Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data

Evangelos Spiliotis[a,*], Spyros Makridakis[b], Anastasios Kaltsounis[a], Vassilios Assimakopoulos[a]

[a] *Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece*
[b] *Institute For the Future, University of Nicosia, Cyprus*

Abstract

Supply chain management depends heavily on uncertain point forecasts of product sales. In order to deal with such uncertainty and optimize safety stock levels, methods that can estimate the right part of the sales distribution are required. Given the limited work that has been done in the field of probabilistic product sales forecasting, we propose and test some novel methods to estimate uncertainty, utilizing empirical computations and simulations to determine quantiles. To do so, we use the M5 competition data to empirically evaluate the forecasting and inventory performance of these methods by making comparisons both with established statistical approaches and advanced machine learning methods. Our results indicate that different methods should be employed based on the quantile of interest and the characteristics of the series being forecast, concluding that methods that employ relatively simple and faster to compute empirical estimations result in better inventory performance than more sophisticated and computer intensive approaches.

Keywords: Probabilistic Forecasting, Sales Forecasting, Time series, Empirical Evaluation, M5 Competition

*Corresponding author
Email address: spiliotis@fsu.gr (Evangelos Spiliotis)

1. Introduction

Capacity planning and inventory management relies on future estimates of product sales (Kourentzes et al., 2020). However, sales are typically characterised by intermittency (sporadic demand including zeros) and erraticness (variable demand sizes), making their prediction a challenging, uncertain task (Syntetos & Boylan, 2005). This is particularly true when forecasting on a daily basis at product-store level, where sales are disaggregated, increasing uncertainty (Zotteri et al., 2005; Nikolopoulos et al., 2011; Petropoulos et al., 2016).

Given the importance of product sales forecasting, numerous forecasting methods have been proposed in the literature for producing accurate, unbiased point forecasts (for a non-systematic review on the forecasting methods used for intermittent demand and count data, including parametric, non-parametric, and classification approaches, see section 2.7 in Petropoulos et al., 2020), like the Croston's method (Croston, 1972) and its variants (Syntetos & Boylan, 2005; Shale et al., 2006; Teunter et al., 2011). However, in supply chain management settings, estimating the right part of the sales distribution is equally important to estimating point forecasts since it is require for determining safety stock levels (do Rego & de Mesquita, 2015; Salinas et al., 2020).

Nevertheless, the research done in the field of product sales probabilistic forecasting is rather limited (Trapero et al., 2019a), attributed to the following three reasons. First, the methods utilized for producing point forecasts are not based on sound statistical models, meaning that probabilistic forecasts cannot be computed analytically (Spiliotis et al., 2019). Second, even when it is possible to compute analytical probabilistic forecasts, this can be difficult in the case of non-linear models (Hyndman et al., 2002). Third, given the irregular patterns that product sales display, making assumptions about the distribution of the demand size and inter-demand interval is challenging (Van der Auweraer et al., 2019), requiring approximating these variables instead (Willemain et al., 2004). Fildes et al. (2019) review the research done on retail demand forecasting and conclude that probabilistic forecasting is an under-investigated area, encouraging research in this direction while suggesting the connection of quantile, density, or volatility forecasting to inventory control.

In order to advance the theory and practice of forecasting in the field, the latest M competition, M5, requested its participants to precisely estimate the uncertainty distribution of the realized values of the more than 42 thousand time series that represent the hierarchical unit sales of Walmart (Makridakis et al., 2020a). Its uncertainty challenge (Makridakis et al., 2020c) attracted 1,137 participants from 94 countries, with the winning submissions, which involved tailored decision-tree-based models and powerful long short-term memory (LSTM) neural networks (NNs) among others, reporting forecast improvements higher than 20% over the top performing benchmark of the competition (ARIMA) according to the evaluation measure used (weighted scaled pinball loss, WSPL). However, these improvements referred mostly to the higher levels of the hierarchy (e.g. at state, store, or product category level), being less significant or even negative at product-store level and the right tail of the distribution, i.e. where forecasting is most challenging and critical from a supply chain management perspective. Moreover, despite the success of the winning submissions, less than 23% of the participants managed to outperform the benchmarks set by the organizers, highlighting the difficulties involved in such forecasting tasks and the lack of effective, "off-the-shelf" probabilistic methods for product sales forecasting. Equally important, the competition focused on forecasting performance measures, providing limited insights concerning the realized service levels and monetary costs. Finally, gains in forecasting performance were not evaluated in terms of additional computational requirements, an important aspect for the practice of forecasting, especially for large retailers that have to produce numerous forecasts on a regular basis and therefore seek for accurate, yet computational efficient forecasting solutions (Seaman, 2018).

In light of the above, this paper proposes a number of new methods that utilize empirical computations and simulations to compute probabilistic forecast, thus offering relatively simple, yet competitive approaches for determining safety levels. In contrast to typical benchmarks used in the field, the proposed methods make few or no assumptions about the distribution of the sales, while being relatively faster to compute and easy to implement in forecasting support system. In order to assess the value added by these methods in supply chain management settings, we evaluate their performance in terms of precision, utility

(inventory performance), and computational resources required to produce the forecasts, by comparing them to established statistical approaches and advanced machine learning (ML) models. To do so, we use a large number of real time series that correspond to the product-store level data of the M5 competition. Moreover, we test the significance of our results for different categories of sales data, as well as for various quantiles of practical interest.

The contribution of our study can be summarized as follows:

- We compare the forecasting performance of various forecasting methods for the case of sales probabilistic forecasting, including linear and non-linear, as well as parametric and non-parametric estimates. Our empirical evaluation considers the 30,490 product-store time series of the M5 competition and four indicative quantiles aimed at approximating the right part of the sales distribution, which is essential for determining inventory service levels.

- We propose some new and relatively faster to compute probabilistic forecasting methods that do not assume that errors are normally distributed. Instead, they estimate uncertainty by producing empirical computations and simulations to generate the quantiles.

- Consequently, we analyze the results of forecasting performance in terms of statistical significance and indicate which forecasting methods are more appropriate depending on (i) the characteristics of the series (intermittency and erraticness) and (ii) the specific quantile being predicted.

- We assess the utility of the examined probabilistic forecasting methods in terms of realized service levels and monetary costs by employing a stock-control simulation, providing insights about which forecasting methods are more appropriate for different categories of series.

- We evaluate the efficiency of the examined probabilistic forecasting methods by comparing the time required for their computation against their forecasting performance.

The rest of the paper is organized as follows. Section 2 presents the forecasting methods utilized in this study, both established and new ones. In section 3, we present the data set used for evaluating the methods considered and describe the experimental design of the study. The results are presented and discussed in section 4. Finally, section 5 concludes the study, referring to its advantages and drawbacks while suggesting some avenues for future research.

2. Forecasting methods utilized

Overall, we consider a number of methods that can be classified into linear and non-linear as well as into parametric and non-parametric, including the new empirical methods proposed in this study, established statistical approaches, advanced ML models, and estimations that are based on the specification of theoretical distributions.

For each forecasting method considered we first describe its main characteristics and then explain how its parameters and settings are determined. Note that some of the methods considered, and especially the non-linear ones, can be heavily parameterized in terms of hyper-parameters and training, resulting to different forecasts for different cases. For these methods we used the best-practices proposed in the literature, adjusting them, however, where necessary based on the particularities of the data set used for evaluation by using a validation process. For reasons of completeness, all the forecasting benchmarks used in the M5 uncertainty competition, being available at the M5 GitHub repository[1], have been included in this study along with the top three winning submissions that serve as the high-performing standards of comparison. The new methods proposed in this study have been made publicly available in a separate GitHub repository[2] to facilitate the replication of our results (Makridakis et al., 2018a).

A. M5 benchmarks

- **Naive**: A random walk model and the easiest one to compute (Hyndman & Athanasopoulos, 2018). The point forecast at time $n+h$, $\hat{Y}_{n+h}$, is equal to the last known

[1]Available here: https://github.com/Mcompetitions/M5-methods

[2]Available here: https://github.com/vangspiliot/Product_sales_probabilistic_forecasting

observation of the time series, Y_n, as follows:

$$\hat{Y}_{n+h} = Y_n,$$

where h is the forecasting horizon considered and n is the length of the series. The respective probabilistic forecast for quantile u, $Q_{n+h}(u)$, can then be computed as follows:

$$Q_{n+h}(u) = \hat{Y}_{n+h} + c \times \text{RMSE} \times \sqrt{h}, \tag{1}$$

where RMSE is the in-sample one-step-ahead root mean squared error of the model and c the approximate value of the u_{th} quantile point of the standard normal distribution (e.g. 0.67 for u=0.750, 0.97 for u=0.835, 1.96 for u=0.975, and 2.58 for u=0.995). Naive was implemented using the `naive()` function of the *forecast* package for R (Hyndman et al., 2020).

Observe that Equation 1 assumes that the forecast errors are normally distributed, which is highly unlikely in practice, especially for intermittent and erratic demand data. Moreover, RMSE is computed based on the in-sample errors of the model which may not represent its out-of-sample performance. In other words, the calculation assumes that the historical patterns that have been modeled will continue into the forecast period, which may not be the case (Tashman, 2000).

- **Seasonal Naive (sNaive)**: Like Naive, but this time the forecast at time $n + h$ is equal to the last known observation of the same period, thus capturing possible seasonal variations (Hyndman & Athanasopoulos, 2018). The point forecast of sNaive is given by

$$\hat{Y}_{n+h} = Y_{n+h-m(k+1)},$$

where k is the integer part of $(h-1)/m$ and m is the seasonal period (e.g. 12 and 7 for monthly and daily series, respectively). The probabilistic forecast for quantile u can then be computed as follows:

$$Q_{n+h}(u) = \hat{Y}_{n+h} + c \times \text{RMSE} \times \sqrt{k+1}. \tag{2}$$

Once again it is assumed that the forecast errors are normally distributed and that the historical patterns that have been modeled will continue into the future. sNaive was implemented using the `snaive()` function of the *forecast* package for R (Hyndman et al., 2020).

