Predicting/hypothesizing the findings of the M5 competition

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

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

Abstract

The scientific method consists of making hypotheses or predictions and then carrying out experiments to test them once the actual results have become available, in order to learn both from successes and mistakes. This approach was followed in the M4 competition with positive results and has been repeated in the M5, with its organizers submitting their ten predictions/hypotheses about its expected results five days before its launch. The present paper presents these predictions/hypotheses and evaluates their realization according to the actual findings of the competition. The results indicate that well-established practices, like combining forecasts, exploiting explanatory variables, and capturing seasonality and special days, remain critical for enhancing forecasting performance, re-confirming also that relatively new approaches, like cross-learning algorithms and machine learning methods, display great potential. Yet, we show that simple, local statistical methods may still be competitive for forecasting high granularity data and estimating the tails of the uncertainty distribution, thus motivating future research in the field of retail sales forecasting.

Keywords: Forecasting competitions, M competitions, Accuracy, Uncertainty, Retail sales forecasting

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

1. Introduction

The objective of the M competitions is to learn from objective, empirical evidence how to improve forecasting performance and use such learning to advance the theory and practice of forecasting. These competitions have been open and replicable, attracting the interest of the forecasting community, influencing the field of forecasting, and serving as a benchmark and standard testing ground to evaluate new forecasting methods and practices (Hyndman, 2020).

The M5 competition (Makridakis et al., 2021), focusing on a retail sales forecasting application, was the continuation of the previous four and consisted of two parallel challenges, namely the "Accuracy" (Makridakis et al., 2020c) and "Uncertainty" ones (Makridakis et al., 2020d). Its main purpose was to learn how to accurately forecast grouped daily unit sales of retail companies and how to precisely estimate the uncertainty distribution of the realized values of the predicted series.

The aim of this paper is to present the ten predictions/hypotheses that the organizers of M5 made before its launch about its findings, including assumptions that were driven by the conclusions of past forecasting studies and competitions, as well as recent technological and methodological advances, but also natural open questions in view of the novel features the setup of the M5 introduced. These predictions/hypotheses make a clear statement of our expectations and possibly those of other forecasters and avoid the problem of rationalizing the findings after the fact. This approach allows the equivalent of the scientific method, widely used in hard sciences, to be applied in a forecasting study, in order to learn from successes as well as avoid mistakes and errors. As such, this paper demonstrates that the M5 competition has not used HARKing (Hypothesizing After the Results are Known; Kerr, 1998) in predicting its findings, i.e., has not presented post hoc hypotheses that are based on actual results as if they were a priori ones, a practice which, at least in some of its forms, is seen as inappropriate[1] (Rubin, 2017, 2019; Lishner, 2021). Thus, this paper continues the major innovation of the M4 competition (Makridakis et al., 2020b), where its organizers assessed the ten predictions/hypotheses they made more than two months before its completion (Makridakis et al., 2020a).

Given that our ten hypotheses/predictions test methods and approaches that could be considered as best practices in the field of forecasting, as well as long-lasting beliefs about which factors affect forecasting accuracy and what is considered important for improving forecasting performance, confirming or falsifying them becomes critical for the forecasting community to make sure that these practices and beliefs hold true in practice in retail sales forecasting applications. Moreover, based on the assessment of these hypotheses/predictions, and in particular those that may be proven wrong or partially correct, unexpected findings and welcome innovations can be effectively identified, thus motivating future research in the field.

The ten predictions/hypotheses of the organizers of M5 were submitted to Pierre Pinson, the EiC of *International Journal of Forecasting*, on February 28th, 2020, i.e., five days

[1]This includes translating Type I errors into theory, not communicating information about what did not work, disguising accommodation as prediction, taking unjustified statistical license, encouraging the adoption of narrow, context-bound theory or the retention of too-broad theory, and inhibiting the identification of alternative hypotheses, among others (Kerr, 1998).

before the launch of the competition. We have now evaluated these predictions/hypotheses, reflecting on their successes and errors, as well as other findings of the competition that our predictions/hypotheses missed or did not specify explicitly.

2. Our predictions/hypotheses

I. Combinations of forecasting methods will continue to outperform submissions that rely on individual methods. However, sophisticated combination schemes will not lead to significantly better forecasts than simpler ones (*this prediction refers to both point and probabilistic forecasts*).

