International Journal of Forecasting 38 (2022) 1346–1364

Contents lists available at ScienceDirect

International Journal of Forecasting

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

M5 accuracy competition: Results, findings, and conclusions

Spyros Makridakis [b], Evangelos Spiliotis [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

ARTICLE INFO

Keywords:
Forecasting competitions
M competitions
Accuracy
Time series
Machine learning
Retail sales forecasting

ABSTRACT

In this study, we present the results of the M5 "Accuracy" competition, which was the first of two parallel challenges in the latest M competition with the aim of advancing the theory and practice of forecasting. The main objective in the M5 "Accuracy" competition was to accurately predict 42,840 time series representing the hierarchical unit sales for the largest retail company in the world by revenue, Walmart. The competition required the submission of 30,490 point forecasts for the lowest cross-sectional aggregation level of the data, which could then be summed up accordingly to estimate forecasts for the remaining upward levels. We provide details of the implementation of the M5 "Accuracy" challenge, as well as the results and best performing methods, and summarize the major findings and conclusions. Finally, we discuss the implications of these findings and suggest directions for future research.

1. Introduction

Forecasts are indispensable for many of the decisions that we make, such as time to get up in the morning in order to not be late for work, or the brand of television to buy that provides the best value for money. Supermarkets require forecasts to support their strategic development, make tactical decisions, and manage their demand and supply planning processes in order to avoid customer service issues and high inventory costs (Fildes et al., 2019). Thus, it is important that these forecasts should be as accurate as possible because stocking too many products incurs extra costs whereas stocking insufficient would lead to lost sales and low profits. The M competitions have been conducted for almost 40 years (Makridakis et al., 1982, 1993; Makridakis & Hibon, 2000; Makridakis et al., 2020c), and they aim to identify ways to improve the accuracy of forecasting by empirically evaluating several forecasting methods and identifying the most accurate. The findings obtained in these competitions have significantly influenced the theory and practice of forecasting by providing valuable insights into how the forecasting accuracy can be improved (Hyndman, 2020). For example, the first three competitions demonstrated the value of combining, the potential of automatic forecasting methods, and the merits of simplicity, and the fourth competition showed that machine learning (ML) methods and a hybrid approach utilizing "cross-learning" (Semenoglou et al., 2021) obtained more successful forecasts than the alternatives.

The M5 competition extended the objectives of the previous four competitions by focusing on a retail sales forecasting application and using real-life, hierarchically structured sales data with intermittent and erratic characteristics (Syntetos & Boylan, 2005; Syntetos et al., 2005). The competition attracted many participants who were eager to experiment with effective forecasting solutions in the real-life situation faced by numerous retail companies on a daily basis. LightGBM is a decision tree-based ML approach with reportedly superior forecasting performance compared with all other alternatives and it was used in practice by all of the top 50 competitors, thereby

* Corresponding author.
E-mail address: spiliotis@fsu.gr (E. Spiliotis).

https://doi.org/10.1016/j.ijforecast.2021.11.013

indicating that this method can be adopted by retail firms to improve the accuracy of their sales predictions and daily operation. However, it was also found that simple to implement and computationally cheap methods such as exponential smoothing were still competitive, especially when used to produce forecasts at the product or product-store level. In addition, the M5 results confirmed most of the key findings obtained in the previous M competitions, thereby further advancing the theory and practice of forecasting in the area of hierarchical retail sales. In the present study, we present the results of the competition and best performing methods, compare their performance with statistical and other benchmarks, and summarize the key findings and conclusions.

2. Implementation and execution

The M5 "Accuracy" competition was organized following the general principles described by Makridakis et al. (2022). The competition began on March 3rd, 2020, when the initial training data set became available to download on the Kaggle platform,[1] and it ended on June 30th, 2020, when the final leaderboard was announced. Moreover, the competition was chronologically divided into two phases, which were used to evaluate the teams. The first "validation" phase was used to allow the teams to receive feedback about their performance and guide the development of their forecasting models. The second "test" phase was used for the final evaluations of the teams. The data set used comprised the unit sales of 3,049 products sold by Walmart in the USA, which were organized in the form of grouped time series aggregated based on their type (category and department) and selling location (stores and states), with a total of 42,840 series in 12 cross-sectional, aggregation levels.

The implementation of the M5 "Accuracy" competition differed from the "Uncertainty" competition (Makridakis et al., 2020d) in terms of the following four aspects: (i) submission template, (ii) performance measure, (iii) prizes, and (iv) benchmarks. The first three aspects are described in the following subsections, and the last in the appendix of the supplementary material.

2.1. Submission

All forecasts were submitted through the Kaggle platform using the template provided by the organizers. The template for the M5 "Accuracy" competition required submitting forecasts that corresponded to only 30,490 series of the lowest cross-sectional aggregation level of the data set (level 12) rather than all 42,840 series in the competition because the M5 series are hierarchically structured, and thus we expected the corresponding forecasts to be coherent (forecasts at the lower levels have to sum up to the forecasts at the higher levels so that the forecasts across different levels are aligned; Spiliotis et al., 2019b). Thus, it was assumed that the forecasting approaches used by the contestants for predicting all 42,840 series in the competition would result in coherent forecasts, so the forecasts at the lowest aggregation level could be appropriately aggregated (summed up) to automatically compute those at the remaining levels.

It should be noted that the submission template did not affect how the forecasts were produced and teams were completely free to use their preferred forecasting method to forecast the individual series. However, the submission template ensured that the forecasts were coherent and in an appropriate form for direct evaluation. For example, a team could simply forecast the series at the most disaggregated level of the competition (level 12) and derive the remaining forecasts using a bottom-up approach. Another team might only forecast the most aggregated series of the competition (level 1) and compute the remainder using proportions (top-down method). A mixture of these two approaches was also possible (middle-out method). Finally, another option involved predicting the series at all levels and obtaining those at the lowest level by using an appropriate weighting scheme (Hyndman et al., 2011). The benchmarks in the competition applied some of these options, including some indicative forecasting approaches that utilized bottom-up and top-down methods, as well as a combination of both (Abouarghoub et al., 2018).

Teams were allowed to submit a maximum of five entries per day on the Kaggle platform. However, for their final evaluation in the "test" phase, each team had to select a single set of forecasts (one submission) because in real life, forecasters have the same problem of selecting a single set of forecasts that they believe will represent the future as adequately as possible. If no particular submission was selected, that with the highest performance during the "validation" phase was automatically selected by the system.

2.2. Performance measure

Various measures have been used for evaluating the point forecast accuracy in previous studies (Hyndman & Koehler, 2006). The first three M competitions employed several of these measures, and M4 examined the overall weighted average of the symmetric mean absolute percentage error (sMAPE; Makridakis, 1993) and a variant of the mean absolute scaled error (MASE; Hyndman & Koehler, 2006). Undoubtedly, no measure is perfect because all have advantages and drawbacks (Goodwin & Lawton, 1999; Kolassa, 2020; Koutsandreas et al., 2021). The comments made about the measures utilized in all previous M competitions by the invited commentators clearly demonstrated this lack of agreement, and they also highlighted the preferences of each forecaster (Makridakis et al., 2020b). Based on the measures used commonly for assessing the forecasting accuracy in previous studies, we consider that those based on scaled errors probably have the most appropriate statistical properties. Thus, the M5 "Accuracy" competition utilized a variant of the MASE originally proposed by Hyndman and Koehler (2006) called the root mean squared scaled error (RMSSE),

[1] https://www.kaggle.com/c/m5-forecasting-accuracy.

which is defined as follows:

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(y_t - \hat{y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(y_t - y_{t-1})^2}},$$

where y_t is the actual future value of the examined time series at point t, $\hat{y}_t$ is the forecast by the method under evaluation, n is the length of the training sample (number of historical observations), and h is the forecasting horizon (28 days). It should be noted that the denominator of RMSSE (in-sample, one-step-ahead mean squared error of the naive method) is computed only for the periods during which the examined products are actively sold, i.e., the periods following the first non-zero demand observed for the series under evaluation, because many of the products included in the data set started to be sold later than the first available date (Makridakis et al., 2022).

Similar to MASE, RMSSE is independent of the data scale and it has predictable behavior, i.e., it becomes infinite or undefined only when all the errors with the Naive method are equal to zero, as well as having a defined mean and finite variance, and it is symmetric in the sense that it penalizes equally positive and negative forecast errors. The choice of this particular measure can also be justified as follows.

- Many of the series in the competition are characterized by intermittency, involving sporadic unit sales with many zeros. Thus, absolute errors that are optimized for the median (Schwertman et al., 1990) would assign lower scores (better accuracy) to forecasting methods that derive forecasts close to zero. However, the objective of the competition is to accurately forecast the average sales. Thus, the accuracy measure employed is built on squared errors that are optimized for the mean (Kolassa, 2016).
- In contrast to other measures with similar statistical properties, such as relative errors and measures (Davydenko & Fildes, 2013), RMSSE can be safely computed for all M5 series because it does not rely on divisions with values that could be equal or close to zero. For example, this is typically the case with percentage errors when y_t is equal to zero or relative errors when the error of the benchmark used for scaling is zero.

After estimating RMSSE for all 42,840 time series in the competition (the average accuracy reported for each series across the complete forecasting horizon), the overall accuracy of the forecasting method was computed by averaging the RMSSE scores across all series in the data set using appropriate weights. This measure, called the weighted RMSSE (WRMSSE), is defined as follows:

$$WRMSSE = \sum_{i=1}^{42,840} w_i \times RMSSE_i, \quad (1)$$

where w_i is the weight and $RMSSE_i$ is the score of the ith series in the competition. Lower WRMSSE scores indicate more accurate forecasts. It should be noted that the estimation of WRMSSE differs from the approaches applied in previous M competitions. In the first three competitions, all of the errors were computed both per series and per forecasting horizon, and then equally averaged together. In M4, the errors were first averaged per series in exactly the same manner as M5, but then averaged again using equal weights.

The weights for the WRMSSE measure were computed based on the last 28 observations of the final training sample in the data set, specifically the cumulative actual dollar sales in each series in that particular period (sum of units sold multiplied by their respective price). Thus, some slow moving series could be assigned zero weights, i.e., they did not contribute to the estimation of the WRMSSE score. This was true for 883 cases, i.e., about 2% of the 42,840 series in the competition, but the realized dollar sales in these series in the testing period were minor, as they accounted for less than 1.3% of the total revenue. Moreover, as reported in Table 3 by Makridakis et al. (2022), the dollar sales computed at the state, store, product category, and product department levels did not change significantly between the "validation" and "test" phases of the competition, so the weights of the WRMSSE measure remained relatively constant for short periods of time and they were indicative of the value that each series represented.

We consider that the weighting scheme used in the M5 competition based on the unit sales of various products with different selling volumes and prices organized in a hierarchical manner is more appropriate for successfully identifying forecasting methods that add significant value to retail companies with an interest in accurately forecasting series that mostly translate to relatively higher revenues. In order for a forecasting method to be considered appropriate in a business context, it must provide accurate forecasts across all aggregation levels, especially for series with high importance, i.e., series that represent significant sales measured in monetary terms. Therefore, we expected the "best" performing forecasting methods to derive lower forecasting errors for the series with more value to a company.

