectively. Both products have the lowest profit margins in their respective categories (see Tables 3 and 4). Additionally, the substitution rates from these products to other products are relatively high. Therefore, the model suggests to "force" customers to substitute for higher-margin products by not ordering the lower-margin products at all. This result should be interpreted purely as an outcome of in-store substitution behavior, not as a recommendation to influence assortment design or customer store choice. We observe that both the separated and integrated approaches follow this logic. To account for longer term service level requirements, we analyze the effect of fill-rate constraints in Section 5.2.4.

Naturally, the order quantities of both the SEO and the IEO approaches are higher, in general, than the ex post optimal order quantity, as they include safety stock to buffer against uncertainty.

To summarize, the integrated method slightly outperforms the separated benchmark on the first dataset, while it is competitive on the second one. In what follows, we provide more insights into the effects that influence the decisions and performance of our approach.

5.2.1. *The effect of demand estimation accuracy*

A plausible reason for the superior performance of our integrated approach is that it is able to leverage more data than the separated exponential smoothing method. Therefore, we employ a separated approach, using an ANN rather than ETS to estimate the demand model. Thus, this model processes the same data, including the features, as our integrated approach.

To measure the effect of the demand estimation accuracy, we compute a set of common accuracy measures, including 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), for both the ETS and ANN forecasts. Table 8 shows the average forecast accuracy over all time series for each method.

We observe that the ANN forecast outperforms the ETS forecast in all accuracy measures across both product categories. The difference is relatively small for breads but relatively large for buns.

We can observe that higher forecast accuracy leads to higher profits. As reported in Fig. 6, using the ANN as the forecast method in the separated approach yields a profit increase of 2% (1%) for buns (breads), relative to the *ETS* forecast. We note that ANN-SEO and IEO perform similarly for buns and that ANN-SEO outperforms IEO by 1% for breads. To better understand the observed relative performances of both approaches, we further compare them for varying settings in Section 6.

Table 8
Forecast performance of the point predictions.

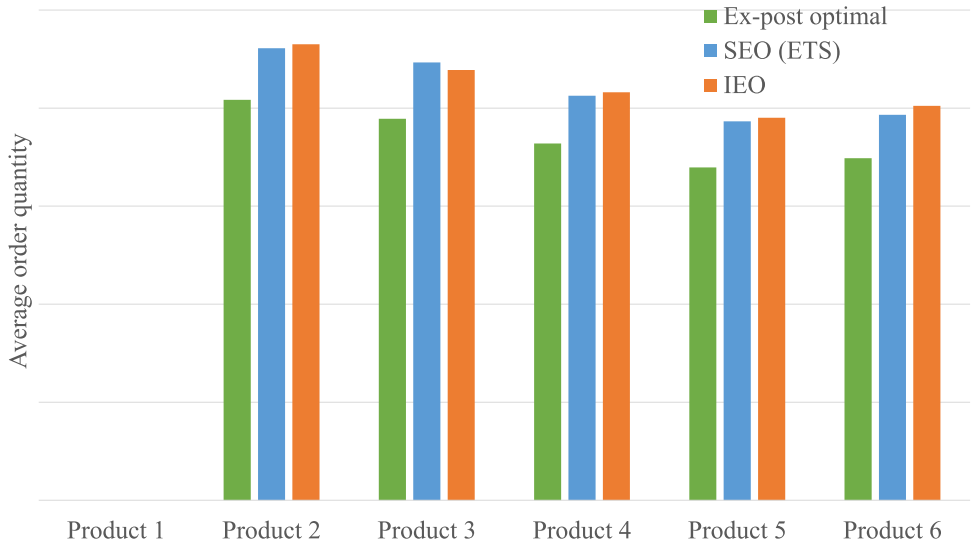
Category	Method	RMSE	MAE	MAPE	SMAPE	MASE
Buns	ETS	75.15	28.50	20.39	18.91	0.84
	ANN	44.13	22.00	16.75	15.59	0.67
Breads	ETS	4.77	3.33	36.47	28.04	0.83
	ANN	4.70	3.25	35.34	27.61	0.82

5.2.2. *The effect of feature data*

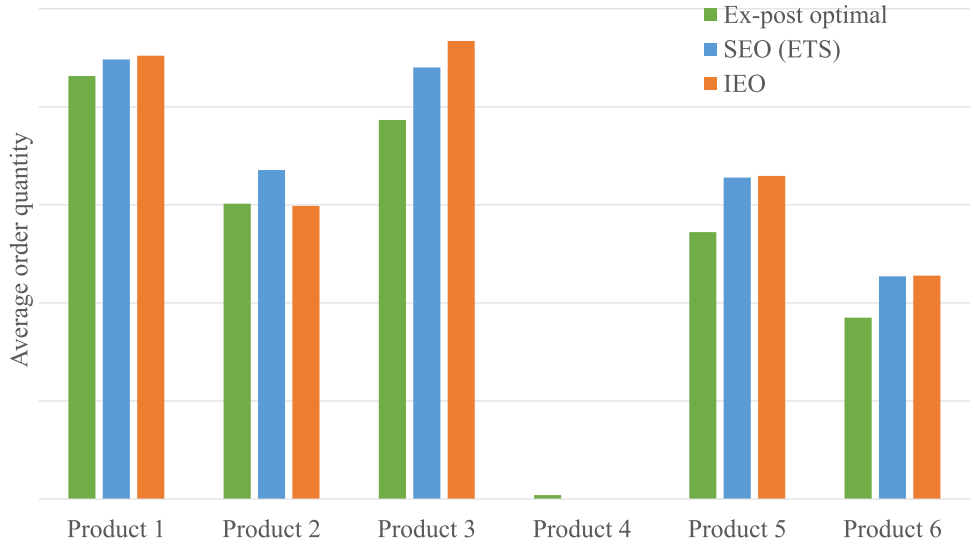
From a practical perspective, we investigate the importance of the individual groups of feature data that our integrated method exploits. We use permutation feature importance (PFI), which was originally introduced for random forest prediction models by Breiman (2001). We adapt the methodology to measure the decrease in profit after feature permutation. To this end, we randomly shuffle the rows of a single set of features (see Table 2) at a time, thereby suppressing its correlation with demand. After permutation, we predict order quantities and compute the corresponding profit. For each set of features, we repeat this process 1000 times and report the average profit loss relative to the unpermuted profit. The permutation of an important feature with high prediction value will result in a high profit loss. The permutation of an unimportant feature will result in a low profit loss if any. The results are depicted in Fig. 8.

We observe that for both categories, calendar data, master data, and transaction data are the most important feature sources, although to varying extents. Surprisingly, weather data have almost no value for optimizing inventory decisions in our datasets, even though they are used in many models for food retail (e.g., Beutel and Minner (2012), Kök and Fisher (2007)). For the category buns, excluding calendar data results in a substantial profit loss of 15.08%, while the loss amounts to only 3.61% for the category breads. Similarly, transaction data have different importances in both datasets. The impact of location data is relatively small (1.37 and 2.04%) in both datasets.

From the feature analysis, we conclude that calendar data, master data, and transaction data are important features in our dataset. Although breads and buns are relatively similar (fresh bakery products),



(a) Category: buns.



(b) Category: breads.

Fig. 7. Average order quantities per product and product category. The scale is not provided for reasons of confidentiality.

the extent of profit loss through PFI varies greatly. Therefore, retailers should carefully select features for inventory optimization.

5.2.3. *The effect of substitution*

Next, we investigate the effect of substitution on the decisions and resulting average profits. To this end, we compare the IEO and SEO approaches to their single-product newsvendor counterparts that ignore substitution effects. We report the percentage of lost profit from ignoring substitution in Fig. 9.