- **Simple Exponential Smoothing (SES)**: The simplest exponential smoothing model, aimed at predicting series without a trend (Gardner Jr., 1985). Although the performance of SES is expected to deteriorate for series displaying intermittency, it is commonly used in practice (Gardner, 2006; Rostami-Tabar et al., 2013). Point forecasts are calculated using weighted averages that decrease exponentially across time, specified through a smoothing parameter, a, as follows:

$$\hat{Y}_{n+h} = \sum_{j=0}^{n-1} a(1-a)^j Y_{n-j} + (1-a)^n l_0,$$

where l_0 is the initial level of the model. The probabilistic forecast for quantile u can then be computed as follows:

$$Q_{n+h}(u) = \hat{Y}_{n+h} + c \times \text{RMSE} \times \sqrt{1 + a^2(h-1)}. \tag{3}$$

Again, the probabilistic forecasts are produced analytically, assuming that the forecast errors are normally distributed. SES was implemented using the `ses()` function of the *forecast* package for R (Hyndman et al., 2020).

- **Exponential Smoothing models (ETS)**: A method which automatically provides the best exponential smoothing model, indicated through information criteria (Hyndman et al., 2002). In contrast to SES, ETS involves a variety of forecasting models that are appropriate for series displaying seasonality and/or trend. However, similarly to SES, probabilistic forecasts are produced analytically based to the underlying statistical model. For a few ETS models that there are no known formulas for producing probabilistic forecasts, a bootstrap method that simulates numerous future sample paths and finds the u_{th} quantile of the simulated data at each forecasting horizon is used instead. ETS was implemented using the `ets()` function of the *forecast* package for R (Hyndman et al., 2020).

- **Autoregressive Integrated Moving Average models (ARIMA)**: A method which automatically provides the best ARIMA model, indicated through information criteria (Hyndman & Khandakar, 2008). Similarly to ETS, this method involves a variety of forecasting models that are appropriate for series displaying seasonality and/or trend, with the probabilistic forecasts being produced analytically when possible. ARIMA was implemented using the `auto.arima()` function of the *forecast* package for R (Hyndman et al., 2020). ARIMA was the top performing benchmark of the M5 uncertainty competition, outperforming more than 75% of the participating teams.

- **Quantile empirical estimation (QEE)**: In this method the probabilistic forecasts are non-parametrically computed from the empirical distribution of the historical data points (Trapero et al., 2019b) as follows:

$$Q_{n+h}(u) = \hat{q}_{|Y}(u), \tag{4}$$

where $\hat{q}_{|Y}(u)$ is the estimated u_{th} quantile of the series. Observe that $Q(u)$ is independent of the forecasting horizon considered as it is assumed that the distribution of the data will remain the same in the forecast period. QEE was implemented using the

`quantile()` function of the *stats* package for R. By default, `quantile()` obtains the sample quantiles by focusing on the mode of the series (`type=7`). This parameterization was also used for constructing the "Kernel" benchmark of the M5 competition. However, in this study we considered `type=8` as its is recommended for estimating median-unbiased quantiles regardless of the distribution of the series, resulting also to slightly more accurate forecasts in the examined data set.

B. Theoretical estimations

- **Poisson theoretical estimation (Pois)**: Probabilistic forecasts are computed from a theoretical Poisson distribution fitted to the entire historical series using the `qpois()` function of the *stats* package for R. Similarly to QEE, the quantiles computed by Pois are independent of the forecasting horizon considered. However, in contrast to QEE, Pois prescribes the distribution of the data.

- **Negative binomial theoretical estimation (NB)**: Probabilistic forecasts are computed from a theoretical negative binomial distribution fitted to the entire historical series using the `qnbinom()` function of the *stats* package for R. The dispersion parameter and the mean of the series were determined using the fitdist() function of the *fitdistrplus* package for R. In rare cases where approximations failed, a Poisson distribution was assumed and quantiles were computed using the Pois method instead. Similarly to QEE and Pois, the quantiles computed by NB are independent of the forecasting horizon considered, following also a prescribed data distribution.

C. Proposed empirical methods

- **SES on empirically estimated quantiles**: QEE produces probabilistic forecasts in an empirical, non-parametric fashion, assuming that the distribution of the data will remain the same in the forecast period. Although this is a reasonable assumption, it is possible for the distribution of the data to change across time. If this is true, the quantiles computed for the complete series will not fully represent the ones of the

forecast period. In order to deal with this issue we consider an alternative method which applies QEE on continuous, non-overlapping windows of the original series to construct a new series which captures the variations of the empirical distribution of the predicted series across time. After constructing the series for the quantile of interest, u, a forecasting method of choice can be used to extrapolate the series and produce probabilistic forecasts. In this study we set the length of the windows equal to the forecasting horizon and used SES for forecasting. We also considered two different loss functions for determining the smoothing parameter of SES: the MSE (SES-QEE-mse) and the pinball loss (SES-QEE-pl), defined as

$$PL(u) = \begin{cases} u(Y - Q(u)) & , Y \geq Q(u) \\ (1-u)(Q(u) - Y) & , Y < Q(u). \end{cases} \tag{5}$$

In the first case, the smoothing parameter is selected so that the forecasts approximate the mean of the quantile being predicted, while in the second so that the forecasts are explicitly optimal for the quantile being predicted.

- **SES with empirically computed residual errors (SES-emp)**: In order to relax the normality assumption made when computing probabilistic forecasts analytically, we consider a method which adjusts the point forecasts produced by the forecasting method of choice based on the respective quantile computed for the empirical distribution of its residual errors, as follows:

$$Q_{n+h}(u) = \hat{Y}_{n+h} + \hat{q}_{|e}(u), \tag{6}$$

where e are the residual errors of the forecasting method. In this study we chose to utilize SES for producing the point forecasts so that the results produced are directly comparable to those of the analytical methods described earlier.

- **SES with simulated forecast errors**: Although SES-emp does not assume that forecast errors are normally distributed, its forecasts are still based on in-sample approximations. In this regard, SES-emp assumes that the patterns observed in the past will continue into the future. Moreover, since these errors refer to one-step-ahead forecasts, they may underestimate uncertainty significantly for the case of multiple-step-ahead forecasts. In order to overcome these issues, we consider a simulation method that allows the estimation of out-of-sample errors and constructs probabilistic forecasts based on their empirical distribution.

  The out-of-sample errors can be simulated by applying a validation procedure (also known as rolling-origin procedure) in which the original train sample is divided into N train and test windows (Tashman, 2000). The length of the test windows can be set equal to h, while the length of the train windows may vary depending on the total number of validations performed and the way the forecast origin is updated. In this study we considered two different alternatives, as follows:

  - *Overlapping windows*: We started with a train sample of $n-h-N+1$ observations and produced point forecasts for the following h periods. The actual data of the series at times $n-h-N+2 \dots n-N+1$ were used for computing the forecast errors. The origin was then increased by one and new forecasts were produced from the updated origin, this time using $n-h-N+2$ observations for training the forecasting method and the following $n-h-N+3 \dots n-N+2$ ones for computing the forecast errors. This process was repeated N times, i.e. until there were no observations left for evaluating the forecasts.
  - *Non-overlapping windows*: Similar to the previous method, but this time the origin was increased by h so that the data used for computing the forecast errors were unique per validation window.

  We set N equal to 100 and used SES for producing the point forecasts in order for the results of the proposed method to be directly comparable to those of the methods

described earlier (SES and SES-emp). We also set the minimum length of the train window equal to 10 so that there are enough data for producing meaningful forecasts. In a few cases where it was not possible to produce 100 validation windows, N was set equal to the maximum number of possible validations.

In addition to the above, we further considered two different methods for producing the probabilistic forecasts. In the first case, the complete sample of errors was used to estimate uncertainty, i.e. $N \times h$ errors. In the second case, uncertainty was estimated per forecasting horizon, i.e. using the N errors available per forecasting horizon. The probabilistic forecasts of each method were then computed as follows:

$$Q_{n+h}(u) = \hat{Y}_{n+h} + \hat{q}_{|e}(u), \tag{7}$$

$$Q_{n+h}(u) = \hat{Y}_{n+h} + \hat{q}_{|e_h}(u), \tag{8}$$

where e and e_h is the total amount of forecast errors and the forecast errors computed for the forecasting horizon h, respectively. By combining the above-mentioned set-ups, we ended up with four different forecasting methods, as follows:

- **SES-sim-o**: Overlapping windows are used for computing the point forecasts $\hat{Y}$, with the probabilistic forecasts $Q(u)$ being computed according to Equation 7.
- **SES-sim-no**: Non-overlapping windows are used for computing the point forecasts $\hat{Y}$, with the probabilistic forecasts $Q(u)$ being computed according to Equation 7.
- **SES-sim-o-fh**: Overlapping windows are used for computing the point forecasts $\hat{Y}$, with the probabilistic forecasts $Q(u)$ being computed according to Equation 8.
- **SES-sim-no-fh**: Non-overlapping windows are used for computing the point forecasts $\hat{Y}$, with the probabilistic forecasts $Q(u)$ being computed according to Equation 8.

- **Linear Quantile Regression (LQR)**: This method is based on the concept of traditional linear regression according to which a linear expression can be determined so that the point forecasts produced by the model are the optimal ones in terms of MSE minimization, i.e. properly approximate the mean of the series, as follows:

$$\hat{Y}_{n+h} = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + \cdots + b_k Y_{t-k},$$

where $b_0 \dots b_k$ are the coefficients of the model. This approach can be expanded for the case of probabilistic forecasting so that the forecasts produced are the optimal ones in terms of the PL function, which is subject to the quantile u being predicted (negative errors are penalized more for higher quantiles and vice versa). As such, the method can estimate the conditional quantiles of the series instead of its mean without assuming a particular parametric distribution for the data, nor a constant variance. The quantiles were determined using the *statsmodels* library for Python which applies the methodological approach proposed by Koenker & Hallock (2001).