It should be noted that according to WRMSSE, all aggregation levels are equally weighted. The reason is that, for instance, the total dollar sales of a product, measured across all three states, are equal to the sum of the dollar sales of this product when measured across all ten stores. Similarly, the total dollar sales of a store's product category are equal to the sum of the dollar sales of the departments that this category consists of, as well as the sum of the dollar sales of the corresponding departments' products. Moreover, given that M5 does not focus on a particular decision-making problem, there is no obvious reason for weighting the individual levels unequally.

An illustrative example of the computation of WRMSSE can be found in the Competitors' Guide to the competition, which is available on the M5 website.[2] The code used for estimating WRMSSE and the exact weight of each series can be found in the GitHub repository[3] for the competition.

[2] https://mofc.unic.ac.cy/m5-competition/

[3] https://github.com/Mcompetitions/M5-methods.

Table 1
Six prizes for the M5 "Accuracy" competition.

Prize name	Description	Amount
1st prize	Best-performing method according to WRMSSE	$25,000
2nd prize	Second best-performing method according to WRMSSE	$10,000
3rd prize	Third best-performing method according to WRMSSE	$5,000
4th prize	Fourth best-performing method according to WRMSSE	$3,000
5th prize	Fifth best-performing method according to WRMSSE	$2,000
Student prize	Best-performing method among student teams according to WRMSSE.	$5,000
Total		$50,000

2.3. Prizes

In order for a team to be eligible for a prize, point forecasts had to be provided for all 30,490 series in the competition's 12th aggregation level (product-store level), which were then aggregated (summed up) to produce forecasts for the remaining levels. Moreover, winning teams had to provide code for reproducing the forecasts that they originally submitted to the competition, as well as documentation to explain the forecasting method employed.

Similar to M4, objectivity and reproducibility were prerequisites for collecting any prize (Makridakis et al., 2018a), and thus except for companies that provide forecasting services and those claiming proprietary software, the winning teams had to upload their code onto the Kaggle platform no later than 14 days after the end of the competition (i.e., July 14, 2020). This material was later uploaded to the M5 public GitHub repository for individuals and companies interested in using the winning methods, and the teams that developed these methods to be credited. Companies that provide forecasting services and those claiming proprietary software had to give the organizers detailed descriptions of how they made their forecasts and a source or execution file for reproducing their forecasts.

After receiving the code and documentation from all the winning teams, the organizers evaluated the reproducibility of their results. ML algorithms typically involve random initializations, so the organizers considered any method as fully reproducible if it had a reproducibility rate, i.e., absolute percentage difference of WRMSSE between the original and reproduced forecasts, lower than 2%.[4] All of the winning methods were found to be fully reproducible, but if this was not the case, then the prizes would have been given to the next best-performing and fully reproducible submission.

The prizes for the M5 "Accuracy" competition are listed in Table 1. No restrictions prevented a team from collecting both a regular prize and a student[5] prize. Moreover, no restrictions prevented a team from collecting a prize in both the M5 "Accuracy" and M5 "Uncertainty" competitions. The awards were given during the virtual online M5 conference on October 29, 2020.

An amount of $40,000 was generously provided by Kaggle, which also waived the fees for hosting the M5 competition. In addition, Google and the Makridakis Open Forecasting Center (MOFC) generously provided $20,000 each, and in addition to the M5 data set, Walmart generously provided an amount of $10,000. Finally, the global transportation technology company Uber generously provided $5,000, while the International Institute of Forecasters (IIF) generously provided another $5,000. The total amount of $100,000 was equally distributed between the accuracy and uncertainty challenges in the M5 competition.

3. Participating teams and submissions

The M5 "Accuracy" competition involved 7092 participants in 5507 teams from 101 countries. Among these teams, 4373 (79.4%) entered the competition during the "validation" phase and 1134 (20.6%) during the "test" phase. Moreover, 1434 teams (26.0%) made submissions during both the "validation" and "test" phases of the competition, but 2939 (53.4%) only during the "validation" phase. In total, the participating teams made 88,136 submissions, most of which (78.3%) were submitted during the "validation" phase. Most of the teams made a single submission, but some made between three and 20 submissions. It is worth mentioning that 1563 participants participated in a Kaggle competition for the first time, including 15 in the top 100.

Unfortunately, due to privacy regulations, no information was made available about the background (academic, research, business, or other) of the participating teams, their experience and skills, and the type of methods utilized (i.e., statistical, ML, combination, hybrid, or other), except for the winning teams and a few others who were willing to share this information with the organizers. However, based on the general characteristics of the Kaggle community, we assumed that most of the teams had an adequate background in statistics and computer science, and they were also familiar with ML forecasting methods, such as neural networks (NNs) and regression trees. This was a fair assumption considering that Kaggle participants self-select for greater data science knowledge than non-participants, and many of them are highly experienced in competitions that require the development of accurate ML methods, including forecasting methods, and they are considered "grandmasters", "masters", or "experts"[6] among the community.

[4] Provided that the winning teams had set the seeds used for generating random numbers in their code, it should have been possible in principle to reproduce their results with a reproducibility rate of 0%. Unfortunately, this was not the case.

[5] In a student team, at least half of the team members were current full-time students. Teams that were eligible for the student prize had a name followed by "_STU".

[6] https://www.kaggle.com/progression

Among the participating teams, 2666 (48.4%) managed to outperform the Naive benchmark, 1972 (35.8%) outperformed the sNaive (a naive method that accounts for seasonality) benchmark, and 415 (7.5%) beat the top performing benchmark (ES_bu; details of the benchmarks used in the M5 "Accuracy" competition and their performance are provided in the appendix of the supplementary material). It is important to note that these numbers refer to the forecasts selected by each team for the final evaluation of their performance and not to the "best" submission made by each case while the competition was still running. In the latter case, 3510 (63.7%), 2685 (48.8%), and 672 (12.2%) teams would have managed to outperform the Naive, sNaive, and ES_bu benchmarks, respectively. Thus, many teams failed to choose the best method that they developed, probably due to misleading validation scores.

The fact that about 92.5% of the participating teams failed to beat ES_bu should not be overlooked. ES_bu can be considered a relatively simple benchmark that is easy to implement and computationally cheap. This method utilizes an algorithm that automatically selects the most appropriate model from the exponential smoothing family of models for each series at the product-store level and the aggregation of the forecasts produced for these series in a bottom-up manner for forecasting the series at higher levels. Exponential smoothing models have been widely used for time series forecasting since the early 1960s (Gardner Jr., 1985) and although several improvements have been made in their estimation since, their principles have remained mostly unchanged. Therefore, as discussed later, the winning teams who utilized more sophisticated ML methods managed to outperform the benchmarks in the competition by a notable margin, but this does not mean that the success of ML methods in general can be taken for granted. When working with ML methods, in addition to skills, a significant amount of time is required for understanding and cleaning the data, engineering features, and defining the architecture and hyper-parameters for the selected method, among other issues. By contrast, exponential smoothing is an "off-the-shelf" forecasting method that is readily available in most types of statistical software for anyone to use, even inexperienced users. Therefore, it is important to consider whether it is necessary for retailers to focus on new, innovative methods without ensuring that their accuracy improvements are substantial compared with simple methods given the extra expenses incurred by skillful data scientists and powerful infrastructure. Clearly, this finding may have been biased to some extent given that many of the teams that failed to outperform ES_bu could have lost interest in the competition or decided to have no further involvement in M5, thereby not improving their initial submissions. Nevertheless, this did not influence the fact that the standard time series forecasting methods were competitive and difficult benchmarks to beat.

Fig. 1 summarizes this information, including the daily number of submissions and cumulative number of participating teams, number of participants per country, distribution of the accuracy for the teams that performed better than the Naive benchmark, the accuracy of the teams that performed better than the top performing benchmark, and their respective ranks.

According to Fig. 1, we can conclude the following.

- The majority of the teams made most of their submissions during the "validation" phase when the public leaderboard was available and live feedback was received. During the "test" phase, most of the teams probably used their own private cross-validation (CV) strategies to fine tune their methods, which were mainly submitted four days before the competition ended.
- The majority of the participants originated from the USA (17%), Japan (17%), India (10%), China (10%), and Russia (6%), and the remaining 40% from 96 other countries. Thus, we conclude that a large, active community is interested in forecasting in both developed and developing countries.
- Only a limited number of teams managed to outperform the top performing benchmark in the competition, and the majority of the teams were outperformed by more than 13% by the top one, ES_bu.
- Among the 415 teams that managed to outperform all of the benchmarks in the competition, five obtained improvements greater than 20%, 42 greater than 15%, 106 greater than 10%, and 249 greater than 5%. These improvements are substantial and they demonstrate the superior performance of the winning M5 methods compared with the standard forecasting benchmarks. Moreover, the five winners of the competition were the only teams to obtain accuracy improvements greater than 20%, thereby achieving a clear victory.

The various tables presented in the following sections are focused on the top 50 performing teams in the competition, as well as the benchmarks considered by the organizers because of the following two reasons. First, it would have been practically impossible to analyze and report in detail the results for all of the teams that participated in the competition. Second, very few teams were willing to share detailed information about the methods utilized, so we considered that there was more to learn from the top performers who provided detailed information. Furthermore, given the general complexity of the data and competition, we considered that it was more useful to draw conclusions from methods that worked well rather than rationalizing why some methods performed poorly.

4. Results, winning submissions, and key findings

4.1. Results

Table 2 presents the accuracy (WRMSSE) achieved by the top 50 teams in the competition in terms of both the overall accuracy and across the 12 aggregation levels. The last column in the table shows the overall (42,840 series) percentage improvement for each team compared with the top performing benchmark (ES_bu), the performance of which is displayed at the bottom of the table.