We observe that ignoring substitution leads to a profit loss of approximately 14.5% for both SEO and IEO in the category buns. For the category breads the corresponding loss is much lower, amounting to approximately 3.4% for both SEO and IEO. The importance of substitution effects in the category buns is mainly due to the specific characteristics of product 1. Product 1 has by far the largest share of demand within the category (64.3%), the lowest profit margin, and the highest rates of substitution for other products (69%). Within the model, these characteristics make it optimal to allocate little or no inventory to product 1, since many customers substitute toward higher-margin alternatives. We see a similar but smaller effect with product 4 in the category breads. It is also the lowest-margin product within the category, but it is not as dominant in terms of demand share (23.8%).

Overall, we find that accounting for substitution is a very important aspect that drives the performance of the methods considered. The large demand shifts away from product 1 in the category buns and from product 4 in the category breads result in significant profit gains. At the same time, we emphasize that such model outcomes should not be interpreted as direct recommendations to eliminate popular products. In practice, reducing the availability of products with a large demand share may negatively affect customer satisfaction and store traffic. We return to this issue in the next section and propose the use of fill-rate constraints to reconcile these objectives.

5.2.4. *The effect of fill-rate constraints*

In Fig. 7, we observed large demand shifts from low-margin products to higher-margin substitutes which are enforced by low, or even

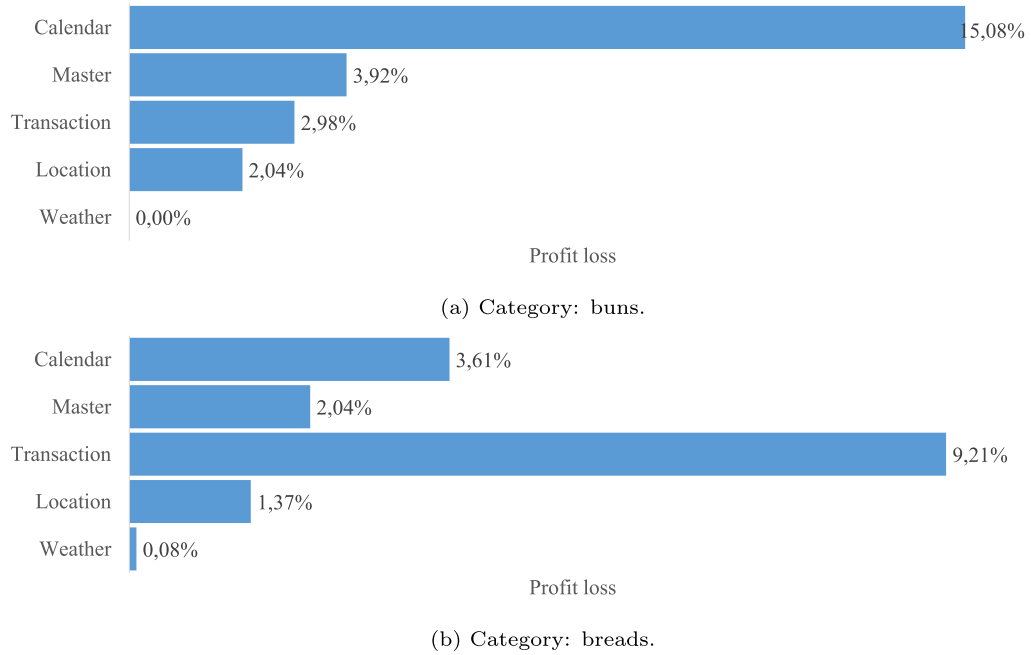


Fig. 8. Profit loss through permutation feature importance (PFI).

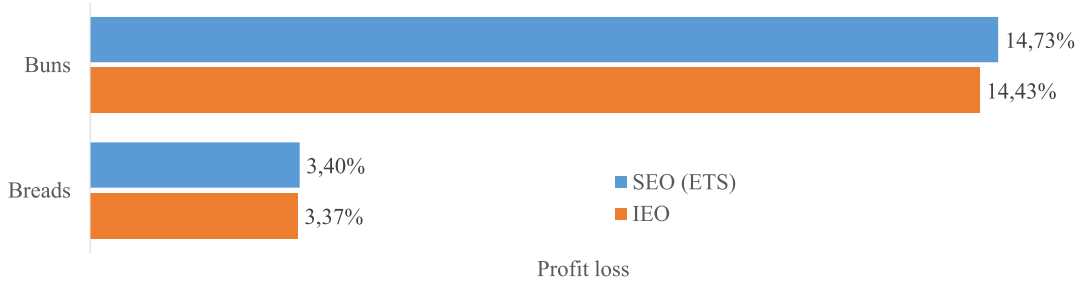


Fig. 9. Profit loss by ignoring substitution.

vanishing, ordering quantities for some products. We see similar effects in other applications of multi-product inventory models to real-world problems. Kök and Fisher (2007) find that "[p]roducts with low profit are dropped from the assortment, the number of facings of products with low marginal return are reduced, and the number of facings of those with higher returns are increased." They state that their recommendation suggests a more than 50% increase in profits compared to the current state. However, these decisions optimize the short-term profit. The long-term effects of stock-outs,e.g., dissatisfied customers and future demand losses, are not captured in these models, although they may drive customers' future store choices (Briesch et al., 2009). Including these long-term effects in the underage costs u_i of each product i is difficult, since long-term shortage costs are notoriously difficult to estimate. Therefore, we instead propose adding fill-rate constraints to the MILP (4) to (10) that account for strategic service level requirements.

We set a target fill-rate per category as

$$\beta^{cat} \leq \frac{\sum_t \sum_i \left(q_i - y_{i,t}\right)}{\sum_t \sum_i \hat{d}_{i,t}}. \tag{13}$$

Alternatively, we introduce target fill-rate constraints per product as

$$\beta_i^{prod} \leq 1 - \frac{\sum_t v_{i,t}}{\sum_t \hat{d}_{i,t}} \quad \forall i. \tag{14}$$

We vary the minimum category fill rate between 0.80 and 0.99 and the minimum product fill rate between 0.80 and 0.95, and we optimize the MILP based on the ETS and ANN distribution forecast. The resulting profits and fill rates are shown in Table 9 and in Fig. 10.

Across all methods and categories, we can observe a trade-off between short-term profit and service, i.e., fill rate, which is illustrated in Fig. 10. As soon as one of the fill-rate constraints is binding, the profit decreases. In the category buns, a product fill rate of 0.80 already reduces the profit by 11 percentage points, as the demand shifts away from product 1 to more profitable products is suppressed. In the category breads, this effect is smaller, as the products are more homogeneous with respect to volume and margin. Low category fill-rate constraints do not harm the profit too much, as they still allow for substitution.

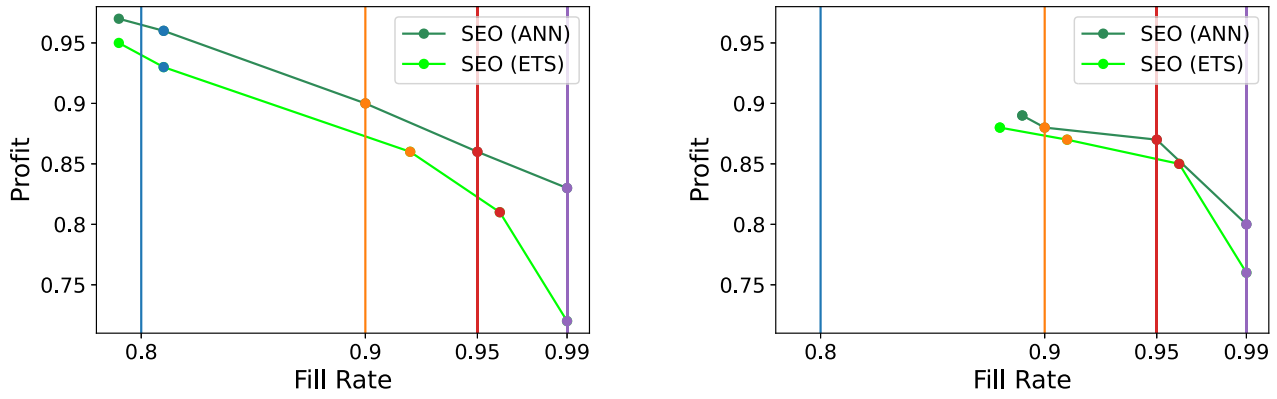
If the long-term objectives of the retailer,e.g., attracting customers by means of a large assortment, are in conflict with short-term profits, fill-rate constraints can help to reach a desirable trade-off.