- **Non-linear Quantile Regression**: Due to its linear nature, LQR cannot be generalized for time series that display complex, non-linear patterns. In order to deal with this issue we consider two non-linear quantile regression methods that build on regression trees (RTs), namely random forest (RF) and gradient boosting trees (GBT). In brief, RF is a combination of multiple RTs, each one depending on the values of a random vector sampled independently and with the same distribution (Breiman, 2001). On the other hand, GBT generates multiple independent trees, one at a time, so that the errors made by the previously trained tree is decreased (Freund & Schapire, 1997). Apart from having bigger learning capacities than single RTs, RF and GBT are more robust to noise and less likely to overfit on the data used for training (Friedman, 2002). Moreover, the vast majority of the top performing methods in the M5 uncertainty competition exploited LightGBM[3], a special, faster variant of GBT, high-

[3]https://lightgbm.readthedocs.io/en/latest/index.html

lighting the potential value of decision-tree-based models in the examined forecasting task.

Given that RF and GBT use a number of hyper-parameters that can significantly affect the generated forecasts, in this study we determined their values using grid search, an automated method that explores a set of different hyper-parameter values and computes the forecasting performance on a predefined validation set to find the optimal ones by minimizing the PL score. We focused on the most important hyper-parameters to be tuned per case, selecting the number of estimated trees from [500, 1000, 2000] and the number of variables randomly sampled at each split from [2, 5, 10] for the RF model, while the learning rate and the total number of trees considered from [0.01, 0.05, 0.1] and [500, 1000, 2000], respectively, for the GBT model. In addition, for both models, the size of the input vector, i.e. the number of historical observations used for computing the forecasts, was selected between [7, 14, 28, 49] in order for the look-back window to cover a reasonable number of full seasonal periods but avoid unnecessary complexity that larger vectors would have introduced to the training process. Both models were implemented using the *Scikit-learn* Python library.

We should clarify that, although a discrete, specialized RF and GBT model was created for each series, the "optimal" values of the hyper-parameters, as well as the size of the input window were the same for all the series per quantile considered to effectively reduce computational cost. In this regard, we used the last h observations of the train sample for validation and scaled the data before training between 0 and 1 to avoid computational problems and enable faster learning. After determining the optimal values of the hyper-parameters at a global level, we re-trained the models using the complete train set available to estimate the quantiles for the forecast period.

Specifically, the models were trained using the standard method of constant size, rolling input and output windows (Smyl, 2020; Spiliotis et al., 2020b). According to this method, the train set of each series, i.e. historical data, is split into multiple input and output vectors of sizes w_{in} and w_{out}, respectively, with each output vector containing

the observations that succeed the ones of the corresponding input vector. Thus, for each series of length n, a total of $n - w_{in} - w_{out} + 1$ train samples can be created. After training the models with these samples, the last w_{in} observations of the series can be used as input to forecast the following w_{out} periods. Since the examined data set involves daily sales data, spanning from Monday to Sunday, w_{in} was set equal to $k \times 7$, with k being selected between 1, 2, 4, and 7. On the other hand, given that standard decision-tree-based models provide a single output, w_{out} was set equal to one, a setting that also facilitates training for short series of limited historical observations (Mukhopadhyay et al., 2012). Overall, our validation results suggested that shorter look-back windows are required for optimally determining the mid-right part of the distribution (14/28 days for GBT/RF models at quantiles 0.750 and 0.835), while longer ones for specifying its right tail (28/49 days for GBT/RF models at quantiles 0.975 and 0.995), probably because larger time periods are required in the latter case for observing a sufficient amount of unusually high sales. However, the optimal values of the hyper-parameters of the models were constant across all the quantiles (500/1000 estimators were considered by the GBT/RF models, with the number of variables randomly sampled at each split being equal to 2 for the RF model and the learning rate of the GBT model equal to 0.01).

When ML models are used for producing multi-step-ahead point forecasts, this is typically done using a recursive approach (Bontempi et al., 2013; Makridakis et al., 2018b) according to which the first point forecast substitutes the last historical observation to produce the second point forecast, then the first two point forecasts substitute the last two historical observations to produce the third point forecast, and so on till all periods have been successfully predicted. Given that this approach is not trivial in probabilistic forecasting settings, with the output of the model (quantile) being different in nature from its input (realized values), in our experiments we considered a simple approach according to which all h periods are being forecast using the first quantile produced by the model. This method is visualized in Figure 1 using a toy

example of series.

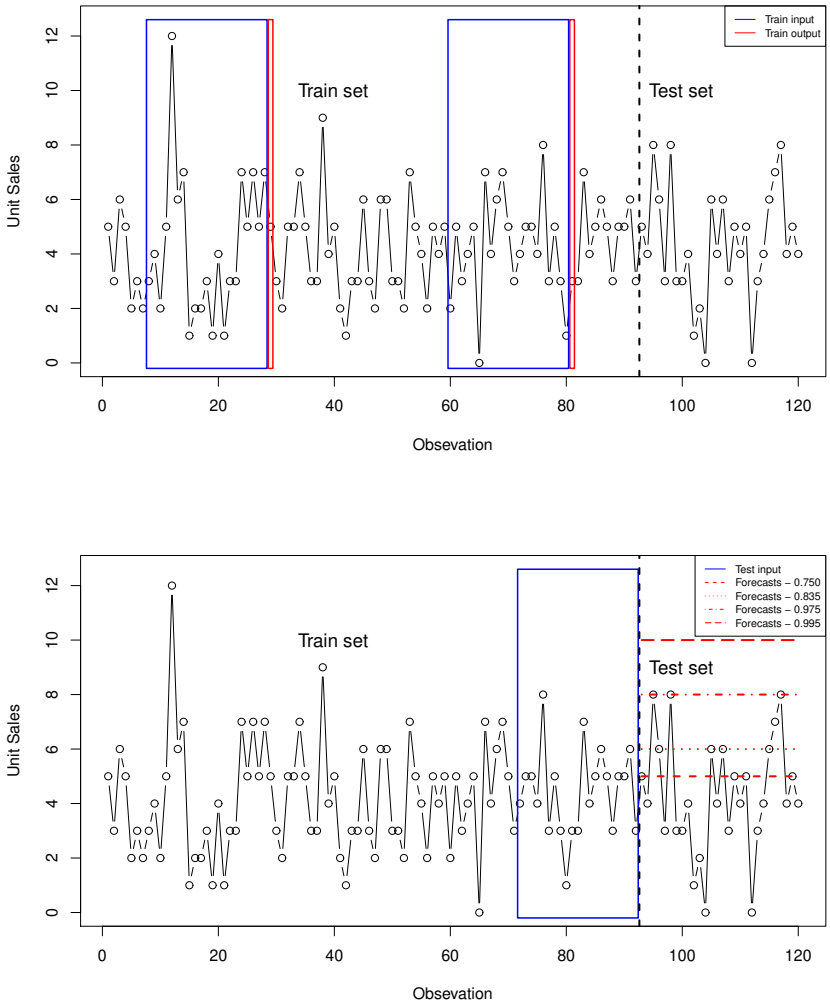

Figure 1: Toy example of the constant size, rolling input and output windows method used for training the RF and GBT models of the present study. Top: Multiple train and test windows of sizes $w_{in} = 21$ and $w_{out} = 1$ are created to form the train set of the forecasting model. The vertical dotted line indicates the forecast origin. Bottom: The last 21 historical observations are used to produce one-step-ahead forecasts for quantiles 0.750, 0.835, 0.975, and 0.995. These forecasts are then repeated $h = 28$ times to predict the complete forecasting horizon.

D. M5 winning submissions

- **Winning submissions of the M5 uncertainty competition**: Given that one of the primary objectives of this study is to evaluate the possible value added by more advanced forecasting methods, both in terms of forecasting performance and utility, we finally consider the top three winning submissions of the M5 uncertainty competition that serve as the competitive standards of comparisons. All the winning methods estimated the quantiles by performing non-linear regression using gradient-boosted trees or LSTM NNs that, in contrast to feed-forward ones, contain feedback connections to account for previous states along with the current input before producing the final output. The top ranked method (M5-w1) involved a total of 126 GBT models, trained for each quantile and aggregation level separately, using multiple hand-picked and crafted features as input (e.g. the days of the week and the month, special events and holidays, rolling means, medians, and quantiles of the past sales, and statistics about the skewness and kurtosis of the series, along with the percentage of zero observations). Sampling and augmentation techniques were also used to ensure model generalization, with the final forecasts incorporating blends of multiple folds, test-time augmentation, and reconciliation of the forecast levels. The runner-up (M5-w2) utilized a hybrid approach of GBT models, time series forecasting models, and statistical methods that estimated the probability distribution of the realized values of the series and applied corrections through external adjustments. The third winning submission (M5-w3) produced probabilistic forecasts by generating point forecasts through an ensemble of a GBT model and three NNs, and tuning them for each quantile using appropriate coefficients that were determined based on the last 28 days of the train set. More details about the winning submissions can be found in the work of Makridakis et al. (2020c).

3. Empirical evaluation

3.1. Data set

The probabilistic forecasts of the 21 methods described in the previous section were evaluated using a data set of 3,490 different products sold by Walmart between 2011-01-

29 and 2016-06-19 (1,969 days or approximately 5.4 years) in ten of its stores located in California (4 stores), Texas (3 stores), and Wisconsin (3 stores). Thus, the data set involves a total of 30,490 series, each one displaying the number of units sold for product i at store j. The data is daily and is organized into three product categories (foods, hobbies, and household) sold in seven departments (sub-categories). The data set is part of what has been used in the M5 uncertainty competition for determining its private leaderboard and corresponds to the 12^{th} (last) hierarchical level of the competition. This lowest, most disaggregate level of M5 was selected as the present study focuses on product sales data that is characterized by intermittency. For more information about the data set, please refer to the work of Makridakis et al. (2020a).