Table 2 shows that all of the top 50 submissions outperformed the overall forecasting accuracy of the top

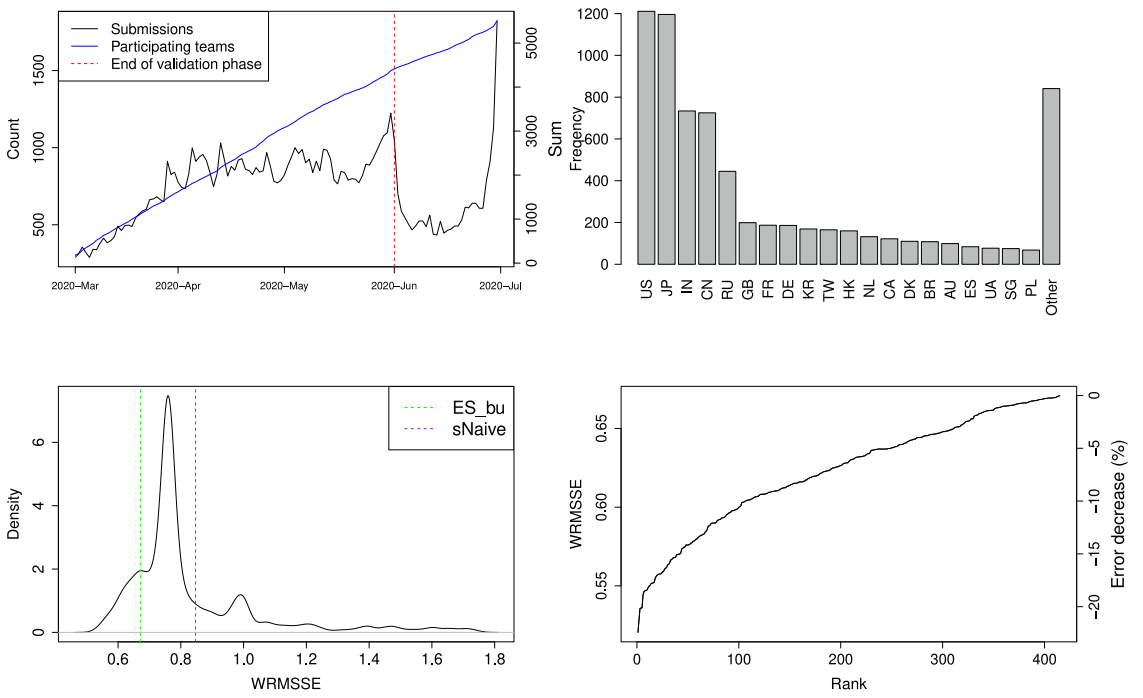

Fig. 1. Summary of the participating teams and submissions. Top left: daily number of submissions (black line, measured based on the left vertical axis) and cumulative number of participating teams (blue line, measured based on the right vertical axis). The red dotted line indicates the end of the "validation" phase. Top right: number of participants per country (top 20 in terms of participation) estimated based on their IP address. Bottom left: distribution of the accuracy (WRMSSE) achieved by teams that performed better than the Naive benchmark. The green dotted line indicates the accuracy of the ES_bu benchmark and the purple dotted line denotes the accuracy of sNaive. Bottom right: accuracy (WRMSSE) and ranks of the teams that performed better than the top performing benchmark (ES_bu). Percentage improvements over ES_bu are also reported. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performing benchmark by more than 14%, and the improvements were higher than 20% for the top five performing teams as well as an impressive 22.4% for the winning team. Considering that the improvements of the winning submissions in the M3 and M4 competitions compared with the corresponding benchmarks were less than 10% (Makridakis et al., 2020c), we can conclude that M5 included more accurate approaches that reduced the error compared with the most accurate benchmark by more than one-fifth. Thus, retail and logistic companies could benefit from utilizing these innovative forecasting approaches in their forecasting practice.

It was also interesting that the winning team (*YJ_STU*) did not obtain the most accurate forecasts across all 12 aggregation levels, where their approach was only the best at levels 3, 7, 8, and 9, and the second best at levels 2 and 6. In particular, this applied to the lowest three aggregation levels of the data set (10, 11, and 12), where among the 50 submissions, *YJ_STU* was ranked 13rd, 12th, and 11th, respectively. The same applied to the runner-up (*Matthias*), which was ranked 1st at levels 2, 4, 5, and 6, but this team obtained almost the worst performance among the 50 methods examined at levels 10, 11, and 12, where they were ranked 48rd, 49th, and 48th, respectively. *Daniela A* ranked 28th overall with the best performance at level 1, *mf* ranked 3rd had the best performance at levels 10 and 11, and *wyzJack_STU* ranked 6th had the best performance at level 12, which contained the vast majority of the series that required forecasting. These results suggest that different forecasting methods may be more appropriate depending on the aggregation level, and that there may be indeed "horses for courses" (Petropoulos et al., 2014). Nevertheless, further analysis would be needed to conclude whether the differences reported for the methods across the various aggregation levels are significant or random.

We also found that the accuracy of the top performing methods deteriorated at a lower aggregation level because the uncertainty increased when forecasting more disaggregated data with volatile sales, and patterns such as trend and seasonality are difficult to capture (Kourentzes, Petropoulos et al., 2014). This finding is clearly visualized in Fig. 2, which presents the distribution of WRMSSE for the 50 top performing teams per aggregation level and the accuracy of ES_bu. The top benchmark was outperformed by all teams at levels 1 to 9, but the improvements reported for the remaining levels were less significant, where some teams even performed worse than the benchmark. For example, the average improvement by the methods compared with the benchmark was 40% at level 1, but the average improvement decreased to about 23% at levels 5, 6, and 7, and 3% at levels 10, 11, and 12. Therefore, we can conclude that the average improvements by the top performing methods were mainly confined to the top and middle parts of the hierarchies, and they were rather limited in terms of WRMSSE at the product, product-state, and product-store levels.

Table 2
Performance of the top 50 teams in the M5 "Accuracy" competition in terms of WRMSSE. The results are presented at both the aggregation level (described in detail in Makridakis et al. (2022)) and overall. Overall percentage improvements are also shown compared with the top performing benchmark (ES_bu). Column-wise minimum values are displayed in bold.

Rank	Team	Aggregation level												Average	Improvement over ES_bu (%)
		1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store		
1	YJ_STU	0.199	0.310	**0.400**	0.277	0.365	0.390	**0.474**	**0.480**	**0.573**	0.966	0.929	0.884	**0.520**	22.4
2	Matthias	0.186	**0.294**	0.416	**0.246**	**0.349**	**0.381**	0.481	0.497	0.594	1.023	0.964	0.907	0.528	21.3
3	mf	0.236	0.319	0.421	0.308	0.397	0.405	0.496	0.505	0.600	**0.950**	**0.917**	0.875	0.536	20.2
4	monsaraida	0.254	0.340	0.418	0.302	0.377	0.411	0.483	0.490	0.579	0.963	0.928	0.886	0.536	20.1
5	Alan_Lahoud	0.213	0.324	0.414	0.272	0.361	0.416	0.494	0.503	0.595	0.995	0.950	0.897	0.536	20.1
6	wyzJack_STU	0.248	0.367	0.431	0.319	0.396	0.436	0.502	0.502	0.584	0.953	0.918	**0.875**	0.544	18.9
7	RandomLearner	0.194	0.317	0.423	0.276	0.404	0.408	0.516	0.503	0.608	1.029	0.968	0.910	0.546	18.6
8	SHJ	0.279	0.357	0.419	0.336	0.406	0.429	0.498	0.497	0.586	0.956	0.922	0.878	0.547	18.5
9	gest_2	0.197	0.322	0.424	0.269	0.406	0.420	0.536	0.513	0.624	1.000	0.953	0.901	0.547	18.5
10	DenisKokosinskiy_STU	0.294	0.363	0.419	0.341	0.401	0.429	0.493	0.494	0.581	0.955	0.921	0.878	0.547	18.4
11	XueWang	0.288	0.358	0.417	0.348	0.423	0.424	0.501	0.490	0.582	0.959	0.921	0.876	0.549	18.2
12	yq_STU	0.226	0.320	0.456	0.294	0.403	0.399	0.496	0.526	0.614	1.010	0.954	0.899	0.550	18.1
13	PoHaoChou	0.212	0.317	0.459	0.322	0.402	0.413	0.504	0.539	0.630	0.968	0.940	0.898	0.550	18.0
14	Tsuru	0.257	0.335	0.402	0.325	0.416	0.421	0.506	0.503	0.608	0.994	0.951	0.900	0.552	17.8
15	bk_18	0.217	0.333	0.420	0.303	0.433	0.431	0.537	0.510	0.615	0.986	0.943	0.893	0.552	17.8
16	N60610	0.195	0.350	0.436	0.298	0.409	0.441	0.530	0.521	0.619	0.976	0.945	0.900	0.552	17.8
17	MonashSL_STU	0.247	0.342	0.446	0.308	0.404	0.412	0.501	0.520	0.622	0.992	0.944	0.892	0.552	17.7
18	leoclement	0.270	0.354	0.410	0.322	0.415	0.434	0.526	0.498	0.603	0.986	0.945	0.896	0.555	17.3
19	minghui_Tju	0.254	0.365	0.428	0.327	0.425	0.439	0.529	0.505	0.602	0.979	0.936	0.886	0.556	17.1
20	zfc613	0.236	0.343	0.470	0.291	0.387	0.412	0.506	0.539	0.627	1.008	0.959	0.907	0.557	17.0
21	Nodalpoints	0.312	0.376	0.423	0.353	0.416	0.443	0.508	0.500	0.587	0.963	0.925	0.880	0.557	17.0
22	CPUkiller	0.263	0.367	0.427	0.333	0.431	0.438	0.529	0.504	0.601	0.979	0.935	0.885	0.558	16.9
23	dont_overfit	0.263	0.367	0.427	0.333	0.431	0.438	0.529	0.504	0.601	0.979	0.935	0.885	0.558	16.9
24	Dan_Hargreaves	0.269	0.350	0.439	0.316	0.405	0.431	0.517	0.522	0.616	0.984	0.946	0.900	0.558	16.9
25	M0T0_STU	0.252	0.346	0.425	0.345	0.431	0.445	0.530	0.526	0.623	0.967	0.932	0.886	0.559	16.7
26	Genryu	0.295	0.368	0.436	0.340	0.410	0.445	0.515	0.523	0.611	0.961	0.924	0.880	0.559	16.7
27	Moscow_Five	0.245	0.349	0.442	0.309	0.435	0.436	0.542	0.526	0.624	0.988	0.941	0.889	0.560	16.5
28	Daniela_A	**0.162**	0.333	0.483	0.278	0.441	0.412	0.569	0.558	0.679	0.988	0.942	0.892	0.561	16.3
29	shuheioka	0.267	0.354	0.440	0.313	0.419	0.430	0.524	0.517	0.616	1.002	0.954	0.903	0.562	16.3
30	sk_2	0.191	0.381	0.511	0.263	0.364	0.470	0.552	0.585	0.661	0.962	0.932	0.887	0.563	16.1
31	nagao	0.279	0.382	0.443	0.328	0.413	0.456	0.531	0.523	0.609	0.975	0.936	0.889	0.564	16.0
32	AjayNagar	0.221	0.324	0.518	0.285	0.412	0.400	0.515	0.573	0.660	1.010	0.954	0.900	0.564	15.9
33	cjwh	0.248	0.348	0.449	0.314	0.420	0.442	0.544	0.535	0.639	1.002	0.955	0.903	0.566	15.6
34	CWD75	0.237	0.326	0.422	0.330	0.452	0.442	0.551	0.526	0.637	1.004	0.960	0.912	0.567	15.6
35	Groot	0.278	0.384	0.443	0.342	0.432	0.458	0.540	0.519	0.611	0.979	0.937	0.887	0.567	15.4
36	Astral	0.299	0.381	0.453	0.342	0.401	0.453	0.520	0.528	0.611	0.984	0.945	0.896	0.568	15.4
37	Logistic	0.278	0.386	0.445	0.344	0.436	0.457	0.541	0.518	0.610	0.979	0.936	0.886	0.568	15.3
38	jdsc_perceiving_team	0.262	0.372	0.461	0.326	0.433	0.445	0.532	0.531	0.623	0.990	0.948	0.897	0.568	15.3
39	Abzal	0.314	0.373	0.434	0.351	0.420	0.447	0.519	0.515	0.603	0.998	0.956	0.906	0.570	15.1
40	Pianus	0.287	0.383	0.473	0.342	0.435	0.451	0.535	0.536	0.626	0.964	0.926	0.880	0.570	15.1
41	NAU	0.277	0.366	0.456	0.310	0.425	0.440	0.537	0.532	0.633	1.002	0.957	0.906	0.570	15.0
42	shirokane_friends	0.300	0.387	0.454	0.347	0.429	0.461	0.540	0.534	0.619	0.965	0.926	0.880	0.570	15.0
43	Alexnet	0.301	0.390	0.444	0.353	0.435	0.463	0.540	0.520	0.610	0.975	0.934	0.885	0.571	14.9
44	Griffin_Series	0.317	0.380	0.469	0.361	0.442	0.448	0.527	0.529	0.618	0.971	0.933	0.887	0.574	14.5
45	Hiromitsu_Kigure	0.291	0.380	0.462	0.342	0.428	0.449	0.533	0.535	0.629	0.991	0.950	0.895	0.574	14.5
46	YK	0.247	0.369	0.464	0.314	0.438	0.453	0.551	0.542	0.644	1.011	0.958	0.904	0.575	14.4
47	PASSTA	0.339	0.396	0.460	0.366	0.421	0.457	0.521	0.532	0.614	0.970	0.933	0.886	0.575	14.4
48	golubyatniks	0.359	0.413	0.455	0.387	0.434	0.466	0.519	0.521	0.600	0.956	0.922	0.879	0.576	14.2
49	belkasanek	0.184	0.329	0.538	0.260	0.427	0.416	0.549	0.608	0.701	1.028	0.964	0.905	0.576	14.2
50	Random_prediction	0.249	0.348	0.455	0.347	0.457	0.460	0.563	0.558	0.655	0.986	0.943	0.890	0.576	14.2
416	ES_bu - Benchmark	0.426	0.514	0.580	0.478	0.557	0.577	0.654	0.643	0.728	1.012	0.969	0.915	0.671	–