6. Assessing the benefits of integrated estimation and optimization

In the previous section, we found that a well-specified SEO approach can match or even slightly outperform IEO on real-world data. To better understand these results, we now investigate the conditions under which one method has an advantage over the other. From a theoretical perspective, IEO has two main benefits: it avoids repeatedly solving the optimization problem (4)-(10) in each period, and it can adapt to feature-dependent changes in demand distributions, whereas SEO

Table 9
Average profit relative to ex post maximum profit and resulting fill-rates under fill-rate constraints.

	Method	Target Fill Rate		Profit	Fill Rate						
		per Cat.	per Prod.		Prod. 1	Prod. 2	Prod. 3	Prod. 4	Prod. 5	Prod. 6	Cat.
Category Buns	SEO (ANN)	–	–	0.97	0.00	1.00	1.00	1.00	1.00	1.00	0.79
		–	0.80	0.86	0.82	1.00	1.00	1.00	1.00	1.00	0.96
		–	0.90	0.84	0.92	1.00	1.00	1.00	1.00	1.00	0.98
		–	0.95	0.84	0.96	0.99	1.00	0.99	1.00	0.99	0.99
		0.80	–	0.96	0.05	1.00	1.00	1.00	1.00	1.00	0.81
		0.90	–	0.90	0.54	1.00	1.00	1.00	1.00	1.00	0.90
		0.95	–	0.86	0.80	1.00	1.00	1.00	1.00	1.00	0.95
		0.99	–	0.83	0.98	1.00	1.00	1.00	1.00	1.00	0.99
	SEO (ETS)	–	–	0.95	0.00	0.99	1.00	1.00	1.00	1.00	0.79
		–	0.80	0.84	0.83	0.99	0.99	0.99	0.99	0.99	0.95
		–	0.90	0.82	0.93	0.99	0.99	0.99	0.99	0.99	0.97
		–	0.95	0.80	0.96	0.99	0.99	0.99	0.99	0.99	0.98
		0.80	–	0.93	0.13	1.00	1.00	1.00	1.00	1.00	0.81
		0.90	–	0.86	0.64	0.99	0.99	0.99	1.00	0.99	0.92
		0.95	–	0.81	0.87	0.99	0.99	0.99	0.99	0.99	0.96
		0.99	–	0.72	0.98	0.99	0.99	0.99	0.99	0.99	0.99
Category Breads	SEO (ANN)	–	–	0.89	0.97	0.96	0.99	0.00	0.99	0.98	0.89
		–	0.80	0.87	0.91	0.93	0.98	0.85	0.98	0.96	0.94
		–	0.90	0.86	0.93	0.93	0.97	0.93	0.97	0.97	0.96
		–	0.95	0.82	0.97	0.95	0.97	0.97	0.98	0.98	0.98
		0.80	–	0.89	0.97	0.96	0.99	0.00	0.99	0.98	0.89
		0.90	–	0.88	0.98	0.96	0.99	0.12	0.99	0.98	0.90
		0.95	–	0.87	0.93	0.94	0.98	0.83	0.98	0.96	0.95
		0.99	–	0.80	0.98	0.96	0.99	0.99	0.98	0.98	0.99
	SEO (ETS)	–	–	0.88	0.97	0.97	0.98	0.00	0.99	0.98	0.88
		–	0.80	0.86	0.90	0.95	0.96	0.84	0.98	0.97	0.94
		–	0.90	0.85	0.93	0.95	0.96	0.93	0.97	0.97	0.96
		–	0.95	0.81	0.97	0.96	0.97	0.96	0.98	0.98	0.98
		0.80	–	0.88	0.97	0.97	0.98	0.00	0.99	0.98	0.88
		0.90	–	0.87	0.97	0.97	0.98	0.23	0.99	0.98	0.91
		0.95	–	0.85	0.94	0.95	0.97	0.87	0.98	0.97	0.96
		0.99	–	0.76	0.99	0.97	0.99	0.99	0.99	0.98	0.99



(a) Category: buns. (b) Category: breads.

Fig. 10. Average profit relative to ex post maximum profit of model-based approaches with fill-rate constraints at the category level. The circular markers report the actual profit and fill rate achieved by an approach (cf. Table 9). The vertical grid lines are visual aids that mark the required minimum fill rates. Markers and grid lines are color-coded to indicate which constraint a result marker belongs to.

assumes stationarity. In this section, we use controlled simulations to isolate and study these mechanisms. To focus on the conjectured differences between the two approaches, we abstract from preprocessing, which is required in both methods, and assume perfect information by simulating demand and substitution rates directly instead of estimating these quantities from sales data.

6.1. Experimental setting

To observe substitution but not overcomplicate our study, we fix the number of products at $n_{prods} = 3$. We consider seven mean service levels[2] $\mu_{SL} \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$. For a given μ_{SL}, we set the underage and overage costs of the three products to $u_{\mu_{SL}} = (u_1, u_2, u_3) = (.99\mu_{SL},\ \mu_{SL},\ 1.01\mu_{SL})$, and $o_{\mu_{SL}} = (o_1, o_2, o_3) = 1 - u_{\mu_{SL}}$, respectively. To test the theoretical benefits of IEO, we construct a setting in which the demand distribution is feature dependent. To this end, we follow the study conducted in Oroojlooyjadid et al. (2020) and simulate demand as a function of four categorical variables $X_i \in \mathcal{K}_i \subseteq \mathbb{N}_0, i = 1, \dots, 4$. The number of combinations of the features $n_{clust} = \prod_{i=1}^{4} |\mathcal{K}_i|$ defines the number of potentially different demand distributions, which we denote as clusters. We construct these distributions as follows. For each cluster $k = 1, \dots, n_{clust}$, we specify a tuple (μ_k, D_k), consisting of a mean demand vector $\mu_k = (\mu_{k,1}, \dots, \mu_{k,n_{prods}})$ and a mean-zero error distribution D_k. Finally, we generate a substitution matrix $\alpha = (\alpha_{i,j})_{i,j=1,\dots,n_{prods}}$ as follows. We consider three levels of substitution: *no*, *moderate* and *strong* substitution, reflected by the following matrices:

$$\alpha_{no} = \begin{bmatrix} 0. & 0. & 0. \\ 0. & 0. & 0. \\ 0. & 0. & 0. \end{bmatrix}, \alpha_{mod} = \begin{bmatrix} 0. & 0.238 & 0.201 \\ 0.182 & 0. & 0.215 \\ 0.146 & 0.297 & 0. \end{bmatrix},$$
$$\alpha_{strong} = \begin{bmatrix} 0. & 0.343 & 0.652 \\ 0.416 & 0. & 0.507 \\ 0.603 & 0.365 & 0. \end{bmatrix}. \tag{15}$$

Without substitution, the order decisions for the three products are mutually independent, and the problem is consistent with the classical single-product newsvendor problem (NVP). The moderate substitution rates are sampled uniformly from $[0, 1/3)$ under the conditions that

(i) $\alpha_{ii} = 0$ for all $i = 1, \dots, n_{prods}$,
(ii) $\sum_{j=1}^{n_{prods}} \alpha_{i,j} \leq 1$, and
(iii1) $\sum_j \alpha_{i,j} u_j < u_i$ for all $i = 1, \dots, n_{prods}$. Condition (iii1) guarantees that there is no substitution in the ex-post solution, i.e., no product is strategically removed from the assortment. Analogously, the strong substitution rates are sampled uniformly from $[1/3, 1)$ under conditions (i), (ii) and
(iii2) $\sum_j \alpha_{i,j} u_j < u_i$ for all $i = 2, \dots, n_{prods}$ and $\sum_j \alpha_{1,j} u_j > u_1$,

which guarantees that in the ex-post solution, product 1 is not offered.