Following the suggestions of Syntetos & Boylan (2005), the series of the data set can be classified into four categories depending on their intermittency and their demand size erraticness. Intermittency is measured by ADI (average inter-demand interval), while demand size erraticness by CV^2 (squared coefficient of variation of the demand when it occurs). We used the thresholds (4/3 for ADI and 0.5 for CV^2) proposed by Kostenko & Hyndman (2006) to classify the series into intermittent, lumpy, smooth, and erratic, as shown in Figure 2. In total, the data set includes 22,355 intermittent, 5,204 lumpy, 879 erratic, and 2,052 smooth series. Observe that the vast majority of the series (90%) implies intermittency, mainly due to the low geographical level for which the sales are reported. Also, a significant proportion of the series (20%) displays erraticness, making the estimation of uncertainty challenging. Figure 3 provides four representative examples of the time series included in the data set, one for each category.

3.2. Experimental design

The empirical evaluation was performed using the first 1,941 days of the data set as the train set and the remaining 28 days as the test set. In this regard, the forecasting horizon considered was the same to the one used in the M5 competition (four weeks). For the case of the non-linear quantile regression forecasting methods, the last 28 days of the train sample were used for validation purposes so that the optimal hyper-parameters were appropriately

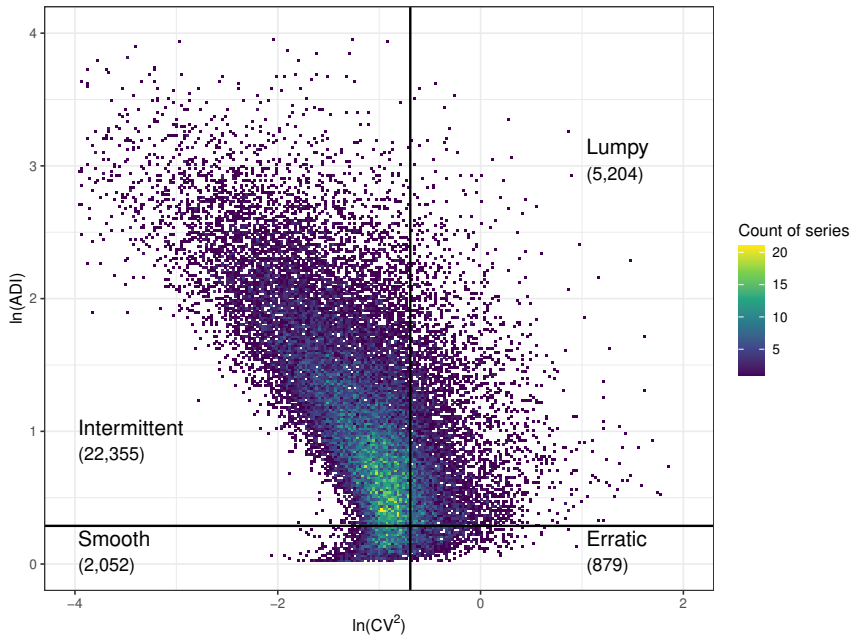

Figure 2: Demand classification of the 30,490 product-store series of the M5 data set based on their intermittency (ADI) and erraticness (CV^2), as suggested by Syntetos & Boylan (2005). The series are presented using a CV^2-ADI scatter-plot with logarithmic axes to allow for a more compact representation given the wide range of values present. In order to account for series density, the space is divided into 225 equal rectangles and a heatmap is used to illustrate the number of series contained in each rectangle. The bold horizontal (ADI=4/3) and vertical (CV^2=0.5) lines correspond to the thresholds proposed by Kostenko & Hyndman (2006) for classifying the series. In total, the data set includes 22,355 intermittent, 5,204 lumpy, 879 erratic, and 2,052 smooth series.

defined, as described in section 2.

The precision of the probabilistic forecasts was evaluated using the scaled pinball loss (SPL) function, as proposed in the guidelines of the M5 competition. The measure is calculated for each series and quantile of interest as follows:

$$\text{SPL(u)} = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} u(Y_t - Q_t(u))\mathbf{1}\left\{Y_t \geq Q_t(u)\right\} + (1-u)(Q_t(u) - Y_t)\mathbf{1}\left\{Y_t < Q_t(u)\right\}}{\frac{1}{n-1}\sum_{t=2}^{n} |Y_t - Y_{t-1}|}, \tag{9}$$

where $\mathbf{1}$ is the indicator function, being 1 if Y is within the postulated interval and 0

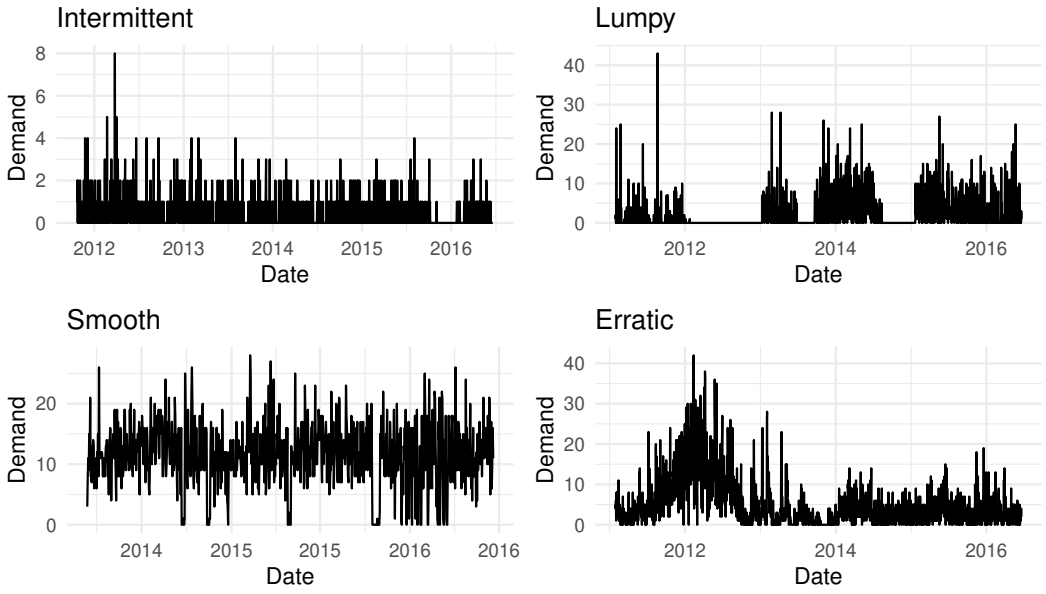

Figure 3: Example time series. From left to right and from top to bottom, an intermittent, a lumpy, a smooth, and an erratic demand time series is presented.

otherwise. Lower SPL values indicate better forecasting performance.

In order to effectively approximate the right part of the distribution of the series, which is essential for determining service levels, in this study we considered four indicative quantiles, namely the ones of a u value of u_1=0.750, u_2=0.835, u_3=0.975, and u_4=0.995. Thus, u_1 and u_2 provide useful information about the mid-right part of the distribution, while u_3 and u_4 provide critical information about its right tail, important in terms of the risk of extremely outcomes. By aggregating the SPL scores computed for each series and quantile separately, the average performance of the considered forecasting methods can be computed for all the 30,490 series of the data set. Similarly, the SPL scores can be averaged so that the forecasting performance of methods is computed per time series category, i.e. intermittent, lumpy, erratic, and smooth data.

Note that, originally, M5 considered nine quantiles, i.e. five more in addition to the ones examined in this study. We decided to exclude those that referred to the left part of the distribution and its median since they are less frequently used in capacity planning and

inventory management settings. Moreover, in the M5, the SPL scores were averaged based on the cumulative dollar sales of the series that approximate the actual value of each product for the company in monetary terms. In this study we decided to use equal weights instead so that the forecasting methods are ranked without putting more emphasis on fast-moving or more expensive products.

In order to assess the utility of the examined probabilistic forecasting methods in terms of realized service levels and monetary costs, we employed a stock-control simulation using an order-up-to policy with lost sales, as recommended by Teunter & Sani (2009) and Petropoulos et al. (2016). We considered four different target service levels of values equal to the quantiles used in the present study and set the inventory review period and lead time to one day. In other words, we assumed that each store is allowed to place new orders on a daily basis, with the products being delivered the following day; a practice which is typical in grocery retail firms. We initialized the stock-control simulation by assuming the initial stock to be equal to the average sales of the last two weeks of the train set and zero orders in the system. Finally, we used the first three days of the simulation as a burn-in period in order to reduce the impact of the selected stock initialization. Monetary cost was computed by tracking the holding stock (sum of stock stored during the simulation period) and the backlog (sum of unserviced units within the simulation period). In order for our results to represent reality as closely as possible, the annual item holding cost and the backlog cost of each product was set equal to 1% (Berling, 2008) and 25%[4] of its unit price, respectively, as determined by the sales price data of the M5 competition.

4. Results and discussion

Table 1 summarizes the performance of the 21 forecasting methods considered in this study per quantile (0.750, 0.835, 0.975, and 0.995). Thus, methods that are appropriate for estimating the mid-right part of the probability distribution and its right tail can be identified. The last four columns of the table display the ranks of the methods for each case.

[4]https://www.statista.com/statistics/269414/gross-profit-margin-of-walmart-worldwide-since-2006/

Table 1: Forecasting performance (SPL) of the 21 methods considered across the 30,490 product-store series of the M5 competition. The results are reported per quantile (0.750, 0.835, 0.975, and 0.995). Bold numbers indicate the top three performing methods per quantile.