In order to further investigate the differences between the top 50 submissions and the top performing benchmark, we employed the multiple comparisons with the best (MCB) test (Koning et al., 2005). The MCB test was used to compute the average ranks of the forecasting methods according to RMSSE across the complete data set in the competition and to assess whether they were statistically different. Fig. 3 presents the results of the analysis. No overlapping of the intervals for two methods indicated a statistically significant difference in performance. Thus, methods that did not overlap with the gray interval in the figures were considered significantly worse than the best, and vice versa. We consider that the MCB test results are indicative and useful for investigating the relative performance of the submitted methods, but it should be noted that this test assumes that the forecasts under comparison are independent. The forecasts we used for conducting the MCB test were related to grouped time series, where the series at the bottom level structured those at the higher levels, so this assumption did not strictly hold, and thus our conclusions may not be entirely valid.

Clearly, *SHJ* (ranked 8th) obtained significantly better forecasts than the other methods examined, although *DenisKokosinskiy_STU* (ranked 10th) and *XueWang* (ranked 11th) did not differ significantly from *SHJ* and they provided similarly accurate forecasts for the majority of the series. None of the five winning teams performed equally as well as *SHJ*, but *wyzJack_STU* was ranked 1st at level 12 according to WRMSSE and they also performed significantly worse overall. Thus, we conclude that the winning teams developed methods that focused mostly on expensive and fast-moving products that minimized WRMSSE, thereby providing less accurate results for the remainder of the series, which probably offered less value to the company. This finding demonstrates that the objective of the competition (minimizing the error measure across all aggregation levels and especially for high-valued series) represented by the error measure employed was critical for determining the winning submissions and optimizing their parameters. Consequently, we found that at the weighting settings applied, the "best" forecasts depended on the accuracy measure used (Kolassa, 2020), especially for flexible ML methods where the loss function can be adjusted accordingly to optimize forecasts based on the selected measure.

Finally, we investigated the impact of the length of the forecasting horizon on the accuracy obtained by the top

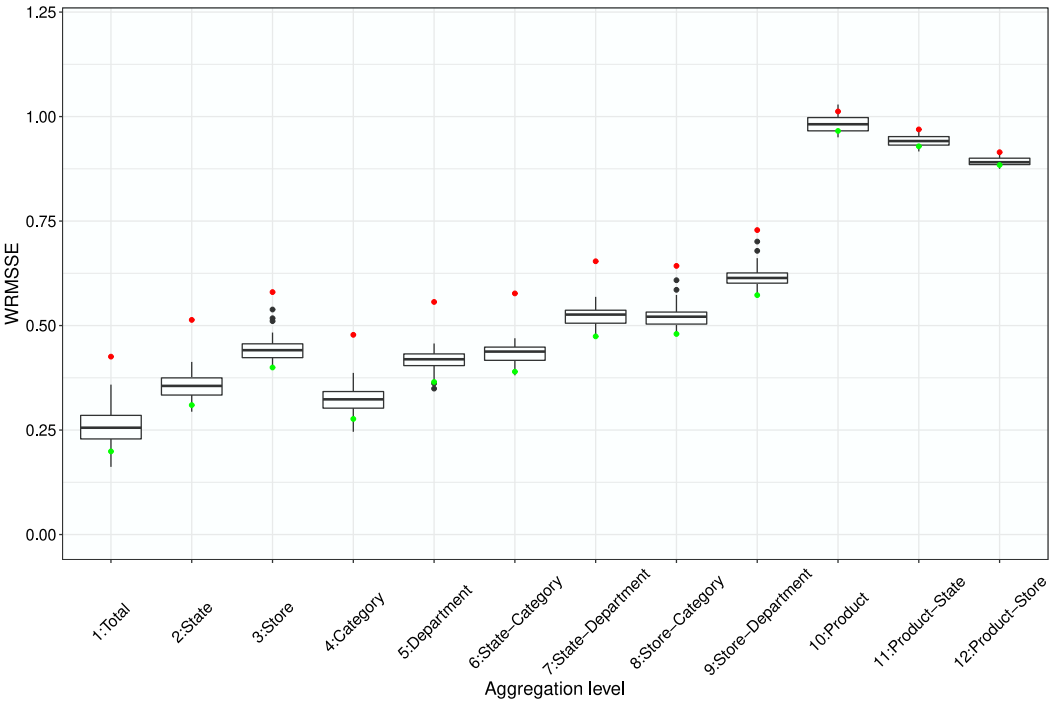

Fig. 2. Forecasting accuracy (WRMSSE) of the top 50 performing teams in the M5 "Accuracy" competition. The results are shown per aggregation level and box-plots display the distribution of the average errors recorded for the methods examined (minimum value, 1st quartile, median, 3rd quartile, maximum value, and outliers denoted by black dots). The red dots indicate the performance of the top performing benchmark in the competition (ES_bu) and the green dots show the performance of the winning team (*YJ_STU*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

50 performing methods in the competition. First, we computed the weighted root squared scaled error (WRSSE) for these methods over each forecasting horizon and series separately, and then aggregated the results per aggregation level and horizon. A summary of the results is presented in Fig. 4. The accuracy remained rather constant in most of the cross-sectional levels and it was even slightly reduced in some cases, but this was not true for the lowest aggregation levels (10, 11, and 12) where the accuracy deteriorated significantly as the forecasting horizon increased. This finding was closely related to the characteristics of the series at each level. At higher levels, trend and seasonality dominated randomness, which did not significantly affect the forecasting accuracy, at least for the relatively short forecasting horizon of 28 days considered in the competition. By contrast, at lower aggregation levels, intermittency, erratic behavior, and a lack of trend and seasonality increased the randomness and negatively affected the forecasting accuracy. In addition, at many aggregation levels, and especially the lowest ones, the errors exhibited some form of periodicity, e.g., larger errors were observed during the weekends, thereby indicating that part of the seasonality present in the data was not appropriately captured by the forecasting methods, even the top ranked ones.

4.2. Winning submissions

Unfortunately, as mentioned previously, only a very limited number of teams that participated in the M5 "Accuracy" competition were willing to share description of their methods with the organizers and the Kaggle community, and even fewer shared their code. The organizers tried to reach at least the top 50 performing teams in the competition by email (a template for describing the main features of the utilized methods was provided), but this information was obtained from only 17 of them by receiving a direct reply, or by observing public discussions and the notebooks posted by these teams on Kaggle. Nevertheless, we still consider that many lessons can be learned from these methods because they all provided more accurate forecasts than the benchmarks considered and the thousands of other participating teams.

Before presenting the five winning methods, we note that most of the methods utilized LightGBM,[7] which is a ML algorithm for performing nonlinear regression using gradient boosted trees (Ke et al., 2017). LightGBM has several advantages compared with other ML alternatives in forecasting tasks, such as those in the M5 "Accuracy" competition, because it allows the effective handling of multiple features (e.g., past sales and exogenous/explanatory variables) of various types (numeric, binary, and categorical). In addition, it is fast to compute compared with other gradient boosting methods (GBMs), does not depend on data pre-processing and transformations, and only requires the optimization of a relatively small number of parameters (e.g., learning rate, number of

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

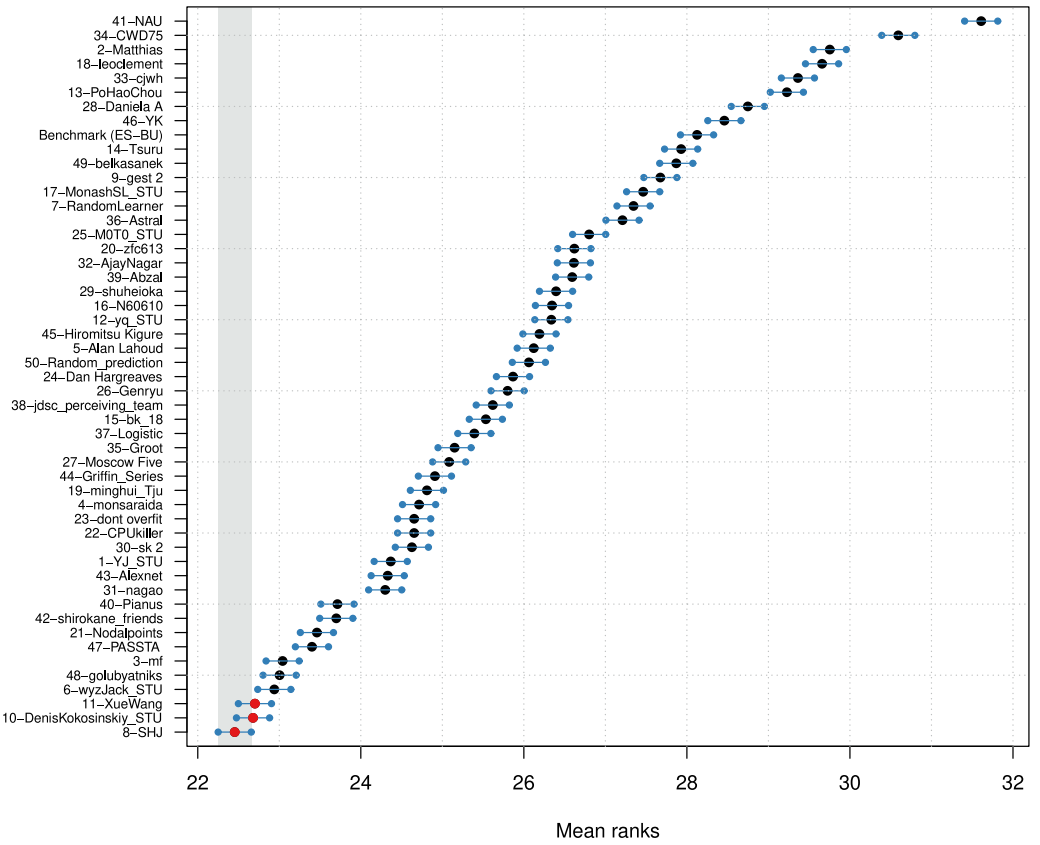

Fig. 3. Average ranks and 95% confidence intervals for the top 50 performing teams in the M5 "Accuracy" competition, and the top performing benchmark (ES_bu) over all series based on the multiple comparisons with the best test (RMSSE used for ranking the methods) proposed by Koning et al. (2005). The overall ranks of the teams in terms of WRMSSE are shown to the left of their names.

iterations, maximum number of bins, number of estimators, and loss functions). In this regard, LightGBM is highly convenient for experimenting and developing solutions that can be accurately generalized to a large number of series with cross-correlations. In fact, LightGBM can be considered the standard method of choice in Kaggle's recent forecasting competitions because the winners of the "Corporación Favorita Grocery Sales Forecasting" and "Recruit Restaurant Visitor Forecasting" competitions built their approaches using this method (Bojer & Meldgaard, 2021), and the discussions and notebooks posted on Kaggle for the M5 "Accuracy" competition focused on using LightGBM and its variants.