For any given set of clusters, we simulate $N = 12,000$ feature-demand pairs $(x_i, d_i)_{i=,\dots,N}$, where d_i is simulated according to the cluster parameters (μ_k, D_k) of the cluster identified by x_i. We split the data into $8,000$ training records, $2,000$ validation records and 10 test sets of 200 records each. The data are perfectly balanced; i.e., every feature combination occurs equally often.

In this environment, we analyze the performance of four models:

- Ex-Post: As explained in Section 5.2, the ex post solution is the solution of the deterministic variant of (4)-(10).
- IEO: Our proposed approach, detailed in Section 4.2.
- SEO: The separated benchmark approach with an ANN in the demand forecasting step as detailed in Section 4.1.
- Ex-Ante: As we generate the data synthetically, we have perfect knowledge of the demand distributions by cluster. This allows us to solve for the ex ante optimal order quantities by minimizing the objective (1) for the given distribution. This ex ante solution mimics the performance of an optimal IEO approach that has perfectly learned the feature-demand relation. Thus, the ex-ante profit serves as an upper bound for IEO performance.

6.2. Motivating example

As a motivating example and to make the general results of Section 6.3 more accessible, we borrow the two-population model introduced in Ban and Rudin (2019). The popular framework is motivated by assuming two distinct customer groups, such as men and women, or shoppers on weekdays vs. weekends. It can be formalized by modeling the demand D via a single binary feature $X \in \{0, 1\}$, the population means μ_1, μ_2 and the corresponding mean-zero error distributions D_1 and D_2:

$$D = (1 - X)(\mu_1 + D_1) + X(\mu_2 + D_2).$$

[2] The term "mean service level" is inspired by the single-product case. The cost parameters have been selected such that the average target service level over all three products corresponds to that cost scenario's μ_{SL}. In the multi-product case, this does no longer hold and the term "mean service level" is mostly to be understood as a convenient shorthand for the corresponding cost scenario.

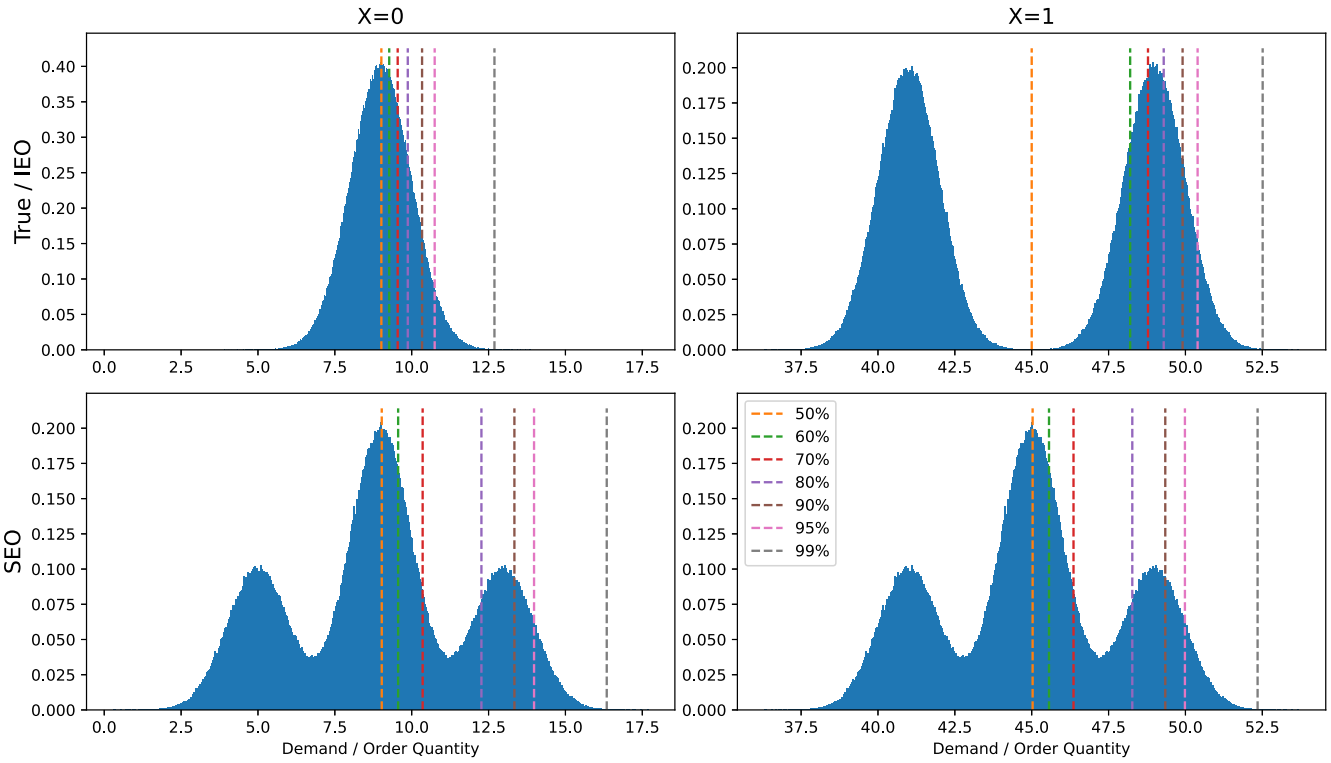


Fig. 11. The single-product NVP in a two-population environment. Top row: True demand, as recognized by an integrated approach. Bottom row: Misspecified demand of a separated approach. The vertical lines indicate different quantiles of the distributions, which for the single-product NVP coincide with the optimal order quantities for the corresponding service levels.

6.2.1. The two-population model in a single-product setting

We start with a brief look at a single-product setting, i.e., a classical NVP. A simple example illustrates how IEO recognizes the feature dependence of the demand distribution while the separated approach is oblivious to it. Fig. 11 visualizes the differences between both approaches where $D_1 \sim \mathcal{N}(0, \sigma_1^2)$ is a standard normal distribution, while $D_2 \sim \mathcal{GM}(\mu, \sigma_2^2) = (1 - B)\,\mathcal{N}(-\mu, \sigma_2^2) + B\,\mathcal{N}(\mu, \sigma_2^2)$ is a Gaussian mixture, denoted $\mathcal{GM}$, where B denotes a $(0, 1)$-Bernoulli variable. While the SEO approach captures the feature dependence of the *mean* demand correctly, it creates a distribution around the estimated means based on all historical prediction errors, regardless of the population to which they belong. The resulting demand *distribution* is independent of the feature value and wrong for both populations, resulting in suboptimal order quantities. The magnitude of the deviation and its impact on the objective function,i.e., cost or profit, depends on the two distributions and the service level.

This analysis is straightforward because the solution to the single-product NVP is available in closed form. This is not the case for the multi-product problem under substitution. In the remainder of Section 6, we demonstrate how the identified benefit of the IEO approach carries over to this problem.

6.2.2. The two-population model under substitution

In this section, we present the results of the two-population model in a multi-product setting under substitution. Formally, we set $\mathcal{K}_1 = \{0, 1\}$, and $\mathcal{K}_i = \{0\}, i = 2, 3, 4$, resulting in $n_{clust} = 2$. We further set $(\mu_1, D_1) = ([23, 22, 21],\ \mathcal{N}(0, 1))$ and $(\mu_2, D_2) = ([22, 21, 23],\ \mathcal{GM}(15, 2))$.

Findings on profits. Fig. 12 displays the results for a moderate level of substitution α_{mod} and varying mean service level μ_{SL}. The results for other levels of substitution are structurally the same. We can observe that the IEO approach always achieves the ex-ante optimal profit, except for a mean service level of 50% where there's a negligible gap. While the SEO approach performs well for small and large service levels, the gap to the ex-ante optimal profit is noticeable for intermediate levels, attaining its maximum of eight percentage points for service levels of 70 and 80%. The ex-ante attainable profit is closer to the ex-post optimal one the larger the service level.