Forecasting	Parametric	Linear	SPL				Rank			
method			0.750	0.835	0.975	0.995	0.750	0.835	0.975	0.995
M5 benchmarks										
Naive	✓	✓	1.278	1.169	0.346	0.092	21	21	21	16
sNaive	✓	✓	0.720	0.612	0.181	0.057	20	20	17	8
SES	✓	✓	0.558	0.476	0.177	0.084	16	12	12	15
ETS	✓	✓	0.554	0.472	0.176	0.084	15	10	11	14
ARIMA	✓	✓	0.570	0.487	0.179	0.084	18	14	14	13
QEE	✗	✓	0.548	0.493	0.159	0.050	13	17	6	4
Theoretical estimations										
Pois	✓	✓	0.541	0.492	0.203	0.097	11	15	19	20
NB	✓	✓	0.552	0.497	0.160	0.050	14	18	7	5
M5 winning submissions										
M5-w1	✗	✗	**0.509**	**0.455**	**0.151**	**0.048**	1	1	1	1
M5-w2	✗	✗	0.610	0.492	0.157	0.055	19	16	4	7
M5-w3	✗	✗	**0.513**	**0.457**	0.165	0.070	3	2	9	11
Proposed empirical methods										
SES-QEE-mse	✗	✓	0.519	0.470	0.181	0.094	6	9	16	17
SES-QEE-pl	✗	✓	**0.513**	**0.464**	0.180	0.096	2	3	15	19
SES-emp	✗	✓	0.522	0.468	**0.155**	**0.049**	8	7	3	3
SES-sim-o	✗	✓	0.519	0.465	0.163	0.058	4	4	8	9
SES-sim-no	✗	✓	0.522	0.468	**0.154**	**0.048**	9	8	2	2
SES-sim-o-fh	✗	✓	0.519	0.466	0.167	0.065	5	5	10	10
SES-sim-no-fh	✗	✓	0.523	0.474	0.178	0.079	10	11	13	12
LQR	✓	✓	0.565	0.519	0.209	0.097	17	19	20	21
RF	✗	✗	0.544	0.480	0.192	0.096	12	13	18	18
GBT	✗	✗	0.521	0.467	0.158	0.054	7	6	5	6

By studying the results of Table 1 we find that although M5-w1 is ranked first across the four quantiles examined, its improvements over the top performing empirical methods, such as SES-QEE-pl, SES-emp, and SES-sim-no, are insignificant, becoming negligible for quantiles 0.975 and 0.995. Similarly, we find that many of the proposed methods, as well as the M5 benchmarks and theoretical estimations, manage to outperform M5-w2 and M5-w3, ranked 2$^{\text{nd}}$ and 3$^{\text{rd}}$ in the M5 uncertainty competition among 892 teams, respectively. More importantly, we find that the relative ranks of the methods change considerably across the quantiles examined, meaning that, in general, no forecasting method outperforms the rest systematically across the four quantiles reported. For instance, SES-QEE-pl is ranked 2$^{\text{nd}}$ and 3$^{\text{rd}}$ at quantiles 0.750 and 0.835, but its performance deteriorates significantly at quantiles 0.975 and 0.995, being ranked 15$^{\text{th}}$ and 19$^{\text{th}}$, respectively. This indicates that each forecasting method is better in estimating different parts of the distribution of the series, with some of them being more appropriate for estimating its mid-right part, while others for approximating its right tail. In this regard, depending on the forecasting task, different methods could be selected to improve the overall forecasting performance.

We also find that all the forecasting methods that assume that the forecast errors are normally distributed (Naive, sNaive, SES, ETS, and ARIMA) are ranked at the lowest part of the table. This finding confirms the concerns expressed in the literature (Van der Auweraer et al., 2019; Willemain et al., 2004), suggesting that analytical computations that make strong assumptions about the distribution of the forecast errors are inappropriate for the case of product sales forecasting, supporting the utilization of empirical estimations. This becomes evident if we compare the performance of SES and SES-emp, where the latter method is found to result in more precise estimations for the four quantiles examined, being also the third-best performing method in modeling quantiles 0.975 and 0.995.

Another interesting finding is that the forecasting methods that estimate uncertainty by simulating the out-of-sample errors of the underlying forecasting method perform at least as well as the ones that build their estimations on in-sample errors. In particular, SES-sim-no is the only simulation method that outperforms SES-emp by a small margin at quantiles 0.975 and 0.995, with SES-sim-o and SES-sim-o-fh being the only simulation

methods that do better than SES-emp for quantiles 0.750 and 0.835, but still to a limited extent. Moreover, by comparing SES-sim-o to SES-sim-o-fh, as well as Sim-no to SES-sim-no-fh, we find that estimating uncertainty for each forecasting horizon separately does not have a positive effect, possibly due to the smaller samples of forecast errors used by the latter approaches for empirically estimating uncertainty. In addition, using non-overlapping windows instead of overlapping ones for simulating out-of-sample errors improves forecasting performance for the tail of the distribution but not its mid-right part, probably because larger parts of the series are taken into account in the second case, allowing SES-sim-no to observe more extreme occurrences of sales when compared to SES-sim-o.

Forecasting methods that empirically estimate the distribution of the series instead of their forecast errors are also found to be promising alternatives. QEE, a simple empirical method that produces forecasts by observing the distribution of the complete series reports good performance for quantiles 0.975 and 0.995, being ranked sixth and fourth, respectively. Accordingly, SES-QEE-pl is ranked second and third for quantiles 0.750 and 0.835, providing also, as expected, better forecasts than SES-QEE-mse, where MSE is used as an optimization criterion for estimating uncertainty instead of PL.

Regarding the quantile regression methods, we find that non-linear implementations are more successful than linear ones. This is particularly true for GBT which, in contrast to RF, significantly outperforms LQR for the four quantiles examined. Moreover, although GBT is not the best alternative for any of the quantiles considered, it is the most consistent empirical method in terms of ranks, both for the mid-right and the right tail of the distribution, ranging between the 5^{th} and 6^{th} position. Our results are therefore in alignment with those of recent studies that support the utilization of ML methods over conventional ones for the case of continuous, fast moving series (Makridakis et al., 2020b).

Interesting conclusions can also be drawn when empirical estimates of the distribution are compared with theoretical ones, namely Pois and NB. We find that, by and large, QEE, which does not prescribe the distribution of the data, provides more precise estimates than Pois and NB, particularly for quantiles 0.975 and 0.995. Moreover, NB outperforms Pois when estimating the tail of the distribution, suggesting that the examined data are more

likely to follow a NB distribution than a Pois one. Nevertheless, the differences reported between these three methods are minor for quantiles 0.750 and 0.835, probably because approximating the mid-right part of the distribution is easier than capturing its right tail.

Nevertheless, the results reported in Table 1 focus on the average performance of the examined forecasting methods per quantile, providing little evidence about their statistical significance. To further explore the results and identify the most appropriate forecasting methods per case, we apply the Multiple Comparisons with the Best (MCB) test that compares whether the average ranking of a forecasting method is significantly different than the others (Koning et al., 2005). If the confidence intervals of two methods overlap, their ranked performances are not statistically different, and vice versa. We perform this test for each quantile separately at a 5% level, using SPL for computing the ranks of the methods for each series.

Figure 4 presents the results of the MCB tests performed for the complete data set of the 30,490 series. In each graph, the methods are displayed from best (top row) to worst (bottom row) based on their average ranks. For each row, the black cell represents the method being evaluated, the blue cells, if any, suggest that the method being evaluated has an average rank that is not statistically different than the method in the respective column, while the white cells suggest statistically significant differences. As seen, depending on the quantile examined, different methods are found to perform significantly better, supporting the conclusion made earlier. More specifically, Pois and QEE are the best methods in estimating quantiles 0.750 and 0.835, providing significantly better forecasts than the rest of the approaches, even the advanced ML ones. In contrast, M5-w1, GBT, and SES-emp provide superior forecasts for quantile 0.975, while M5-w3 and Pois significantly better rankings for quantile 0.995. It is interesting to note that the results of the MCB tests are not in complete agreement with those reported in Table 1. This indicates that although some methods provide better forecasts for the majority of the series (better average rank), they tend to produce extreme errors in individual cases, thus leading to lower performance in terms of SPL. Undoubtedly, from an inventory-control point of view, the impact of these occurrences is expected to be subject to the value (size of demand multiplied by price) forecast wrongly.

In general, however, we find that relatively simple methods, estimating uncertainty either theoretically or empirically through simulations, are robust and can provide similar or even better results than more sophisticated ones.