The forecasting methods used by the five winning teams can be summarized as follows.

- **First place (*YJ_STU*; YeonJun In):** The winner of the competition was a senior undergraduate student at Kyung Hee University, South Korea, who used an equal weighted combination (arithmetic mean) of various LightGBM models, which were trained to produce forecasts for the product-store series using data pooled per store (10 models), store-category (30 models), and store-department (70 models). Two variations were considered for each type of model, where the first applied a recursive approach and the second a non-recursive forecasting approach (Bontempi et al., 2013). In total, 220 models were built and each series was forecast using the average of six models, where each exploited a different learning approach and training set. The models were optimized without considering early stopping and by maximizing the negative log-likelihood of the Tweedie distribution (Zhou et al., 2020), which is considered an effective approach when handling data with a probability mass of zero and non-negative, highly right-skewed distribution. The method was fine tuned using the last four 28 day-long windows of available data for CV and by measuring both the mean and the standard deviation of the errors produced by the individual models and their combinations. In this manner, the final solution was selected such that it provided both accurate and robust forecasts. Among the features used, the models considered various identifiers, calendar-related information, special days, promotions, prices, and unit sales data in both recursive and non-recursive formats.
- **Second place (*Matthias*; Matthias Anderer):** This method was also based on an equally weighted combination of various LightGBM models, but it was externally adjusted through multipliers according to

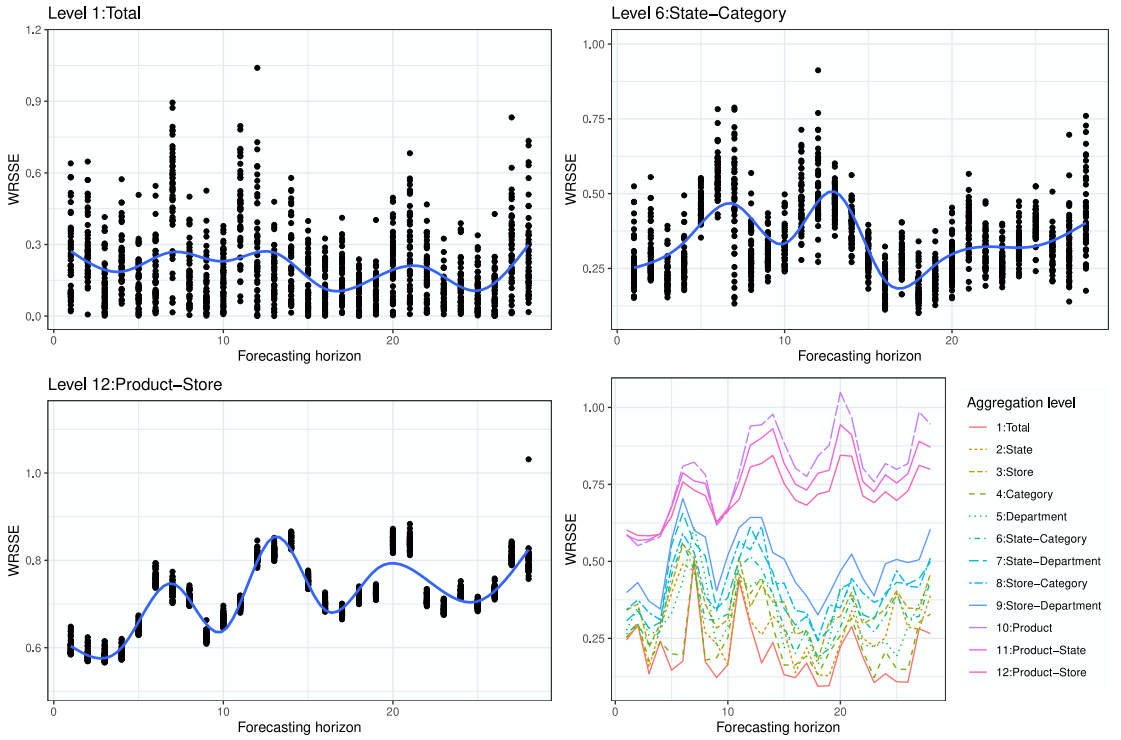

Fig. 4. Impact of the length of the forecasting horizon on the forecasting accuracy. Top left: forecasting accuracy (WRSSE) of the top 50 performing methods in the competition over each forecasting horizon for the top level of the data set. The blue line represents locally estimated scatter plot smoothing. Top right: this is similar to the figure at the top left but the results are shown for the middle aggregation level in the data set (state-category). Bottom left: this is similar to the figure at the top left but the results are shown for the lowest aggregation level in the data set (product-store). Bottom right: average error for all top 50 best-performing methods across all 12 aggregation levels and forecasting horizons. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the forecasts produced by N-BEATS (deep-learning NN for time series forecasting; Oreshkin et al., 2019) for the top five aggregation levels of the data set. Essentially, LightGBM models were first trained per store (10 models) and then five different multipliers were used to adjust their forecasts and to correctly capture the trend. In total, 50 models were built and each series of the product-store level in the data set was forecast using a combination of five different models. A custom, asymmetric loss function was used. The last four 28 day-long windows of available data were used for CV and model building. The LightGBM models were trained using only some basic features of calendar effects and prices (past unit sales were not considered), and the N-BEATS model was based solely on historical unit sales.

- **Third place (*mf*; Yunho Jeon & Sihyeon Seong):** This method used an equally weighted combination of 43 deep-learning NNs (Salinas et al., 2020), where each comprised multiple long short-term memory layers, which were employed to recursively predict the product-store series. Among the models trained, 24 considered dropout, whereas the other 19 did not. These models originated from only 12 models and they corresponded to the last, more accurate instances observed for these models during training, as specified by CV (last fourteen 28 day-long windows of available data). Similar to the winner, the method considered Tweedie regression, but it was modified to optimize weights based on the sampled predictions instead of actual values. The Adam optimizer and cosine annealing were used for learning rate scheduling. The NNs considered 100 features with similar characteristics to those used by the winning submission (sales data, calendar-related information, prices, promotions, special days, identifiers, and zero-sales periods).
- **Fourth place (*monsaraida*; Masanori Miyahara):** This method produced forecasts for the product-store series in the data set using non-recursive LightGBM models trained per store (10 models). However, in contrast to the other methods, each week in the forecasting horizon was forecast separately using a different model (four models per store). Thus, 40 models were built to produce the forecasts. The features used as inputs were similar to those applied by the winning submission, except for the recursive features. Tweedie regression was applied for training the models with no early stopping, and the training parameters were not optimized. The last five 28 day-long windows of available data were used for CV.
- **Fifth place (*Alan Lahoud*; Alan Lahoud):** This method used recursive LightGBM models, which were trained per department (seven models). After producing the forecasts for the product-store series,

they were externally adjusted such that the mean of each of the series at the store-department level was the same as that for the previous 28 days, which was achieved using appropriate multipliers. The models were trained using Poisson regression with early stopping and validated using a random sample of 500 days. The features used as inputs were similar to those employed by the winning submission.

Among the other top 50 best-performing methods with an available method description, it should be noted that most of them adopted similar approaches to the winning submission by training recursive and non-recursive LightGBM models per store, department, or store-department. The main exceptions were *N60610* ranked 16th who predicted the product-store series of the data set using both LightGBM and a Kalman filter and selected the most appropriate approach per series, *MonashSL_STU* ranked 17th who used an equal-weighted combination of LightGBM and a pooled regression model, *Nodalpoints* ranked 21st who employed a weighted combination of LightGBM and NNs trained across all series or per store, and *Astral* ranked 36th who considered a non-recursive Prophet-like model that mixed classical statistics practices with nonlinear optimization ML techniques, i.e., XGBoost and LightGBM.

4.3. Key findings

The main findings related to the performance of the top five methods are summarized as follows.

Finding 1: Superior performance of ML methods. Over many years, empirical studies have demonstrated that simple methods are as accurate as complex or statistically sophisticated methods (Makridakis et al., 2020c). Limited data availability, inefficient algorithms, the need for preprocessing, and restricted computational power are just some of the factors that reduce the accuracy of ML methods compared with statistical methods (Makridakis et al., 2018b). M4 was the first forecasting competition to show that two ML-based approaches were significantly more accurate than simple statistical methods, thereby highlighting the potential value of ML methods for obtaining more accurate forecasts (Makridakis et al., 2020c). The first method that won the M4 competition was a hybrid approach based on mixed recurrent NNs and exponential smoothing (Smyl, 2020), and the second ranked method used XGBoost to optimally weight the forecasts produced by standard time series forecasts (Montero-Manso et al., 2020). Both of the winning M4 submissions were based on ML but they were built on statistical, series-specific functionalities, and their accuracy was also similar to a simple combination of the median of four statistical methods (Petropoulos & Svetunkov, 2020). Therefore, M5 was the first competition where all of the top-performing methods were both "pure" ML approaches and better than all statistical benchmarks and their combinations. It was shown that LightGBM can be used effectively to process numerous correlated series and exogenous/explanatory variables, and to reduce the forecast errors. Moreover, deep learning methods like DeepAR and N-BEATS, using advanced, state-of-the-art ML implementations, have shown forecasting potential, motivating further research in this direction.

Finding 2: Value of combining. The M5 "Accuracy" competition confirmed the findings of the previous four M competitions as well as those of numerous other studies by demonstrating that the accuracy can be improved by combining forecasts obtained with different methods, even relatively simple ones (Petropoulos & Svetunkov, 2020). The winner of the M5 "Accuracy" competition employed a very simple, equal-weighted combination, involving six models, where each exploited a different learning approach and training set. Similarly, the runner-up utilized an equal-weighted combination of five models, where each obtained a different estimate of the trend, and the third best-performing method used an equal-weighted combination of 43 NNs. Simple combinations of models were also used by the methods ranked 14th, 17th, 21st, 24th, 25th, and 44th. Among these combination approaches, only that ranked 25th considered unequal weighting of the individual methods. The value of combining was also supported by comparisons made between the benchmarks in the competition. As shown in the appendix of the supplementary material, the combination of exponential smoothing and AutoRegressive Integrated Moving Average (ARIMA) models performed better than the individual methods, while a combination of top-down and bottom-up reconciliation methods outperformed both the top-down and bottom-up methods. Therefore, our results support the long-standing belief that combining forecasts obtained with different methods can improve the forecasting accuracy and they confirm that there is no guarantee that an "optimal" forecast combination will perform better than a simpler, equal-weighted one (Claeskens et al., 2016).