Fig. 12 supports three conclusions. First, the integrated approach perfectly recognized the feature-demand-dependency and operates on the same level as the ex ante solution. We conclude that an ANN can, in principle, be trained on the multi-product newsvendor profit function (11). Second, the IEO approach outperforms its SEO counterpart across all service levels and substitution rates. We conclude that the theoretical superiority of the integrated approach in the classical NVP is transferable to the case under substitution. Third, the IEO-SEO gap depends on the service level. In particular, with increasing service levels, the gap first rises, peaks for moderate service levels and then decreases again. There are two reasons for this result. In our two-population model, both error distributions not only have zero mean but are symmetric. As a consequence, the 50% quantile coincides with the mean of the distribution, and the SEO approach makes very small mistakes when estimating those quantiles. Therefore, for small service levels, the profit gap is small. Larger service levels require the estimation of quantiles at the upper tail of the distribution, so the deviations of the SEO decisions from the ex ante optimal order quantities increase with the service level. This is depicted in Fig. 13a. Initially, this also inflates the profit gap. However, an increasing service level also means a reduction of the overage cost in proportion to the underage cost, so the profit is less sensitive to excess orders than to shortages. Since the SEO approach creates its distribution by considering all errors, it will always assume a distribution with a support that contains the support of the true distribution (cf. Fig. 11). As a consequence, the SEO error becomes increasingly dominated by overages, which for a high service level are cheaper than underages, and the profit gap starts to decrease again.

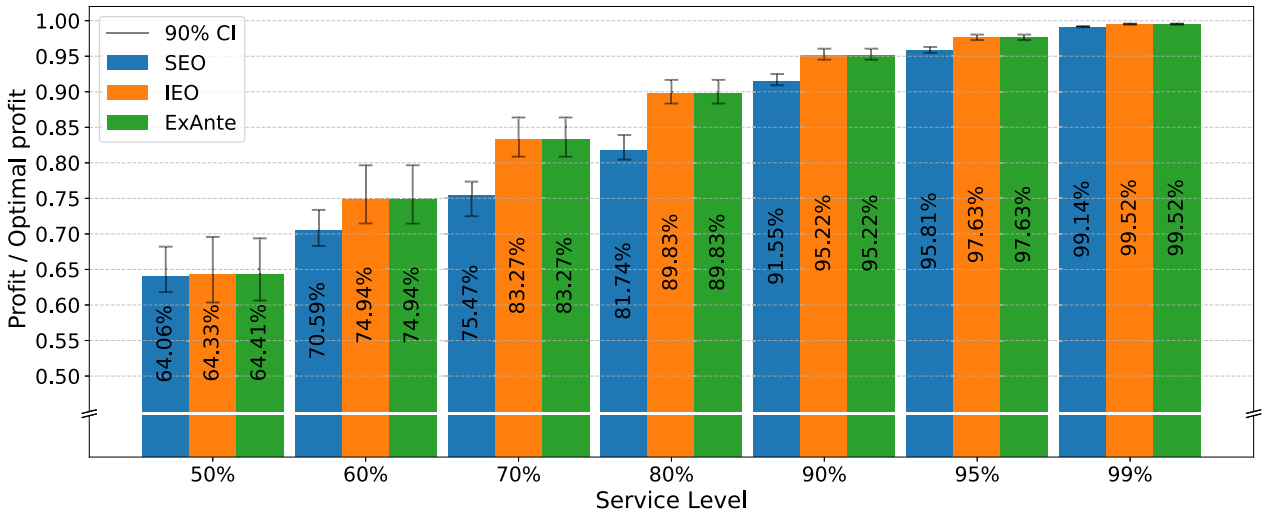
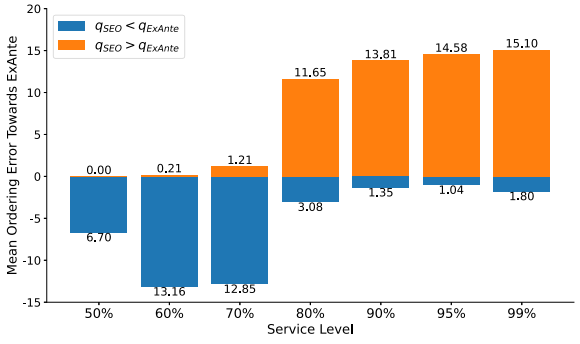
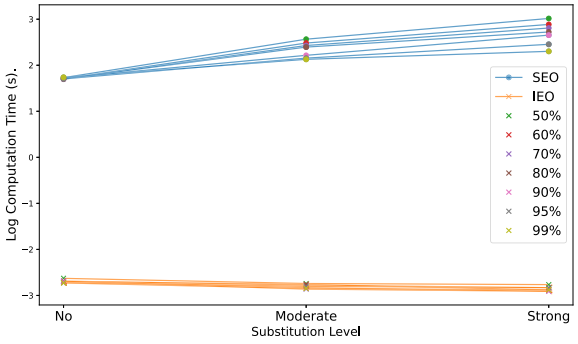


Fig. 12. Results under moderate substitution in the two-population model. The bars show the ratio of profit to ex post profit.



(a) Deviations of the SEO approach from the ex ante optimal order quantities under moderate substitution in the two-population model. Both the total deviation and the share of excess orders increase with the service level.



(b) Mean evaluation time (log scale, in seconds) as a function of the level of substitution. Each line connects the markers belonging to the same service level.

Fig. 13. Average profit relative to ex post maximum profit of model-based approaches with fill-rate constraints at the category level.

Findings on computation times. The second fundamental benefit of an integrated method is the reduced computational effort at test time. This computational advantage is positively correlated with the level of substitution. Indeed, without substitution, the ordering decisions for the different products are not interrelated and the linear program (4)-(10) dissolves into n_{prods} classical NVPs. When we increase the level of substitution, the interdependence between the products and therefore the solution space becomes more complex, which increases the computation time required to solve the problem. The results supporting these hypotheses are shown in Fig. 13b.

Fig. 13b shows the average log-computation time in seconds for 200 test periods. The average is taken over the 10 test datasets. The IEO approach evaluates in real-time, regardless of the level of substitution, with an average evaluation time of 0.002 seconds per 200 periods. The evaluation time of the SEO approach increases with the level of substitution, from an average of 50 seconds for the classical NVP to approximately 800 seconds under strong substitution and moderate service levels. We note that the latter setting – moderate service levels and strong substitution – is common in fast-moving consumer goods (cf. Sections 2 and Oroojlooyjadid et al. (2020)). In addition, by examining the MILP formulation (4)-(10), it is clear that the complexity increases further with the number of products, while the evaluation time of the IEO approach is little affected. With only three products, our study environment can be considered a toy example along that dimension. This underlines the second theoretical advantage: The IEO model *learns* the relation between input features and the solution of the optimization problem, while the SEO approach needs to *re-solve* the problem in each period.

6.3. Multi-population model

The two-population setting analyzed in the preceding subsections is simplistic from a practical perspective. Therefore, we generalize our observations in this subsection by increasing the number of populations and considering various different error distributions. An important intuition gained from the two-population model is that the SEO error should increase with a stronger feature-dependence of the error distribution. In this section, we quantify this demand-feature-dependency (in the following referred to as DFD) and its effect on the IEO-SEO profit gap.

To formalize our observations, we require a metric for the strength of the feature-dependence of the demand distribution. We note that in both the standard regression and the newsvendor literature, the impact that features have on a target variable's distribution is largely limited to changes in the variance, a phenomenon referred to as heteroscedasticity (Harsha et al., 2021; Huber et al., 2019). In our study, we take a more general perspective and allow the features to impact the distribution family itself,cf. Fig. 11. This framework can model standard heteroscedasticity, but also covers scenarios in which the distribution family changes while the variance remains constant. In addition, the

solution of the NVP,with or without substitution, depends on the distribution in a more subtle way than on its variance alone. As a consequence, to quantify feature-dependence of the demand distribution, we need a measure that examines more than only the first or second moments. Because of its statistical consistency and ease of implementation, we choose the maximum mean discrepancy (MMD, Gretton et al. (2012)).