The main finding of the first M competition (Makridakis et al., 1982), dating back to the early '80s, was that combining simple time series forecasting methods (e.g., Simple, Holt, and Damped exponential smoothing) results in superior forecasts that outperform both the individual methods used for combining and other, more sophisticated methods (e.g., ARIMA models). The same finding was confirmed in all M competitions that followed (Makridakis et al., 1993; Makridakis & Hibon, 2000; Makridakis et al., 2020b), where the vast majority of the top-performing submissions utilized combinations of various forecasting methods. For instance, the Theta method (Assimakopoulos & Nikolopoulos, 2000), that won the M3 competition, is based on a framework that decomposes data into two Theta lines which are individually extrapolated using Simple exponential smoothing (SES) and linear regression on a time-trend indicator and combined using equal weights to enhance forecasting accuracy. Similarly, in the M4 competition (Makridakis et al., 2020b), some of the top-performing teams considered simple combinations of time series forecasting methods (Petropoulos & Svetunkov, 2020), i.e., used standard operators such as the arithmetic mean, the geometric mean, and the median for combining the base forecasts, with the runner-up (Montero-Manso et al., 2020), who utilized an XGBoost algorithm to optimally determine the combination weights of nine forecasting methods, being the only team to consider a sophisticated combination scheme, i.e., a framework where combination weights are specified based on the accuracy of the methods to be combined, possibly using meta-data (e.g., time series features) as additional input. The beneficial value of combining has a long history (Bates & Granger, 1969; Claeskens et al., 2016), expanding with new applications in the form of cross-sectional (Hyndman et al., 2011; Wickramasuriya et al., 2019), temporal (Athanasopoulos et al., 2017; Spiliotis et al., 2019), and cross-temporal aggregation (Spiliotis et al., 2020c), bagging (Petropoulos et al., 2018), boosting (Barrow et al., 2020), and ensembles (Kourentzes et al., 2014; Oreshkin et al., 2019), among others.

The M5 competition contributes to the literature by re-confirming the benefits of combining and indicating that simple combinations of methods, pooled using robust and objective cross-validation (CV) strategies, can result in superior retail sales forecasts. Specifically, in the "Accuracy" challenge, the top three performing methods utilized simple, equal-weighted combinations of multiple decision-tree-based models or neural networks (NNs) (for more details on the "Accuracy" winning methods, see section 4.2 in Makridakis et al., 2020c). This was also true for many other teams among the top 50 performing ones. Similarly, in the "Uncertainty" challenge, the methods ranked 1st, 3rd, and 7th, considered relatively simple combinations of regression trees or NNs, including equal-weighted and geometric mean combinations (for more details on the "Uncertainty" winning methods, see section 4.2 in

Makridakis et al., 2020d). Interestingly, all combination approaches employed in M5 by the top-performing teams can be classified into two main categories: (i) combinations of different types of methods (e.g., ensembles of regression trees and NNs) and (ii) combinations of methods of the same type, trained across different subsets of data (e.g., models specialized to accurately predict the series at a store, store-category, store-department, or product-store level). Undoubtedly, further research is required to replicate this finding and better understand the combination approaches with the highest potential value. Nevertheless, the results suggest that combining forecasts from a robust pool of diverse methods, trained across selected series of similar characteristics, may be more important than investing in sophisticated combination schemes.

II. The most accurate forecasting methods will utilize "cross-learning", i.e., extract information from the global data set (including data of different aggregation levels) to predict the individual series (*this prediction refers to both point and probabilistic forecasts*).