Finding 3: Value of "cross-learning." In the previous M competitions, most of the series were uncorrelated, with a different frequency and domain, and chronologically unaligned. Therefore, both of the top-performing M4 submissions utilized "cross-learning" from multiple series concurrently instead of one series at a time, but their approach was difficult to implement effectively in practice, and it did not demonstrate the full potential of "cross-learning". By contrast, the M5 comprised aligned, highly-correlated series structured in a hierarchical fashion, so "cross-learning" was much easier to apply and superior results were achieved compared with methods trained in a series-by-series manner. It should be note that in addition to producing more accurate forecasts, "cross-learning" implies the use of a single model instead of multiple models, where each is trained using data from a different series, thereby reducing the overall computational cost and mitigating difficulties related to limited historical observations (Semenoglou et al., 2021). Essentially, all top 50-performing methods in M5 utilized "cross-learning" by exploiting all of the information in the data set.

Finding 4: Notable differences between the winning methods and benchmarks used for sales forecasting.

The M5 "Accuracy" competition considered 24 benchmarks of various types that are typically used in sales forecasting applications, including traditional and state-of-the-art statistical methods, ML methods, and combinations. As shown in Fig. 3 and Table 2, the winning submissions provided more accurate forecasts in terms of ranks compared with these benchmarks and they were also more than 20% better in terms of the average WRMSSE. The differences were much smaller at lower aggregation levels and in some cases negative, but the results clearly demonstrated their overall superiority, thereby motivating additional research into the area of ML forecasting methods that can be used to predict complex, nonlinear relationships between series, as well as including exogenous/explanatory variables. However, it should be noted that this finding is based on the performance of the winning teams alone. When the whole sample of participating teams was considered, we found that the vast majority (about 92.5%) failed to outperform the top performing benchmark, despite the latter being considerably simpler. This finding suggests that standard time series forecasting methods, such as exponential smoothing, may still be useful for supporting decisions related to the operation of retail companies and that the utilization of ML methods does not necessarily guarantee better performance, at least if the methods employed are not built and trained correctly, which was the case for the M5 winning teams. Similarly, we found that it was possible to utilize more sophisticated methods to improve the forecasting accuracy at a particular cross-sectional level, but the impact was minor at other levels, especially the most granular levels. Therefore, the adoption of more sophisticated methods should be carefully assessed by investigating whether the value added by these approaches in terms of accuracy is meaningful compared with their costs (Gilliland, 2020).

Finding 5: Beneficial effects of external adjustments. Forecast adjustments are typically used when forecasters exploit external information as well as inside knowledge and their expertise to improve forecasting accuracy (Davydenko & Fildes, 2013). Such adjustments were applied in the M2 competition and it was found that they did not improve the accuracy of pure statistical methods (Makridakis et al., 1993). In the M5 "Accuracy" competition, some of the top-performing methods including those ranked 2nd and 5th utilized adjustments in the form of multipliers to enhance the forecasts derived by the ML models. These adjustments were not completely based on judgment but instead on the analytical alignment of the forecasts produced at the lowest aggregation levels with those at the higher levels, and these adjustments proved to be beneficial, where they helped the models to reduce bias and better consider the longer-term trends that are easier to observe at higher aggregation levels (Kourentzes, Petropoulos et al., 2014). The concept of reconciling the forecasts produced at different aggregation levels is not new in the field of forecasting, and numerous studies have empirically demonstrated its benefits, especially when forecasts and information from the complete hierarchy are exploited (Hyndman et al., 2011; Spiliotis et al., 2020b). Therefore, further investigation is required to evaluate the actual value of the external adjustments used in M5 to determine how they should be preferably selected in order to improve the accuracy in a more consistent and unbiased manner. Several studies have shown that judgmental adjustments are often unnecessary and they can degrade the forecasting accuracy (Lawrence et al., 2006). Fildes et al. (2009) analyzed the forecasts produced by four supply-chain companies, including a retailer, and found that small positive adjustments generally reduced the accuracy, thereby suggesting a general bias toward optimism, whereas larger negative adjustments were more likely to be beneficial.

Finding 6: Value added by effective CV strategies. When dealing with complex forecasting tasks, adopting effective CV strategies is critical for objectively capturing the post-sample accuracy, avoiding overfitting, and mitigating uncertainty (Tashman, 2000). The importance of adopting such strategies is demonstrated by the results of the M5 "Accuracy" competition, which indicate that a significant number of teams failed to select the most accurate set of forecasts from those submitted while the competition was still running (see Section 3). However, various CV strategies can be adopted and different conclusions can be drawn based on their design (Bergmeir et al., 2018). Selecting the time period when the CV will be performed, the size of the validation windows, how these windows will be updated, and the criteria used to summarize the forecasting performance are just some of the factors that forecasters must consider. In the M5 "Accuracy" competition, the top four best-performing methods and the vast majority of the top 50 submissions employed a CV strategy where at least the last four 28-day-long windows of available data were used to assess the forecasting performance, thereby providing a reasonable approximation of the post-sample accuracy. In addition to this CV scheme, the winner measured both the mean and standard deviation of the models that he developed. According to his validations, the recursive models in his approach were more accurate on average than the non-recursive models but with greater instability. Thus, he decided to combine those two models to ensure that the forecasts produced were both accurate and stable. Spiliotis et al. (2019a) stressed the necessity to consider the full distributions of forecasting errors and especially their tails when evaluating forecasting methods, thereby indicating that robustness is a prerequisite for achieving high accuracy. We hope that the M5 results will encourage more research in this area and contribute to the development of more powerful CV strategies.

Finding 7: Importance of exogenous/explanatory variables. Time series methods are usually sufficient for identifying and capturing historical data patterns (level, trend, and seasonality) and they can produce accurate forecasts by extrapolating these patterns. However, methods that rely solely on identifying and extrapolating historical data fail to effectively account for the effects of holidays, special days, promotions, prices, and possibly the weather. Moreover, these factors can affect historical data and distort the time series pattern unless they are removed before use for forecasting. In these settings, the information from exogenous/explanatory variables is of critical importance for improving the forecasting accuracy (Ma et al.,

2016). In the M5 "Accuracy" competition, all of the winning submissions utilized external information to improve the forecasting performance of their models. For example, *monsaraida* and other top teams found that several price-related features were significantly important for improving the accuracy of their results. Furthermore, the importance of exogenous/explanatory variables was supported by comparisons made between the benchmarks in the competition, as shown in the appendix of the supplementary material. For instance, ESX used information about promotions and special days as exogenous variables within exponential smoothing models and performed 6% better than ES_td, which employed the same exponential smoothing models but without considering exogenous variables. The same was true in the case of the ARIMA models, where ARIMAX was found to be 13% more accurate than ARIMA_td.

5. Discussion, limitations, advantages, and directions for future research

5.1. Discussion

The M5 "Accuracy" competition clearly showed that ML methods have entered the mainstream of forecasting applications, at least in the area of retail sales forecasting. The potential benefits of these methods are substantial and there is little doubt that retail firms will need to adopt them to improve the accuracy of their forecasts and support better decision making related to their operations and supply chain management.

Table 3 provides a simple comparison of Croston's method (CRO), which is widely used for forecasting intermittent demand data, with the sNaive, SES, ES_bu, ES_td, and ESX benchmarks (for more information regarding the benchmarks, please see the appendix of the supplementary material). On average, sNaive (a naive method that accounts for seasonality) was 11.5% more accurate than CRO, but its improvements were extremely uneven across various cross-sectional levels. These improvements started at 37.8% at the highest aggregation level, but dropped to 9.7% at level 9, and then become negative at levels 10, 11, and 12, where CRO was more accurate than sNaive by 13.0%, 20.2%, and 27.0% at these levels, respectively. These results demonstrate the value of CRO for forecasting intermittent demand data, as well as highlighting its limitations when applied to continuous series characterized by seasonality and trend. A similar comparison of CRO with SES (a simple exponential smoothing method that does not account for seasonality) also show the value added by CRO where SES was 1.3% less accurate on average than CRO in this case, and it only provided slightly better forecasts at level 10. Finally, after comparing the accuracy of CRO with the three top-performing exponential smoothing benchmarks in the competition (ES_bu, ES_td, and ESX), which are all capable of accounting for seasonality, we observed an average improvement of about 28%, starting at 54% at the top level and dropping to 1.1% at the lowest comprising the product-store level. It should be noted that the improvements were consistent for the three exponential smoothing methods across all aggregation levels.

Clearly, most of the improvements reported between CRO and the three top-performing exponential smoothing benchmarks were due to the ability of the latter to adequately capture seasonality, as well as the capacity to exploit explanatory/exogenous variables. In order to separate the effects of these two factors, we compared ES_td with ESX because both of these methods employ the same exponential smoothing models, but the latter also considers some indicative explanatory/exogenous variables. The average improvement by using ESX compared with ES_td was 5.7%, starting with 25.5% at the top level, dropping to 2.7% at level 9, and becoming negative at levels 10, 11, and 12. Thus, external information could improve the forecasting accuracy, but the seasonality observed mainly at higher aggregation levels was the most critical factor for improving the overall forecasting performance.

The improvements were much more substantial when CRO was compared with the top-performing method in the competition. The improvements started at 77.9% at the top level, and dropped to 10.8%, 7.3%, and 4.7% at levels 10, 11, and 12, respectively. However, the average overall improvement was 45.6%, with superior values up to the eighth level. According to these results, the superiority of the winning method is not likely to be disputed, particularly up to level 8, at least until new and more accurate ML methods for handling similar forecasting tasks are developed. In addition, CRO and other standards used for forecasting intermittent demand, such as SBA, TSB, ADIDA, and iMAPA (for more details about these methods, please see the appendix of the supplementary material), seem to have some value for low cross-sectional levels, especially at the product-store level.

Tables 3 and 4 provide useful information. The ML method employed by the winning team in the competition was based on the LightGBM algorithm and it obtained substantial improvements compared with the CRO method, with significantly better forecasts at the higher aggregation levels but also accurate ones at the product-store level, which was the hardest to predict due to its higher randomness. Key aspects of these improvements were the correct modeling of seasonality and exploitation of useful explanatory/exogenous variables. The remaining improvements can probably be attributed to the "cross-learning" approach applied and to the nonlinear nature of the models exploited by the winning team. Two interesting questions that need to be answered are whether LightGBM is truly the most suitable ML method for applying "cross-learning" in these forecasting applications, and how ML methods might be adjusted to provide significantly better forecasts than the existing approaches at both high and low aggregation levels.

The main advantages of the ML methods used by all 50 top-performing methods were probably their versatility in accurately predicting all 30,490 product-store series in the competition concurrently using "cross-learning" and their flexibility to be fine tuned based on the idiosyncrasies of the data set. In addition, few teams tried to identify and use a single "best" model to predict the series. Instead, many alternative models were built and subsequently averaged to obtain forecasts, where the winner developed 220 different models and used six models to

Table 3
Percentage improvements (according to WRMSSE) reported between indicative benchmarks in the competition, i.e., CRO, sNaive, SES, ES_bu, ES_td, and ESX. Positive numbers indicate accuracy gains for the second method compared with the first and negative numbers denote degraded accuracy.