The MMD quantifies the distance between two distributions by computing the distance of their means when embedded in a – usually higher-dimensional – feature space. More specifically, for two distributions P, Q on a space $\mathcal{X}$ and a feature map $\varphi : \mathcal{X} \to \mathcal{H}$ mapping $\mathcal{X}$ to a reproducing Hilbert space (RKHS) $\mathcal{H}$, the maximum mean discrepancy is given by

$$\mathrm{MMD}^2(P, Q) = \|\mathbf{E}_{X\sim P}[\varphi(X)] - \mathbf{E}_{Y\sim Q}[\varphi(Y)]\|^2_{\mathcal{H}}.$$

Using the so-called kernel trick (Steinwart & Christmann, 2008), we can exploit complex feature maps while avoiding calculations in high-dimensional spaces. That is, letting $k(x, y) = \langle\varphi(x), \varphi(y)\rangle_{\mathcal{H}}$ denote the kernel that is uniquely identified with the feature map φ, the MMD reduces to

$$\mathrm{MMD}^2(P, Q) = \mathbf{E}_{X,X'\sim P}[k(X, X')] + \mathbf{E}_{Y,Y'\sim Q}[k(Y, Y')] - 2\mathbf{E}_{X\sim P,Y\sim Q}[k(X, Y)].$$

This is readily estimated given samples of P and Q by replacing the expectations with empirical means. However, this naive estimation approach yields an estimator that is biased and whose computing time scales quadratically in the number of samples. Instead, for samples $S_P = (x_1, \ldots, x_n)$ from P and $S_Q = (y_1, \ldots, y_n)$ from Q, we set $m = \lfloor n/2 \rfloor$ and compute

$$MMD_l^2(S_P, S_Q) = \frac{1}{m_2}\sum_{i=1}^{m_2}\big[k(x_{2i-1}, x_{2i}) + k(y_{2i-1}, y_{2i}) - k(x_{2i-1}, y_{2i}) - k(x_{2i}, y_{2i-1})\big]. \tag{16}$$

MMD_l scales linearly in the number of samples and is an unbiased estimator of the true MMD (Gretton et al., 2012). Simplistic feature maps and their kernels,such as the identity $\varphi : \mathbb{R} \to \mathbb{R}, x \mapsto x$, lead to a naive distance measure that evaluates to 0 even if the distributions are not identical. More powerful kernels such as the Gaussian kernel

$$k_\sigma(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-y)^2}{2\sigma^2}}$$

guarantee that the MMD is 0 if and only if P is equal to Q, ensuring that in fact the MMD is a metric in the mathematical sense (Gretton et al., 2012; Steinwart & Christmann, 2008). As a consequence, we choose the MMD with a Gaussian kernel and a bandwidth parameter of $\sigma = 5$. For a more detailed treatment of kernels, feature maps and reproducing kernel Hilbert spaces, we refer to Steinwart and Christmann (2008).

Inspired by the Goldfeld-Quandt test for heteroscedasticity (Goldfeld & Quandt, 1965), we use the MMD to quantify the DFD of a dataset in the following way:

$$\mathrm{DFD}(D_1, \ldots, D_{n_{clust}}) = \frac{1}{n_{clust}}\sum_{i=1}^{n_{clust}} \mathrm{MMD}(D_i, D_{\mathrm{SEO}}),$$

where D_{SEO} denotes the error distribution as estimated by the SEO approach. For instance, returning to our initial example (cf. Fig. 11), we have

$$\begin{aligned}\mathrm{DFD}(D_1, D_2) &= \frac{1}{2}\big[\mathrm{MMD}(\mathcal{N}(0, 1), D_{\mathrm{SEO}}) + \mathrm{MMD}(\mathcal{GM}(4, 1), D_{\mathrm{SEO}}\big] \\ &\approx \frac{1}{2}(0.0390 + 0.0401) = 0.0395.\end{aligned}$$

Our goal is to quantify the relationship between the DFD and the IEO-SEO profit gap. To this end, we consider seven error distribution families: Normal, Gaussian mix, Logistic, Beta with shape parameters $a = 1.5, b = 1$, Beta with shape parameters $a = b = 2$, and Gamma with shape parameter $a = 2$, each centered at 0 and scaled to five different levels of variance (1, 5, 10, 15, 20). In total, this yields 35 distinct error distributions. We ran the experiment according to the setting outlined in Section 6.1 50 times. For each run, the number of clusters n_{clust} was drawn uniformly between 2 and 35, and the n_{clust} error distributions were drawn randomly from the 35 distributions detailed above without replacement. We note that in this section, we are not interested in the performance of a specific IEO or SEO variant. We analyze the benefit of IEO over SEO under the assumption that both approaches have been trained perfectly. Therefore, we refrain from training IEO and SEO approaches for every run but rather report the Ex-Ante versus SEO profit gap. In Section 6.2.2 we saw that, in this setting, the IEO approach can be trained, in principle, and, given sufficient training data, achieves a performance indistinguishable from that of the ex ante optimal solution. Also, we assume an SEO approach that perfectly captures a cluster's mean demand and only constructs erroneous error distributions by lumping together samples of all clusters' error distributions (see Fig. 11). In this sense, the reported Ex-Ante-SEO profit gap mirrors the IEO-SEO gap.

Fig. 14 displays the results of our extensive computational study. For six different mean service levels, we plot the Ex-Ante-SEO profit gap on the y-axis versus the DFD on the x-axis. The three different substitution levels are represented by three different colors. Each marker represents the results of one experiment run as described above. The generalized linear regression lines are fitted using `sklearn`'s `TweedieRegressor` function.

Fig. 14 confirms the conjectured positive correlation between the demand-feature-dependency and the profit gap of the SEO approach with respect to the ex ante optimal solution. The specifics of the correlations depend on the service level and, for large service levels, on the substitution level as well to a small degree. Nonetheless, for all seven mean service levels and all three levels of substitution, we can confirm a strong positive correlation. Regardless of the level of substitution, the IEO-SEO profit gap scales exponentially in the DFD for moderate service levels, between 50% and 80%, and linearly for high service levels, above 90%.

We conclude that the integrated approach outperforms its separated counterpart when the error distribution depends on the features and is competitive otherwise. Our analysis also helps to explain the suboptimal performance of our IEO method in Section 5. Our real-world data possess a DFD of 0.1189 in the category buns and 0.0815 in the category breads, which is indeed very low, giving the integrated approach only a small advantage at best. Furthermore, Section 6.2.2 demonstrates that our integrated approach can indeed achieve ex ante optimal performance; i.e., it is perfectly trainable on the non-standard multi-product loss function (11). Failure to achieve such performance is most likely attributable to insufficient hyperparameter optimization and/or lack of data. These issues can explain the results of Section 5.2.1, where the ANN-SEO approach outperforms the integrated approach in the category bread by 1%.