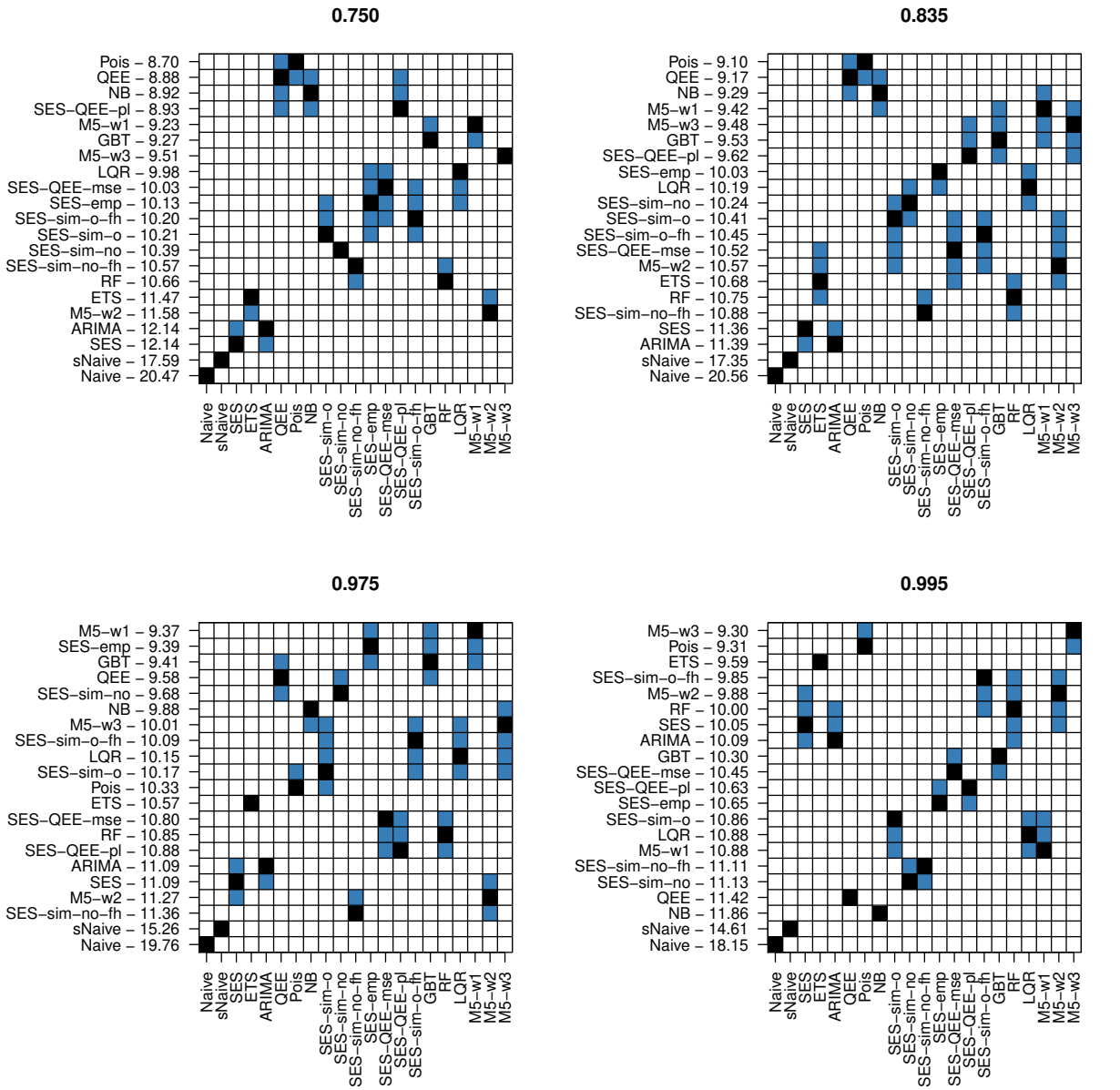

Figure 4: MCB significance tests for the 21 forecasting methods considered. The results are presented for the complete data set (30,490 series) for each quantile (0.750, 0.835, 0.975, and 0.995) separately. The ranks are computed according to SPL.

Given the individual characteristics of the series involved in the data set, we have expanded our study to investigate the performance of the various forecasting methods per time series category, i.e. intermittent, lumpy, smooth, and erratic ones. We have exploited the time series categorization made in the previous section and applied the MCB test for each sub-set of series separately, considering the average performance of the forecasting methods across the four quantiles. The results of the MCB tests are presented in Figure 5. As it can be seen, the category of the series proves to be equally important to the quantile being predicted for determining the most appropriate forecasting method. When the series are intermittent, simple empirical and theoretical estimations, like QEE and Pois, are found to provide the best results. On the other hand, when the series are smooth, erratic, or lumpy, M5-w1 manages to provide superior average ranks. Moreover, for smooth series, established statistical methods, like ETS, provide comparable ranks to advanced ML models, while for lumpy and erratic data, empirical and simulation methods, like SES-emp and SES-sim-o, are ranked just below the top performing methods of the M5 competition.

To assess the efficiency of the examined forecasting methods, we also compare their forecasting performance (SPL) across the 30,490 time series (indicatively for quantile 0.975) versus the total computational time (minutes) required for their estimation. Figure 6 shows the results in a logarithmic scale to facilitate comparisons. It can be observed that there is not any clear negative relationship between the two, indicating that more computationally expensive methods do not assure better post-sample accuracy. For instance, QEE, which results are computed in less that one minute, is more accurate than SES-sim-o and M5-w3, which forecasts are estimated in more than one and six hours, respectively. We also find that most of the proposed empirical methods manage to provide more accurate results than the approaches that build on theoretical estimations or assume normally distributed errors, while requiring, at the same time, lower computational times. Similarly, the forecasting performance of the proposed empirical methods is on par if not better than that of more sophisticated ML approaches, despite being significantly less computationally intensive. Given these findings, decision makers would need to consider balancing computational cost against the expected accuracy in order to increase the overall efficiency of their inventory manage-

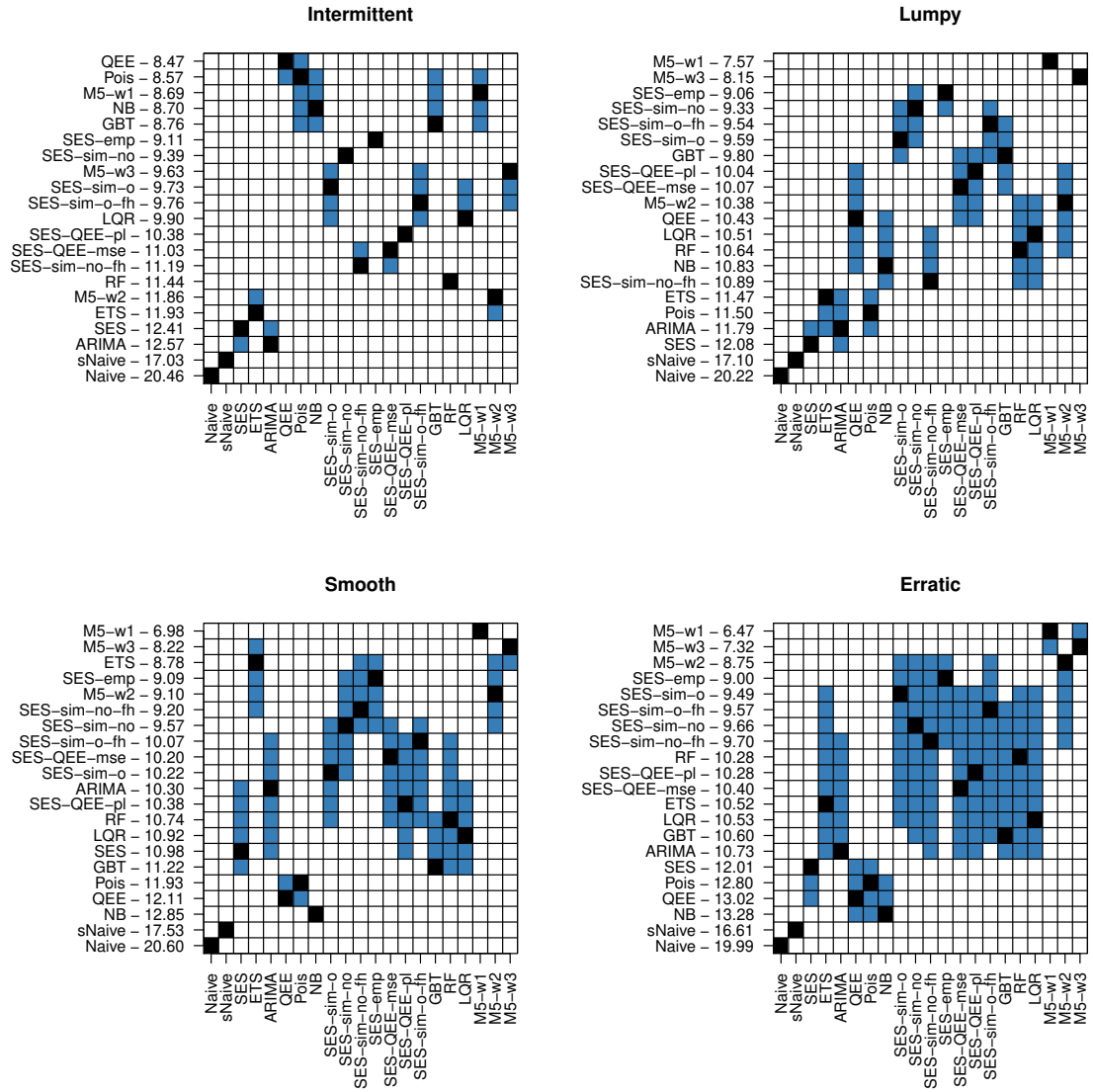

Figure 5: MCB significance tests for the 21 forecasting methods considered. The results are presented for intermittent, lumpy, smooth, and erratic demand series separately. The average SPL score of the forecasting methods across quantiles 0.750, 0.835, 0.975, and 0.995 is used for ranking their performance.

ment activities.