Methods compared	Aggregation level												Average
	1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store	
CRO vs. sNaive	37.8%	26.4%	22.2%	31.5%	27.0%	19.2%	16.0%	14.8%	9.7%	−13.0%	−20.2%	−27.0%	11.5%
CRO vs. SES	−2.3%	−2.6%	−2.3%	−2.0%	−1.3%	−2.0%	−1.5%	−1.7%	−1.1%	1.1%	0.0%	−0.6%	−1.3%
CRO vs. ES_bu	52.7%	43.8%	37.1%	47.4%	42.7%	38.7%	33.8%	31.6%	25.9%	6.5%	3.2%	1.2%	29.9%
CRO vs. ES_td	47.8%	39.9%	28.0%	41.7%	34.1%	33.1%	26.4%	23.7%	18.5%	5.0%	2.8%	1.2%	24.7%
CRO vs. ESX	61.1%	46.0%	32.0%	51.8%	41.5%	37.3%	29.9%	26.4%	20.7%	5.2%	2.7%	1.0%	29.0%
ES_td vs. ESX	25.5%	10.1%	5.6%	17.3%	11.3%	6.2%	4.7%	3.4%	2.7%	0.2%	−0.1%	−0.2%	5.7%

Table 4
Percentage improvements (according to WRMSSE) reported between the winning submission (*YJ_STU*) and Croston's method (CRO). Column-wise minimum values are displayed in bold.

Methods compared	Aggregation level												Average
	1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store	
Winning team	**0.199**	**0.310**	**0.400**	**0.277**	**0.365**	**0.390**	**0.474**	**0.480**	**0.573**	**0.966**	**0.929**	**0.884**	**0.520**
CRO	0.900	0.915	0.923	0.909	0.971	0.941	0.987	0.940	0.983	1.083	1.002	0.926	0.957
Improvement	77.9%	66.1%	56.7%	69.6%	62.4%	58.6%	52.0%	49.0%	41.7%	10.8%	7.3%	4.5%	45.6%

predict each series. Furthermore, the approaches used by the top-performing teams to determine the most accurate forecasting method were unstructured and data-driven approaches based on CV strategies, which required little knowledge about forecasting and statistics. By contrast, the structured approaches widely used before the M5 competition focused on identifying a statistical method or combination of statistical methods that could provide the most accurate forecasts for each series, as well as the hybrid approaches employed by the winning M4 submissions, which mixed elements of both statistical and ML methods. The unstructured, agnostic approaches used in M5 required less experience of modeling and even less knowledge about the forecasting application considered, where they mostly depended on experimentation and data mining, without the need to understand the data itself and its characteristics. This was demonstrated by the fact that the winning method was developed by a student with little forecasting knowledge and little experience in building sales forecasting models. However, he effectively managed to win the competition and outperform thousands of competitors, including experienced Kaggle grandmasters among others. ML and computer science have generally gained more acceptance in the field of forecasting, so it would not be surprising to see the value of knowledge and experience become less important in developing and using forecasting models.

In the M4 competition, the five methods that achieved the most accurate results were also among the top five in terms of their average ranks according to the MCB test (Makridakis et al., 2020c), thereby indicating a high degree of correspondence between these two evaluation measures, i.e., the methods that provided the most accurate forecasts on average were also those that generally provided the most accurate forecasts separately for each series. However, this was not the case in the M5 "Accuracy" competition. As shown in Fig. 3, the winning method was ranked 13th according to the MCB test, the runner-up was ranked 48th, and the third, fourth, and fifth were ranked 6th, 17th, and 28th, respectively. It seems that the top five winning methods in M5 achieved their objective by minimizing the overall WRMSSE, and weighting the more expensive and fast-moving products more heavily to achieve their single objective rather than trying to provide accurate forecasts for every single series in the competition. As mentioned above, it remains to be seen whether these methods can be effectively adjusted to accurately predict all the series in the competition equally, especially those at the most disaggregated level.

Regarding the applicability of the competition's results, it would be interesting to see how long it takes until LightGBM and other ML methods are accepted by academics and widely utilized in practice by retail sales firms. In the academic world, exploring and adopting new methods does not usually require a long time because information is disseminated rapidly through journals and conferences. Clearly, the results of relevant studies must be replicated by other researchers and unless they disagree with those of the M5, they will hopefully be accepted with little delay. However, change occurs more slowly in the business world. First, it will take some time until practitioners learn the results of the M5 "Accuracy" competition and they will then need to be persuaded of their superior value. Second, a software program will have to be developed either in-house or by a consulting firm to implement the competition's winning methods or their variants. Third, the software should not require any special fine tuning to produce forecasts with the same accuracy as those reported in the competition. Fourth, the computational cost should not be prohibitively expensive so hundreds of thousands or even millions of forecasts can be produced on a weekly basis (Seaman, 2018), and finally, the software should be easy to integrate with the enterprise resource planning systems of firms in order to retrieve the raw data and provide the corresponding forecasts (Petropoulos, 2015). If these requirements are impossible to meet, then the most accurate benchmarks in the competition, which are relatively simple to implement and computationally cheap, will be utilized instead.

5.2. Limitations

The M5 "Accuracy" competition and other empirical studies conducted in the past provide valuable information about the accuracy of various forecasting approaches to guide academic research and give advice to practitioners about methods that can improve their forecasting performance and allow them to make better decisions. However, the value of this information depends greatly on the extent to which the data used for conducting the empirical comparison is representative of reality. It is difficult to argue otherwise when 100,000 times series that cover most data frequencies and various domains are utilized (Spiliotis et al., 2020a), but it is still possible even for large data sets to differ on average to those used in particular forecasting applications, such as those containing high-frequency data or cross-correlated series. In addition, when certain data sets cover a specific aspect of reality, such as daily Wikipedia page visits or daily retail sales, their findings are likely to be more representative of the application examined but cannot be generalized beyond the specific area covered by the data, except to draw some general conclusions about the methods used or how they were selected or evaluated. Therefore, there is a fundamental difference between the data used in M4, which covered six different data frequencies and six different domains, and those used in M5, which comprised retail sales data structured across 12 cross-sectional levels. The M5 data set refers to a specific forecasting application, but we still consider that it will be of great interest to a large number of retail firms that are specifically concerned with how to best forecast their daily sales and determine their inventory levels accordingly (Seaman, 2018). This was also the case in previous Kaggle competitions that involved daily and weekly product sales of large retail firms (Bojer & Meldgaard, 2021). The potential usefulness of the results obtained from the M5 competition is also supported by the findings reported by Theodorou et al. (2021) who explored the representativeness of its time series data by comparing their characteristics with those of two other grocery retailers that operate in different regions, sell different product types, and consider different marketing strategies, where they concluded that only minor discrepancies could be observed between them.

Another limitation of the M5 "Accuracy" competition is that it focused on the point forecast accuracy of the submitted methods, which were not directly linked to Walmart's underlying operational costs. Empirical studies have shown that minor improvements in accuracy can lead to substantial reductions in stock holding and higher service levels (Ghobbar & Friend, 2003; Pooya et al., 2019; Syntetos et al., 2010), but making a connection and translating forecasting error reduction into cost savings is far from trivial because the findings would depend greatly on the context of the task where the forecasts are applied (e.g., store versus warehouse replenishment), the type of the products being forecast (e.g., slow versus fast moving goods), and the forecasting horizon (e.g., monthly versus daily forecasts), among others. This is because different savings can be assumed for the same gains in accuracy depending on the supply chain of each company as well as its facilities, holding costs, and replenishment policies. Moreover, many assumptions must be made, particularly about the backlog and lost sales costs. Unfortunately, this detailed information was not made available to the M5 participants, and thus it was not possible for the organizers to evaluate the implications of the accuracy improvements reported by the winning submissions in monetary terms. However, we hope that the competition results will inspire relevant studies and motivate further research in the field. For instance, Spiliotis et al. (2021) recently used the M5 data to investigate the connection between forecasting accuracy and inventory performance, and concluded that simple empirical methods may result in similar if not lower monetary costs than more sophisticated approaches that achieve better accuracy on average.

The training and testing data used in the competition were made publicly available at the end of the competition, thereby ensuring openness and objectivity by allowing anyone to try to replicate the results of the competition, test alternative forecasting methods, and propose new, more accurate methods. This type of openness and objectivity is not possible in Kaggle competitions where the test data are not made available after the competition has ended, and the participants are not required to reveal the methods developed as the basis for their forecasts or to share their code for use by others (Bojer & Meldgaard, 2021). In contrast to the first four M competitions, this was also a serious problem with the M5 "Accuracy" competition because only 17 of the developers of the 50 top-performing methods shared information about their forecasting approaches, even after the organizers sent several emails requesting this information. The competition's rules stated that in order for the winners to receive their prizes, they had to reveal the method used and make their code available in order for the organizers to reproduce their results and compare the accuracy achieved to that of the originally submitted forecasts. The forecasts were successfully reproduced, which allowed us to provide the detailed descriptions of the five top performing methods presented in Section 4.2 and to ensure that the respective code for those participants that do not belong to a business firm will be uploaded onto GitHub, from where it can be downloaded and used for free.

5.3. Directions for future research

The findings obtained in the M4 and M5 "Accuracy" competitions, as well as in the latest two Kaggle sales forecasting competitions (Bojer & Meldgaard, 2021), indicate that ML methods are becoming more accurate than statistical methods, and thus their theoretical value and potential usage by organizations require reassessment. In the academic domain, more research is needed to verify that the more accurate results apply to areas other than hierarchical, retail sales forecasting and that other ML methods are not more accurate than the winning methods in these competitions.

From a practical perspective, it is necessary to determine the extra costs incurred to run ML methods versus the standard statistical methods, and whether their accuracy improvements would justify higher costs (Fry &

Brundage, 2020; Nikolopoulos & Petropoulos, 2018). If both of these concerns can be satisfied, two other issues require further investigation. The first issue is related to understanding how ML methods produce their forecasts and account for factors such as price, promotions, and special days. Managers are typically unwilling to make decisions when they cannot understand the logic of the methods that they plan to use. This is a major problem that affects all ML models and it eventually needs to be solved. At present, interim solutions must be found by comparing the forecasts obtained by ML methods and those using known benchmarks, as shown in Tables 3 and 4, thereby allowing the contribution of each factor to be indirectly determined. The second issue concerns which and how many ML models must be combined to achieve improvements in accuracy. Instead of developing ensembles of hundreds of models, it is possible that eliminating the worst or those less likely to improve forecasting performance could improve the overall accuracy without greatly increasing the computational cost.