We close the discussion of the DFD with a note on potentially more sophisticated separated approaches. The fundamental reason for the inferiority of the SEO method is its naive assumption that the target's error distribution is constant. One might be tempted to improve upon that naive approach by selecting historical errors more carefully. Assuming that the error distribution depends on the features, one would then try to include only those past errors that correspond with feature realizations that are in some sense "similar" to the current one. This has two downsides. First, it raises a number of new non-trivial questions regarding the error selection process: How many past errors should we select? Too few would result in an inaccurate and volatile approximation of the error distribution, while too many would water down the dependence between the created error distribution and the features, which defeats the purpose. Furthermore, what criterion should be used to select past errors? A simple similarity measure such as the Euclidean norm oversimplifies the problem, as it is unlikely that all feature variables are equally relevant in determining the error distribution. A more promising solution is a supervised clustering method, such as a decision tree

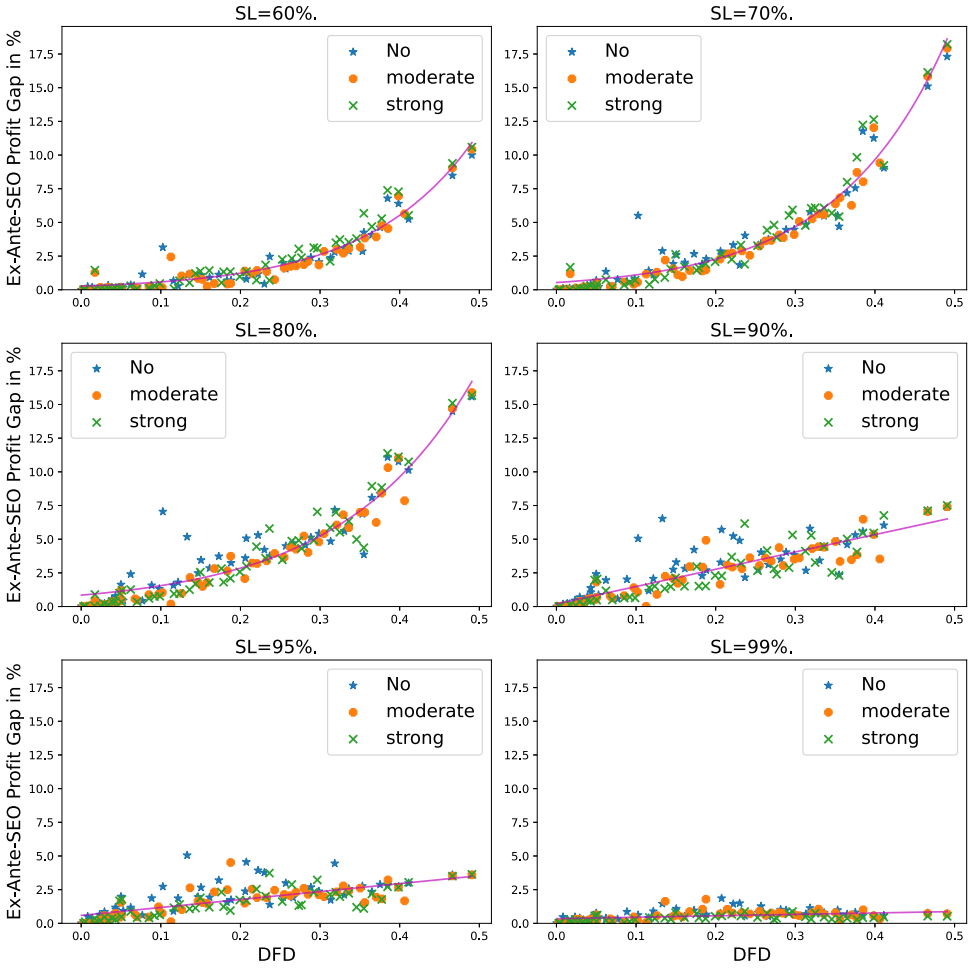


Fig. 14. Ex-Ante-SEO profit gap as a function of DFD for six service and three substitution levels.

(Bertsimas & Kallus, 2020). In essence, this approach introduces another set of methods and hyperparameters that needs careful tuning and complicates the problem. The IEO approach resolves those issues implicitly. Second, even if those questions are addressed satisfactorily, the SEO approach still needs to solve the optimization problem in each period, while the IEO method readily prescribes a decision for any combination of input features (cf. our discussion in Section 6.2.2).

7. Conclusions

In this paper, we study the multi-product newsvendor problem with unknown demand distributions. We introduce a solution approach that, after necessary preprocessing of sales data (e.g., demand decensoring and substitution estimation), integrates estimation and optimization within a single framework. We compare our method to a traditional separated approach in an empirical study based on real-world data from a bakery business. We also evaluate how performance depends on demand estimation accuracy, feature data, and fill-rate constraints. Subsequently, we perform a second study using systematically generated data to pinpoint what drives the performance gap between our approach and the traditional one.

Our key result is that our approach outperforms state-of-the-art SEO methods when the demand distribution depends on the feature data and is competitive otherwise. We introduce an MMD-based measure to quantify this relationship and observe that, depending on the cost factors, the benefit of our approach scales either exponentially or linearly as a function of this measure. In addition to the benefit of better performance, our approach prescribes a decision for any combination of feature data values once it is trained. There is no need to solve an optimization problem in each period; only infrequent updates of the parameters (for our real-world data, every two weeks) are needed. Similar to previous research (Kök & Fisher, 2007), we find in our empirical analysis that it is optimal for the maximization of short-term profits to delete a low-margin product from the assortment of each category. However, this might conflict with the long-term objectives, e.g., attracting customers through a large assortment, of the retailer. To reconcile these goals, we propose the use of fill-rate constraints.

Our empirical analyses also contribute to the literature on substitution estimation and the role of auxiliary data in inventory management. We estimate substitution rates from transaction data and find them to be consistent with prior survey-based studies on bakery products (van Woensel et al., 2007), demonstrating the viability of data-driven substitution estimation. Furthermore, we quantify the impact of different feature data sources, such as master data, transaction data, calendar effects, weather, and store location, on profits. While the relative importance varies by product category, calendar, master, and transaction data emerge as particularly valuable.

The interface between ML and OR is an active field of research that provides many opportunities for future work. Looking ahead, this study connects to the growing body of work on data-driven and end-to-end methods for operations (see, e.g., Elmachtoub and Grigas (2022), Elmachtoub et al. (2023), Mandi et al. (2024)), which emphasize integrating predictive and prescriptive elements. Promising avenues for future research include extending our method to multi-period settings with replenishments, developing approaches that are robust to inaccuracies in preprocessing steps (especially substitution rate estimation), and

experimenting with more advanced neural network architectures such as recurrent networks or transformers, which are increasingly applied in demand forecasting. Finally, applying our approach to other data-rich settings such as e-commerce could provide further insights into its practical benefits and limitations.

CRediT authorship contribution statement

Sebastian Müller: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization; **Jakob Huber:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation; **Ralph Alexander Bubak:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization; **Moritz Fleischmann:** Writing – review & editing, Supervision, Conceptualization; **Heiner Stuckenschmidt:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

None

References

Anupindi, R., Dada, M., & Gupta, S. (1998). Estimation of consumer demand with stock-out based substitution: An application to vending machine products. *Marketing Science*, *17*(4), 406–423. https://doi.org/10.1287/mksc.17.4.406

Ban, G. Y., & Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, *67*(1), 90–108. https://doi.org/10.1287/opre.2018.1757

Barrow, D., Crone, S. F., & Kourentzes, N. (2010). An evaluation of neural network ensembles and model selection for time series prediction. In *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.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Proceedings of the 24th international conference on neural information processing systems* (pp. 2546–2554). Curran Associates Inc.

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th international conference on machine learning - volume 28* (pp. 115–123). JMLR.org.

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044.

Besbes, O., Phillips, R., & Zeevi, A. (2010). Testing the validity of a demand model: An operations perspective. *Manufacturing and Service Operations Management*, *12*(1), 162–183.

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.

Briesch, R. A., Chintagunta, P. K., & Fox, E. J. (2009). How does assortment affect grocery store choice? *Journal of Marketing Research*, *46*(2), 176–189. https://doi.org/10.1509/jmkr.46.2.176

Campo, K., Gijsbrechts, E., & Nisol, P. (2000). Towards understanding consumer response to stock-outs. *Journal of Retailing*, *76*(2), 219–242. https://doi.org/10.1016/S0022-4359(00)00026-9

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.

Chen, N., Lagzi, S., & Milner, J. (2022). Using neural networks to guide data-driven operational decisions. Working paper. Available at SSRN: https://ssrn.com/abstract=4217092 or http://dx.doi.org/10.2139/ssrn.4217092.

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.

den Boer, A. V., & Sierag, D. D. (2021). Decision-based model selection. *European Journal of Operational Research*, *290*(2), 671–686. https://doi.org/10.1016/j.ejor.2020.08.025

Elmachtoub, A. N., & Grigas, P. (2022). Smart "predict, then optimize". *Management Science*, *68*(1), 9–26.