As a final step in our analysis, we assess the inventory performance of the forecasting methods examined, as determined by the stock-control simulation performed. To simplify the comparisons and facilitate discussion, we focus on an indicative sample of methods that

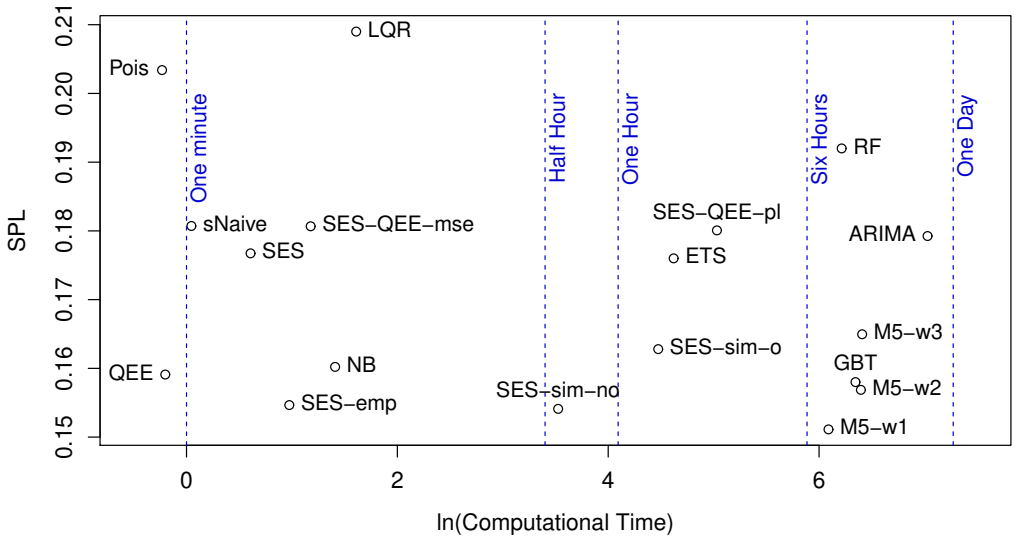

Figure 6: Forecasting performance (SPL) for the forecasting methods considered in this study compared to the computational time required (minutes) to obtain probabilistic forecasts for all the 30,490 time series. The statistics are computed indicatively for quantile 0.975. Computational times were estimated using a system of the following characteristics: 4 cores and 8 logical processors at 3.60 GHz, 16 GB RAM, 1 TB HDD, Microsoft Windows 10.

were previously found to provide significantly more precise estimates, either for a particular quantile or time series category, considering, however, in addition relatively simpler methods with similar forecasting performance. The results are summarized in Figure 7 where the left graph compares the realized service levels to the target ones and the right graph displays the monetary costs against the deviations of the target service level. Ideally, a method should achieve the target service level with the smallest amount of cost.

By observing the left graph of Figure 7 we find that all of the selected methods result in higher service levels than originally specified for quantile 0.750. In contrast, they underestimate demand for service levels 0.975 and 0.995, with sNaive being the only exception, possibly because it generally tends to produce wider intervals and, as a result, overestimates on average demand to a greater extent. Moreover, we find that the winning method of the

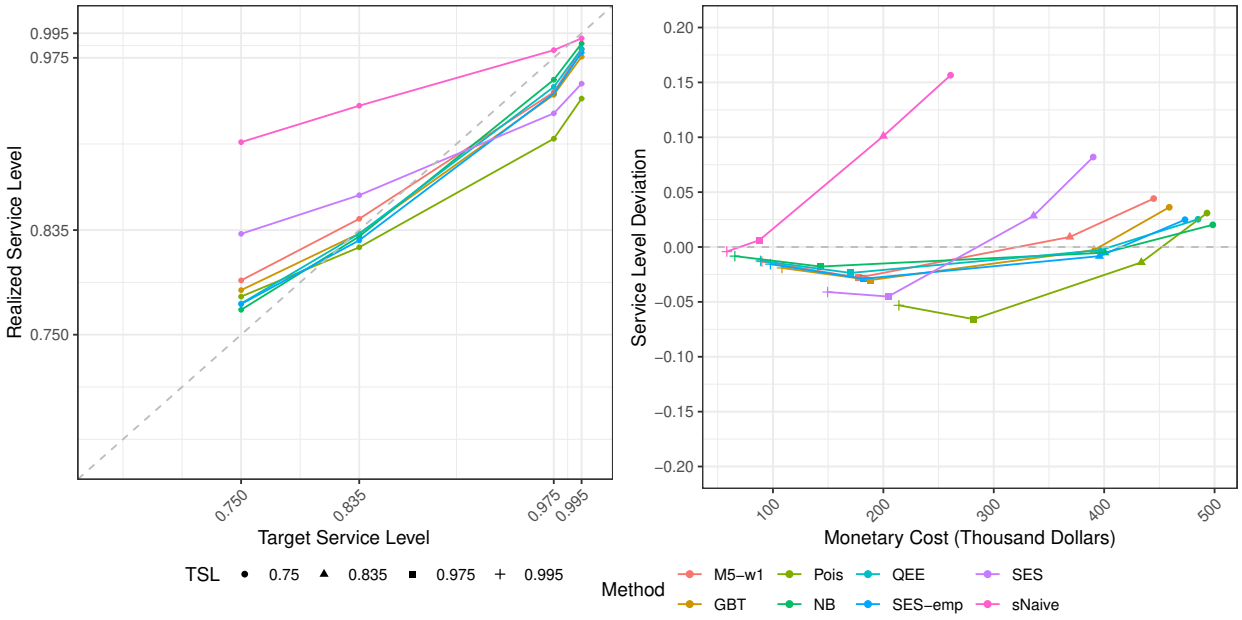

Figure 7: Inventory performance of selected forecasting methods based on the simulation performed. Left: Realized service levels against target service levels (TSL). Right: Monetary costs (sum of backlog and holding cost) against service level deviations. The results are presented for the complete data set (30,490 series).

M5 uncertainty competition displays a poor calibration for the two lowest service levels, being outperformed both by GBT, a local ML model that does not consider any explanatory variables, and approaches that build on empirical or theoretical computations. For service levels 0.975 and 0.995, by and large, all of the selected methods display similar inventory performance with the exception of the Pois method which significantly underestimates sales. On average, however, SES-emp, QEE, and NB provide better matches between the target and the realized service levels.

Similar conclusions can be drawn based on the right graph of Figure 7. As seen, NB and sNaive result in lower costs for target service levels 0.975 and 0.995, followed by QEE, M5-w1, SES-emp, and GBT. Since the vast majority of the methods underestimate demand at these levels, the value added by NB and sNaive is mostly driven by the lower amount of sales being lost. In contrast, sNaive, SES, and M5-w1 result in lower costs for service levels 0.835 and 0.750. However, this is because these methods (significantly) overestimate

demand, thereby generating less lost sales compared to the rest of the methods. Therefore, we conclude that GBT, QEE, and NB could be more appropriate for minimizing costs when the target level is set to 0.835 or 0.750.

We expand the analysis performed on the utility of the selected forecasting methods by comparing the realized service levels against the target service levels considered, for each time series category separately. Our results, visualized in Figure 8, suggest that demand intermittency and erraticness can significantly affect inventory performance, requiring the utilization of different types of forecasting methods. Specifically, we find that ML and empirical methods are well calibrated when used to forecast lumpy series, in contrast to statistical ones that overestimate demand significantly for target service levels 0.750 and 0.835. In addition, all of the selected methods tend to overestimate the mid-right part of the uncertainty distribution when used to predict intermittent data, despite estimating precisely the tail of the distribution. Moreover, with the exception of sNaive, all methods underestimate the demand of the smooth and erratic series across all parts of the uncertainty distribution. We also observe that the inventory performance of the methods is similar among the intermittent-lumpy and smooth-erratic categories. This finding indicates that inter-demand interval has a greater impact on realized service levels compared to demand size erraticness, meaning that the existence of sporadic demand is critical for the selection of appropriate forecasting methods.

Overall, our simulation results suggest that superior forecasts do not always guarantee better inventory performance. In other words, although precise forecasting methods are more likely to result in more calibrated service levels and lower monetary costs, it is still possible for less sophisticated models to allow for better inventory control. Moreover, our results indicate that different methods may be more suitable for supporting supply chain management decisions, depending on the target service level set and the category of the demand patterns. Therefore, empirical approaches, such as QEE and SES-emp, could be used to forecast intermittent and lumpy series, while simple statistical methods, like sNaive, to predict smooth and erratic data. Additionally, we should also consider the computational requirements of forecasting methods. Taking into account that QEE, NB, and SES-emp