Until recently, forecasting methods were mostly employed in a series-by-series fashion, meaning that forecasting models were trained and optimized using only the information contained in the single series to be predicted (Makridakis et al., 2018). This was also true in retail sales and intermittent demand forecasting settings (Kourentzes, 2013; Seeger et al., 2016). "Local" training strategies (Januschowski et al., 2020) can be effective when the series consist of many observations and, as a result, the model can effectively identify established patterns/relationships (Barker, 2020). However, when data is limited or sparse, and/or series are highly correlated, cross-learning, i.e., learning from multiple series how to accurately predict individual ones, can significantly improve overall forecasting performance (Oreshkin et al., 2019; Ma & Fildes, 2021, 2020; Spiliotis et al., 2020b; Montero-Manso & Hyndman, 2021). Moreover, if a single, cross-learning (or "global") model is used to predict numerous series, the computational cost of the forecasting process is significantly lower (Semenoglou et al., 2020). The beneficial value of cross-learning was highlighted in the results of the M4 competition, where both the winner (Smyl, 2020) and runner up (Montero-Manso et al., 2020) employed cross-learning approaches to extract valuable information from the complete data set, followed by several other studies that exploit both the local and global patterns of the time series data available (Sen et al., 2019; Li et al., 2020; Godahewa et al., 2021). In the area of retail sales forecasting, Spiliotis et al. (2020b) considered a data set of 3,300 daily demand series and found that cross-learning methods provided better results than local approaches. In a similar fashion, in the area of hierarchical forecasting, base forecasts produced at different cross-sectional or temporal levels are typically combined or adjusted in order for the final forecasts to reflect the patterns observed across the complete hierarchy, while also being coherent (Hyndman et al., 2011; Spiliotis et al., 2021).

The results of the M5 competition re-confirm these findings and indicate that cross-learning is the natural way of applying forecasting methods to improve overall forecasting performance. In fact, all top-performing methods of the "Accuracy" and "Uncertainty" challenges employed cross-learning by extracting information from multiple series at various aggregation levels, such as store, product, product-category, or product-department (for more details on the hierarchical structure of the M5 data set, see section 4 in Makridakis et al., 2021). Thus, in contrast to the previous M competitions, the M5 winning meth-

ods extended cross-learning to effectively account for correlations naturally observed both among the series of the same aggregation level and the series of different aggregation levels, a practice which has long been considered useful in the forecasting literature when dealing with hierarchical and grouped series (Panagiotelis et al., 2021). What seemed to be of particular importance, thereby requiring further research, was the selection of the pools of series used for training the cross-learning methods. For example, the winner of the "Accuracy" challenge, who considered pools of product-store series using data per store (10 models), store-category ($10 \times 3 = 30$ models), and store-department ($10 \times 7 = 70$ models), managed to produce significantly better forecasts than the rest of the teams which utilized similar baseline forecasting models (LightGBM), using however different pools of series for their training. Thus, it would be interesting to investigate how the number and the correlation of the series being pooled affects the performance of the cross-learning methods and if there are any trade-offs between these factors.

III. Methods exploiting explanatory/exogenous variables will provide significantly better forecasts than those that do not (*this prediction refers to both point and probabilistic forecasts*).

The M5 was the first of the M competitions to include explanatory/exogenous variables along with time series data, namely information about holidays, special days, Supplement Nutrition Assistance Program (SNAP) activities, prices, and calendar effects (for more details on the explanatory/exogenous variables included in the M5 data set and their overall impact on unit sales, see section 4 in Makridakis et al., 2021). Such variables offer valuable information to the forecasting models that cannot be extracted from the time series data itself and, therefore, allow the improvement of forecasting performance (Ma et al., 2016; Fildes et al., 2019). The results of the M5 competition confirmed the value of explanatory/exogenous variables by finding that the vast majority of the winning methods used all or part of the provided variables as input. Similarly, benchmarks that utilized this information, like ESX and ARIMAX, were found to be more accurate than their standard counterparts, i.e., ES_td and ARIMA_td[2], by a margin of 6% and 13%, respectively (for more details on the performance of the "Accuracy" benchmarks, see the appendix in Makridakis et al., 2020c). Moreover, many teams reported that some variables, like changes in prices and special days, were critical for improving the overall performance of their submissions. Nevertheless, it seems that the importance of the individual explanatory/exogenous variables varied among the teams, depending strongly on the forecasting methods employed. Therefore, there is a need for future research to investigate in depth the effect of these variables on post-sample forecasting performance by identifying the most critical ones, and connect their importance with the base forecasting models employed, as well as the rest of the features used as input. Additionally, since the explanatory/exogenous variables involved in the competition did not require forecasting their future values, assumed to be known in advance, it would be interesting to investigate whether their beneficial value would be

[2]ES_td and ARIMA_td employ an algorithm to automatically select the most appropriate exponential smoothing or ARIMA model, respectively, for predicting total unit sales (level 1), indicated through information criteria (Hyndman et al., 2002). The top-down method is then used to forecast the rest of the series (levels 2-12). ESX and ARIMAX employ the same methods, but this time two exogenous/explanatory variables (SNAP activities and special days identifiers) are used as regressors, in addition to historical data.

significantly affected if they had to be predicted or not.