Elmachtoub, A. N., Lam, H., Zhang, H., & Zhao, Y. (2023). Estimate-then-optimize versus integrated-estimation-optimization versus sample average approximation: A stochastic dominance perspective. Working paper. Available at https://arxiv.org/abs/2304.06833.

Emmelhainz, L. W., Emmelhainz, M. A., & Stock, J. R. (1991). Logistics implications of retail stockouts. *Journal of Business Logistics*, *12*(2), 129–142.

Farahat, A., & Lee, J. (2018). The multiproduct newsvendor problem with customer choice. *Operations Research*, *66*(1), 123–136.

Fisher, M., & Vaidyanathan, R. (2014). A demand estimation procedure for retail assortment optimization with results from implementations. *Management Science*, *60*(10), 2401–2415. https://doi.org/10.1287/mnsc.2014.1904

Goldfeld, S. M., & Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, *60*(310), 539–547. http://www.jstor.org/stable/2282689.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(25), 723–773.

Gruen, T. W., Corsten, D. S., & Bharadwaj, S. (2002). Retail Out-of-Stocks: A Worldwide Examination of Extent, Causes and Consumer Response. Technical Report Grocery Manufacturers of America, Washington DC.

Harsha, P., Natarajan, R., & Subramanian, D. (2021). A prescriptive machine-learning framework to the price-setting newsvendor problem. *INFORMS Journal on Optimization*, *3*(3), 227–253. https://doi.org/10.1287/ijoo.2019.0046

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The elements of statistical learning. Springer.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. .

Huber, J., Müller, S., Fleischmann, M., & Stuckenschmidt, H., et al. (2019). A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, *278*(3), 904–915. https://doi.org/10.1016/j.ejor.2019.04.043

Hübner, A., Kuhn, H., & Kühn, S. (2016). An efficient algorithm for capacitated assortment planning with stochastic demand and substitution. *European Journal of Operational Research*, *250*(2), 505–520. https://doi.org/10.1016/j.ejor.2015.11.007

Hyndman, R. J., & Athanasopoulos, G. (2014). Forecasting: principles and practice. OTexts.

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.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations 2015*. San Diego.

Kök, A. G., & Fisher, M. L. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, *55*(6), 1001–1021. https://doi.org/10.1287/opre.1070.0409

Kök, A. G., Fisher, M. L., & Vaidyanathan, R. (2015). Assortment planning: Review of literature and industry practice. In N. Agrawal, & S. A. Smith (Eds.), *Retail supply chain management* (pp. 99–153). Springer US (*vol. 122*). https://doi.org/10.1007/978-0-387-78902-6

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.

Liu, C., Letchford, A. N., & Svetunkov, I. (2022). Newsvendor problems: An integrated method for estimation and optimisation. *European Journal of Operational Research*, *300*(2), 590–601. https://doi.org/10.1016/j.ejor.2021.08.013

Liu, M., Bai, Y., Qi, M., & Shen, Z.-J. M. (2024). Inventory management with LLM: Automated decision-making for order timing and quantity. Working paper. Available at SSRN: https://ssrn.com/abstract=3888897 or http://dx.doi.org/10.2139/ssrn.3888897.

Liyanage, L. H., & Shanthikumar, J. G. (2005). A practical inventory control policy using operational statistics. *Operations Research Letters*, *33*(4), 341–348. https://doi.org/10.1016/j.orl.2004.08.003

Mandi, J., Kotary, J., Berden, S., Mulamba, M., Bucarey, V., Guns, T., & Fioretto, F. (2024). Decision-Focused learning: Foundations, state of the art,benchmark and future opportunities. *Journal of Artificial Intelligence Research*, *81*, 1623–1701.

Musalem, A., Olivares, M., Bradlow, E. T., Terwiesch, C., & Corsten, D., et al. (2010). Structural estimation of the effect of out-of-stocks. *Management Science*, *56*(7), 1180–1197. https://doi.org/10.1287/mnsc.1100.1170

Müller, S., Huber, J., Fleischmann, M., & Stuckenschmidt, H. (2020). Data-driven inventory management under customer substitution. Working paper. Available at SSRN: https://ssrn.com/abstract=3624026 or http://dx.doi.org/10.2139/ssrn.3624026.

Netessine, S., & Rudi, N. (2003). Centralized and competitive inventory models with demand substitution. *Operations Research*, *51*(2), 329–335. https://doi.org/10.1287/opre.51.2.329.12788

Olivares-Nadal, A. V. (2024). Constructing decision rules for multiproduct newsvendors: An integrated estimation-and-optimization framework. *European Journal of Operational Research*, *315*(3), 1021–1037. https://doi.org/10.1016/j.ejor.2024.01.014

Oroojlooyjadid, A., Snyder, L. V., & and, M. T. (2020). Applying deep learning to the newsvendor problem. *IISE Transactions*, *52*(4), 444–463. https://doi.org/10.1080/24725854.2019.1632502

Parlar, M., & Goyal, S. K. (1984). Optimal ordering decisions for two substitutable products with stochastic demand. *OPSEARCH*, *21*(1), 1–15.

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.

Qi, M., Grigas, P., & Shen, Z.-J. M. (2023a). Integrated conditional estimation-optimization. Working paper. Available at https://arxiv.org/abs/2110.12351.

Qi, M., Shi, Y., Qi, Y., Ma, C., Yuan, R., Wu, D., & Shen, Z.-J. M. (2023b). A practical end-to-end inventory management model with deep learning. *Management Science*, *69*(2), 759–773. https://doi.org/10.1287/mnsc.2022.4564

van Ryzin, G. J., & Mahajan, S. (1999). On the relationship between inventory costs and variety benefits in retail assortments. *Management Science*, *45*(11), 1496–1509.

Sachs, A. L. (2015). Data-driven order policies with censored demand and substitution in retailing. In *Retail analytics* (pp. 57–78). Springer International Publishing. https://doi.org/10.1007/978-3-319-13305-8

Schlapp, J., & Fleischmann, M. (2018). Technical note - Multiproduct inventory management under customer substitution and capacity restrictions. *Operations Research*, *66*(3), 740–747.

Steinwart, I., & Christmann, A. (2008). Support vector machines. In *Information science and statistics*.

Team, R. C. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

Topaloglu, H. (2013). Joint stocking and product offer decisions under the multinomial logit model. *Production and Operations Management*, *22*(5), 1182–1199. https://doi.org/10.1111/j.1937-5956.2012.01423.x

Vulcano, G., van Ryzin, G., & Chaar, W. (2010). Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management*, *12*(3), 371–392. https://doi.org/10.1287/msom.1090.0275

Vulcano, G., van Ryzin, G., & Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research*, *60*(2), 313–334. http://pages.stern.nyu.edu/$\sim$gvulcano/PrimaryDemandEM.pdf.

Wan, M., Huang, Y., Zhao, L., Deng, T., & Fransoo, J. C., et al. (2018). Demand estimation under multi-store multi-product substitution in high density traditional retail. *European Journal of Operational Research*, *266*(1), 99–111. https://doi.org/10.1016/j.ejor.2017.09.014

van Woensel, T., van Donselaar, K., Broekmeulen, R., & Fransoo, J., et al. (2007). Consumer responses to shelf out-of-stocks of perishable products. *International Journal of Physical Distribution and Logistics Management*, *37*(9), 704–718. https://doi.org/10.1108/MRR-09-2015-0216

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, J., Xie, W., & Sarin, S. C. (2021a). Multiproduct newsvendor problem with customer-driven demand substitution: A stochastic integer program perspective. *INFORMS Journal on Computing*, *33*(3), 1229–1244. https://doi.org/10.1287/ijoc.2020.0996

Zhang, J., Xie, W., & Sarin, S. C. (2021b). Robust multi-product newsvendor model with uncertain demand and substitution. *European Journal of Operational Research*, *293*(1), 190–202. https://doi.org/10.1016/j.ejor.2020.12.023