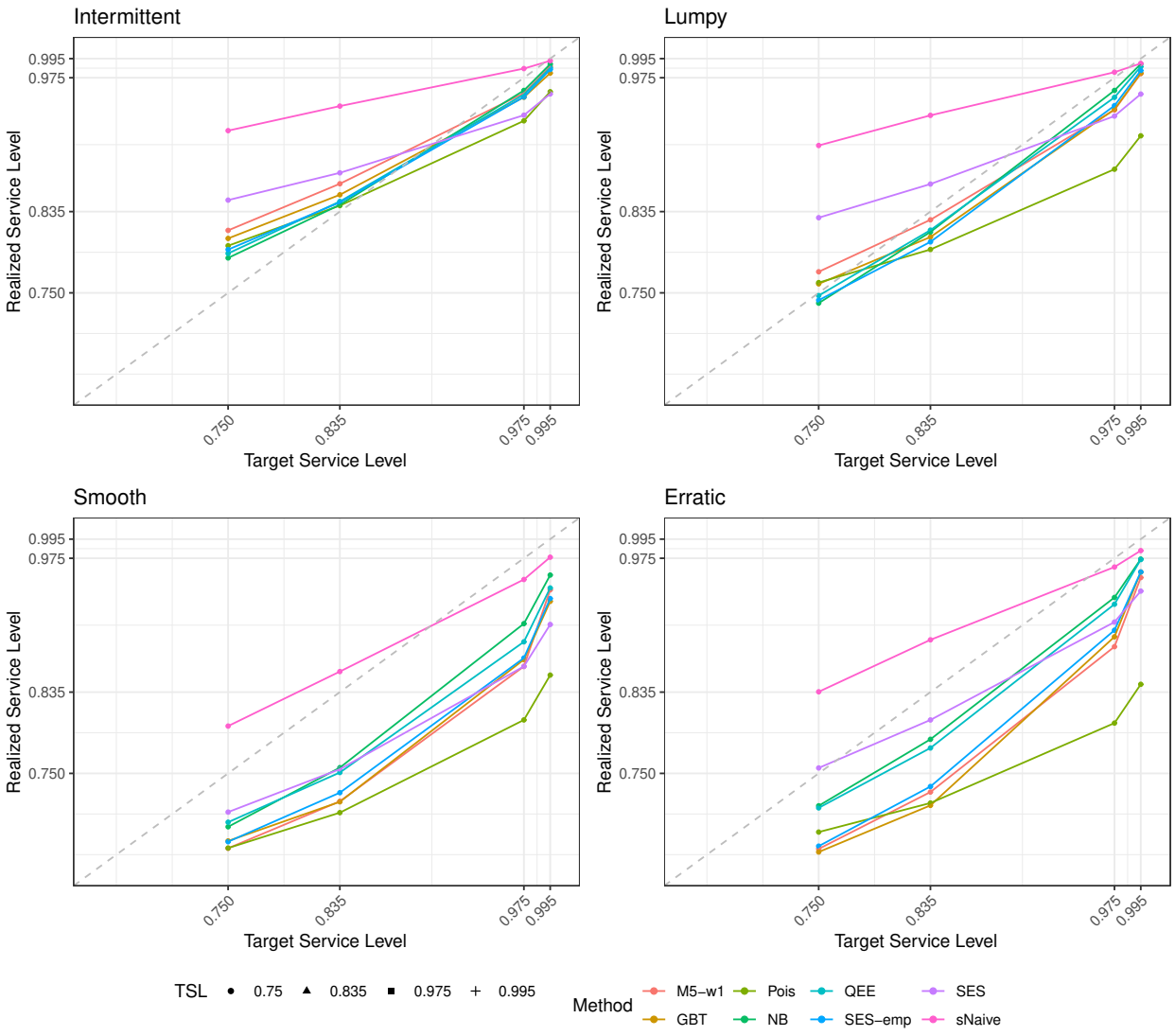

Figure 8: Realized service levels of selected forecasting methods based on the simulation performed against target service levels (TSL). The results are presented for intermittent, lumpy, smooth, and erratic demand series separately.

were found being several orders of magnitude cheaper than M5-w1, it becomes evident that, from an operational perspective, relatively simpler methods that estimate uncertainty through empirical computations and simulations may still be relevant in inventory control settings. This is particularly true for large retailers like Walmart or Target where hundreds of thousands or even billions of forecasts must be produced on a weekly basis (Seaman,

2018), thereby requiring the utilization of robust methods that are fast to run and easy to integrate with the rest of the firms' computerize systems.

5. Conclusions

Product sales probabilistic forecasts are indispensable for effective capacity planning and inventory management. Given the limited work done in this area, our large-scale study introduced several novel methods that estimate quantiles through empirical computations and simulations. Consequently, the performance of the proposed methods was compared to that of established statistical and advanced machine learning approaches using the large data set of the M5 competition. In addition, their computational requirements were investigated to evaluate their applicability and ease of use. In both cases the results were encouraging as the accuracy of the empirical methods was similar or better than the accuracy of the statistical and machine learning ones, while their computational time was considerably lower, opening up the usage of these methods to estimate probabilistic forecasts practically.

The most important finding of our study was that no forecasting method outperformed the rest across all the quantiles considered, implying that different methods could be utilized for each quantile in order to improve overall performance. The same conclusion was drawn when considering their performance for different categories of series in terms of intermittency and erraticness. This finding confirms George Box's famous quote, "all models are wrong, but some are useful", as well as the belief that there are "horses for courses" (Petropoulos et al., 2014), i.e. methods that are better tailored to some types of data (Spiliotis et al., 2020a) while at the same time being computationally cheaper to apply (Nikolopoulos & Petropoulos, 2018).

An additional, useful finding was that methods that assumed that the forecast errors are normally distributed were outperformed by those that empirically estimated the probability distributions when applied to predict intermittent and lumpy series. Furthermore, it was concluded that methods that approximated uncertainty by estimating the distribution of the series empirically perform better than those computing the distribution of the forecast errors analytically. Yet, this was not always the case for smooth and erratic data, where

statistical methods displayed an advantage. Given these findings, new probabilistic forecasting methods could estimate the uncertainty present by focusing on the characteristics of the data series themselves rather than relying explicitly on the abilities of the forecasting methods used to do so. Moreover, improvements were found for the case of non-linear quantile regression models. Interestingly, not only was the performance of these methods good on average, but it was also consistent across the various quantiles. Although recently introduced, machine learning has been proven an effective approach for estimating probabilistic forecasts (Makridakis et al., 2020b), with cross-learning algorithms being an example of how it should be further advanced to improve the estimation of sales distributions.

The findings of this study can be of value to both the academic community and practitioners, even though future research maybe required to verify them and determine if they can be expanded to other areas beyond daily, retail demand forecasting of "brick-and-mortar" stores that drive their sales through a "constant low prices" strategy (Makridakis et al., 2020a). Moreover, research would be needed to suggest ways for selecting or combining the available probabilistic forecasting methods and the new ones introduced by this study, a practice commonly used for improving the accuracy of point forecasts, but so far largely ignored in the probabilistic forecasting literature (Makridakis et al., 2020b). Finally, the results of this study suggest that the value added by more sophisticated methods need to be carefully examined based on their impact on actual inventory management, keeping in mind that relatively simpler and computationally less expensive approaches, like the empirical methods proposed in this study, may provide similar or even better results in terms of inventory service levels, at lower computational costs.

References

Berling, P. (2008). Holding cost determination: An activity-based cost approach. *International Journal of Production Economics*, *112*, 829–840. Special Section on RFID: Technology, Applications, and Impact on Business Operations.

Bontempi, G., Ben Taieb, S., & Le Borgne, Y.-A. (2013). Machine learning strategies for time series forecasting. In M.-A. Aufaure, & E. Zimányi (Eds.), *Business Intelligence: Second European Summer School,*

eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures (pp. 62–77). Berlin, Heidelberg: Springer Berlin Heidelberg.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, *23*, 289–303.

Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, .

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*, 367–378.

Gardner, E. S. (2006). Exponential smoothing: The state of the art—part ii. *International Journal of Forecasting*, *22*, 637–666.

Gardner Jr., E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*, 1–28.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2020). *forecast: Forecasting functions for time series and linear models*. R package version 8.12.

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *26*.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. (2nd ed.). Melbourne, Australia: OTexts.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439–454.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, *15*, 143–156.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*, 397–409.

Kostenko, A. V., & Hyndman, R. J. (2006). A note on the categorization of demand patterns. *Journal of the Operational Research Society*, *57*, 1256–1257.

Kourentzes, N., Trapero, J. R., & Barrow, D. K. (2020). Optimising forecasting models for inventory planning. *International Journal of Production Economics*, *225*, 107597.

Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018a). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*, 835–838.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018b). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*, 1–26.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020a). *The M5 competition: Background, organization and implementation*. Working paper available at: https://drive.google.com/drive/u/1/folders/1S6IaHDohF4qalWsx9AABsIG1filB5V0q.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020b). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*, 54–74.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R. L. (2020c). *The M5 Uncertainty competition: Results, findings and conclusions*. Working paper available at: https://drive.google.com/drive/u/1/folders/1S6IaHDohF4qalWsx9AABsIG1filB5V0q.

Mukhopadhyay, S., Solis, A. O., & Gutierrez, R. S. (2012). The accuracy of non-traditional versus traditional methods of forecasting lumpy demand. *Journal of Forecasting*, *31*, 721–735.

Nikolopoulos, K., & Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? *Computers & Operations Research*, *98*, 322–329.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., & Assimakopoulos, V. (2011). An aggregate–disaggregate intermittent demand approach (adida) to forecasting: an empirical proposition and analysis. *Journal of the Operational Research Society*, *62*, 544–554.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Bergmeir, C., Bessa, R. J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., Baets, S. D., Dokumentov, A., Fiszeder, P., Franses, P. H., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Arenas, J. R. T., Wang, X., Winkler, R. L., Yusupova, A., & Ziel, F. (2020). Forecasting: theory and practice.

Petropoulos, F., Kourentzes, N., & Nikolopoulos, K. (2016). Another look at estimators for intermittent demand. *International Journal of Production Economics*, *181*, 154–161. SI: ISIR 2014.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). 'horses for courses' in demand forecasting. *European Journal of Operational Research*, *237*, 152–163.

do Rego, J. R., & de Mesquita, M. A. (2015). Demand forecasting and inventory control: A simulation study on automotive spare parts. *International Journal of Production Economics*, *161*, 1–16.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., & Ducq, Y. (2013). Demand forecasting by temporal aggregation. *Naval Research Logistics (NRL)*, *60*, 479–498.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*, 1181–1191.

Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, *34*, 822–829.

Shale, E. A., Boylan, J. E., & Johnston, F. R. (2006). Forecasting for intermittent demand: the estimation of an unbiased average. *Journal of the Operational Research Society*, *57*, 588–592.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*, 75–85.

Spiliotis, E., Assimakopoulos, V., & Nikolopoulos, K. (2019). Forecasting with a hybrid method utilizing data smoothing, a variation of the Theta method and shrinkage of seasonal factors. *International Journal of Production Economics*, *209*, 92–102.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020a). Are forecasting competitions data representative of the reality? *International Journal of Forecasting*, *36*, 37–53.

Spiliotis, E., Makridakis, S., Semenoglou, A.-A., & Assimakopoulos, V. (2020b). Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research: An International Journal*, (pp. 1–25).

Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, *21*, 303–314.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*, 437–450.

Teunter, R., & Sani, B. (2009). On the bias of Croston's forecasting method. *European Journal of Operational Research*, *194*, 177–183.

Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, *214*, 606–615.

Trapero, J. R., Cardós, M., & Kourentzes, N. (2019a). Empirical safety stock estimation based on kernel and garch models. *Omega*, *84*, 199–211.

Trapero, J. R., Cardós, M., & Kourentzes, N. (2019b). Quantile forecast optimal combination to enhance safety stock estimation. *International Journal of Forecasting*, *35*, 239–250.

Van der Auweraer, S., Boute, R. N., & Syntetos, A. A. (2019). Forecasting spare part demand with installed base information: A review. *International Journal of Forecasting*, *35*, 181–196.

Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, *20*, 375–387.

Zotteri, G., Kalchschmidt, M., & Caniato, F. (2005). The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, *93-94*, 479–491.