Another alternative to explore further is the concept of "horses for courses" (Petropoulos et al., 2014), i.e., the idea that different methods could potentially be used to forecast various aggregation levels separately based on their corresponding performance. If *Daniela*'s (ranked 28th) method was used to forecast the top level, followed by *Matthias* (ranked 2nd) for levels 2 to 6, *YJ_STU* for levels 7 to 9 (ranked 1st), *mf* for levels 10 and 11 (ranked 3rd), and *wyzJack_STU* for level 12 (ranked 6th), the overall accuracy of the winning submission would have been improved by an additional 2.3%. However, this selective approach would have required the identification in advance of the best-performing method at each level. It would have also required reconciling the forecasts produced by these methods so they were coherent across the various aggregation levels. This task proved much more challenging than expected in the M5 "Accuracy" competition, where many teams tried to apply well-known reconciliation methods (Hyndman et al., 2011) but failed, probably due to the size of the data set, complexity of the underlying hierarchy, and other limitations (non-negative forecasts and additional computational cost). These insights indicate that there is much potential in this particular area of forecasting and that substantial improvements in accuracy may be obtained by developing new hierarchical forecasting methods with the capability of reconciling forecasts so the average forecast error is minimized, but also separately at each aggregation level. The potential of these methods is highlighted when we consider that the simple, equal weighted combination of the above-mentioned five submissions, which directly provided coherent forecasts, yielded 2% more accurate forecasts on average than the winning submission, although they were not always the best at each individual level, as shown in Table 5.

6. Conclusions

It has been almost 40 years since the first M forecasting competition, which was the first of its type in a new scientific field (Makridakis et al., 1982). At that time, only seven contestants tested their methods against each other and predicted up to 1,001 time series to determine the most accurate, which contrary to expectations, was a simple exponential smoothing method rather than the statistically sophisticated Box–Jenkins method applied to ARIMA models known as the "king" of that era. In addition, the competition established the value of combining, where empirical tests demonstrated that combining the forecasts produced by more than one method improved the accuracy and reduced uncertainty. This was an important finding at the time because it was considered that a single appropriate model existed for each time series and that this model could be identified judgmentally by inspecting the characteristics of the series. The M2 (Makridakis et al., 1993) was also a small-scale competition where five participants competed in real time between 1987 and 1989, so the contestants could incorporate their judgments by adjusting their statistical forecasts using inside company information and knowledge about the current economy. In contrast to expectations, the competition results showed that human judgment did not improve the accuracy of the statistical forecasts and combining was the most accurate method for predicting the 29 series in the competition. In the M3 (Makridakis & Hibon, 2000) conducted in 2000, the number of time series increased substantially to 3003 and the number of participants was 15, where they employed simple and statistically sophisticated methods, as well as rule-based and NN methods. Simple methods still outperformed the relatively more complex ones and a new simple method called Theta (Assimakopoulos & Nikolopoulos, 2000) was the most accurate on average, and forecast combinations continued to produce more accurate results than combining individual methods.

The M4 (Makridakis et al., 2020c) was conducted in 2018 and there were dramatic increases to 100,000 time series and 49 participants. In addition to the accuracy competition, it was also necessary to estimate the uncertainty by asking the participants to provide the 95% prediction intervals around their point forecasts for each of the 100,000 series. As mentioned above, the M4 ended a long forecasting winter by reversing the findings of the previous three competitions, where it was concluded that two sophisticated methods based on a mixture of statistical and ML concepts outperformed all of the others in terms of both accuracy and uncertainty. The forecasting spring continued with the M5, which demonstrated the superiority of ML methods, particularly LightGBM, where the 50 top-performing methods achieved more than 14% better performance compared with the most accurate statistical benchmark and more than 20% better for the top five. All five M competitions have demonstrated that combining models improves the forecasting accuracy. However, M1, M2, and M3 showed that simple statistical methods were more accurate than more complex, sophisticated methods. In M4, only two sophisticated methods were found to be more accurate than simple statistical methods, where the latter occupied the top positions in the competition. By contrast, all 50 top-performing methods were based on ML in M5. Therefore, M5 is the first M competition in which all of the top-performing methods

Table 5
Percentage improvements (according to WRMSSE) reported between the winning submission (*YJ_STU*), best-performing submission at each aggregation level, i.e., *Daniela* for level 1, *Matthias* for levels 2 to 6, *YJ_STU* for levels 7 to 9, *mf* for levels 10 and 11, and *wyzJack_STU* for level 12, and the simple, equal weighted combination of these five methods. Column-wise minimum values are shown in bold.

Methods compared	Aggregation level												Average
	1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store	
Winning team	0.199	0.310	0.400	0.277	0.365	0.390	0.474	0.480	0.573	0.966	0.929	0.884	0.520
Best team per level	**0.162**	0.294	0.400	**0.246**	**0.349**	0.381	0.474	0.480	0.573	**0.950**	**0.917**	**0.875**	**0.508**
Combination (COMB)	0.181	**0.283**	**0.390**	0.260	0.365	**0.371**	**0.473**	**0.472**	**0.572**	0.960	0.921	0.876	0.510
Improvement of COMB over winner	8.9%	8.5%	2.5%	6.1%	0.1%	4.7%	0.3%	1.7%	0.1%	0.6%	0.9%	0.9%	2.0%
Improvement of COMB over the best team per level	−11.9%	3.5%	2.5%	−5.7%	−4.5%	2.6%	0.3%	1.7%	0.1%	−1.0%	−0.5%	−0.1%	−0.4%

were both ML methods and better than all of the statistical benchmarks and their combinations. It was demonstrated that LightGBM can be used to effectively process numerous correlated series and exogenous/explanatory variables, and reduce the forecast error. Moreover, deep learning methods such as DeepAR and N-BEATS, which provide advanced state-of-the-art ML implementations, can potentially further improve the forecasting accuracy in hierarchical retail sales applications.

In summary, the M5 "Accuracy" competition provided the forecasting community with the following three important new findings.

- The superior accuracy of the LightGBM method for predicting hierarchical retail sales resulted in substantial improvements compared with the benchmarks considered.
- The external adjustments utilized in some methods were beneficial for improving the accuracy of the baseline forecasting models.
- Exogenous/explanatory variables were important for improving the forecasting accuracy of time series methods.

In addition, M5 reaffirmed the value of the following three findings obtained in the previous M competitions regarding forecasting accuracy improvement.

- Combining
- "Cross-learning"
- Cross-validation

The exceptional performance of statistical methods versus ML method found by Makridakis et al. (2018b), as well as in the early Kaggle competitions (Bojer & Meldgaard, 2021), first shifted toward both ML and statistical methods in the M4 competition, and then to exclusively ML methods in the Kaggle competitions after 2018 and M5, as described in this study. It will very interesting to see whether ML methods continue to dominate statistical methods in the future, particularly for other types of data that are not exclusively related to hierarchical retail sales applications.

Finally, the integration of statistics and data science into a unique field covering all academic aspects of forecasting and uncertainty is important, as well as determining how to increase the usage of forecasting in organizations by persuading executives of the benefits of systematic forecasting for improving their bottom line (Makridakis et al., 2020).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2021.11.013.

References

Abouarghoub, W., Nomikos, N. K., & Petropoulos, F. (2018). On reconciling macro and micro energy transport forecasts for strategic decision making in the tanker industry. *Transportation Research Part E: Logistics and Transportation Review*, *113*, 225–238.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*, 521–530.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83.

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*, 587–603.

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). Springer Berlin Heidelberg.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*, 754–762.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, *29*, 510–522.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*, 3–23.

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

Fry, C., & Brundage, M. (2020). The M4 forecasting competition — A practitioner's view. *International Journal of Forecasting*, *36*, 156–160.

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

Ghobbar, A. A., & Friend, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. *Computers & Operations Research*, *30*, 2097–2114.

Gilliland, M. (2020). The value added by machine learning approaches in forecasting. *International Journal of Forecasting*, *36*, 161–166.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*, 405–408.
Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*, 7–14.
Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*, 2579–2589.
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 3146–3154). Curran Associates, Inc..
Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, *32*, 788–803.
Kolassa, S. (2020). Why the "best" point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, *36*, 208–211.
Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*, 397–409.
Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*, 291–302.
Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2021). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, 1–18.
Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*, 493–518.
Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, *249*, 245–257.
Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, *9*, 527–529.
Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.
Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*, 835–838.
Makridakis, S., Bonnell, E., Clarke, S., Fildes, R., Gilliland, M., Hoover, J., & Tashman, L. (2020). The benefits of systematic forecasting for organizations: The UFO project. *Foresight: The International Journal of Applied Forecasting*, *59*, 45–56.
Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*, 5–22.
Makridakis, S., & Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*, 1–26.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). Responses to discussions and commentaries. *International Journal of Forecasting*, *36*, 217–223.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*, 54–74.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*, 1325–1336.
Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R. L. (2020). The M5 uncertainty competition: Results, findings and conclusions. *Int. J. Forecast.*.
Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*, 86–92.
Nikolopoulos, K., & Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? *Computers & Operations Research*, *98*, 322–329.
Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *CoRR*, arXiv:1905.10437.
Petropoulos, F. (2015). Forecasting support systems: Ways forward. *Foresight: The International Journal of Applied Forecasting*, (39), 5–11.
Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). 'Horses for courses' in demand forecasting. *European Journal of Operational Research*, *237*, 152–163.
Petropoulos, F., & Svetunkov, I. (2020). A simple combination of univariate models. *International Journal of Forecasting*, *36*, 110–115.
Pooya, A., Pakdaman, M., & Tadj, L. (2019). Exact and approximate solution for optimal inventory control of two-stock with reworking and forecasting of demand. *Operational Research: An International Journal*, *19*, 333–346.
Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*, 1181–1191.
Schwertman, N. C., Gilks, A. J., & Cameron, J. (1990). A simple noncalculus proof that the median minimizes the sum of the absolute deviations. *The American Statistician*, *44*, 38–39.
Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, *34*, 822–829.
Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, *37*, 1072–1084.
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., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality? *International Journal of Forecasting*, *36*, 37–53.
Spiliotis, E., Makridakis, S., Kaltsounis, A., & Assimakopoulos, V. (2021). Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data. *International Journal of Production Economics*, *240*, Article 108237.
Spiliotis, E., Nikolopoulos, K., & Assimakopoulos, V. (2019). Tales from tails: On the empirical distributions of forecasting errors and their implication to risk. *International Journal of Forecasting*, *35*, 687–698.
Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2019). Improving the forecasting performance of temporal hierarchies. *PLOS ONE*, *14*, 1–21.
Spiliotis, E., Petropoulos, F., Kourentzes, N., & Assimakopoulos, V. (2020). Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Applied Energy*, *261*, Article 114339.
Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, *21*, 303–314.
Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society*, *56*, 495–503.
Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting*, *26*, 134–143.
Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*, 437–450.
Theodorou, E., Wang, S., Kang, Y., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Exploring the representativeness of the M5 competition data. *International Journal of Forecasting*.
Zhou, H., Qian, W., & Yang, Y. (2020). Tweedie gradient boosting for extremely unbalanced zero-inflated data. *Communications in Statistics. Simulation and Computation*, 1–23.

Further reading

Bergmeir, C., & Benítez, J. M. (2012). RSNNS: neural networks using the stuttgart Neural Network Simulator (SNNS). R package version 0.4.12.

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.

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*, 1–22.

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.

Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, *41*, 4235–4244.

Liaw, A., & Wiener, M. (2018). Randomforest: Breiman and Cutler's random forests for classification and regression. R package version 4.6.14.

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, *6*, 525–533.

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., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, *66*, 914–924.

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

Svetunkov, I. (2020). Smooth: forecasting using state space models. R package version 2.5.6.

Teunter, R. H., & Duncan, L. (2009). Forecasting intermittent demand: a comparative study. *Journal of the Operational Research Society*, *60*, 321–329.

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.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks:: the state of the art. *International Journal of Forecasting*, *14*, 35–62.