IV. The most accurate forecasting methods will be machine learning[3] (ML) or deep learning (DL), or depend heavily on ML/DL methods (*this prediction refers to both point and probabilistic forecasts*).

In the M4 competition, the top two performing methods were hybrids that utilized ML models, namely recurrent NNs (Smyl, 2020) and XGBoost (Montero-Manso et al., 2020). Although both methods involved innovative, effective ways for applying cross-learning and improving overall forecasting performance, they were not "pure" ML in nature, i.e., used ML models along with standard statistical methods, while the architectures of the models used were not deep. At the same time, some combinations of standard statistical methods were very competitive, providing statistically insignificant different results compared to the winning submissions. Therefore, the organizers of M4 concluded that hybrid methods and combinations of statistical and ML methods was the way forward and that pure ML approaches were still underperforming, possibly because DL was still in its infancy in the area of time series forecasting. However, soon after the end of the competition, Oreshkin et al. (2019) introduced N-BEATS, a deep neural architecture that is based on backward and forward residual links and a very deep stack of fully-connected layers, claiming superior forecasting accuracy than the M3 and M4 winning methods. Similarly, Salinas et al. (2020) introduced DeepAR, a methodology for producing probabilistic forecasts using deep, autoregressive recurrent NNs on a large number of related time series, reporting superior precision to standard methods used in time series probabilistic forecasting. These studies demonstrated that pure ML methods and especially DL ones could be exploited to improve forecasting performance, thus introducing a new era in forecasting research and practice.

Based on the above, it was predicted that the M5 competition would be dominated by pure ML methods, DL models, and combinations of ML/DL approaches. This was actually the case as the majority of the top 50 competitors' methods were based on LightGBM, a decision-tree-based model. At the same time, the runner-up of the "Accuracy" challenge, *Matthias*, employed an equal-weighted combination of various LightGBM models that were adjusted based on the forecasts produced by N-BEATS. Similarly, *mf*, ranked 3$^{\text{rd}}$ in the "Accuracy" competition, employed an equally weighted combination of 43 DeepAR models. In the "Uncertainty" challenge, *Wal Dash Mart*, ranked 13$^{\text{th}}$, and *RandomObserver*, ranked 27$^{\text{th}}$, employed DeepAR and GRU-based models that are fed with embeddings, respectively. Moreover, many of the top 50 performing teams of both challenges considered LSTM NNs of deeper architectures than those employed in M4. The results of the M5 competition

[3]We categorize as "ML" any method that allows for data relationships and patterns to be identified, memorized, and estimated automatically, thus being generic in nature. In contrast, we categorize as "statistical" any method that prescribes the underlying data generating process (e.g. in terms of trend and seasonal patterns), thus making strong structural assumptions about the series being forecast. In this regard, methods like exponential smoothing, ARIMA, Croston's, and Theta are considered statistical, while methods such as neural networks and decision trees are considered ML. This distinction is similar to the "model-driven" versus "data-driven" categorization proposed by Januschowski et al. (2020), the "structured" versus "unstructured" categorization proposed by Barker (2020), as well as the one adopted in the M4 competition by its organizers for discussing its results. Although in some edge cases this distinction may be challenging, we believe it is intuitive for the vast majority of the researchers (NNs and XGBoost models are widely accepted as ML by the data science community) and helpful for discussing the results of the M5 competition.

demonstrate that ML methods and approaches that build on deep NNs can provide more accurate forecasts and more precise estimators of uncertainty than statistical or even hybrid approaches, confirming their beneficial value in the field of retail sales forecasting.

V. Simple forecasting methods in terms of of trainable parameters, such as the benchmarks included in M5, will be significantly less accurate than the top-performing methods (*this prediction refers to both point and probabilistic forecasts*).

The M4 was the first of the M competitions to show that sophisticated methods, such as NNs, decision-tree-based models, and hybrids, can provide better forecasts than simple ones, such as exponential smoothing. Based on this finding, the long-lasting belief that simple methods were at least as good as more sophisticated ones for generic time series forecasting tasks (Makridakis et al., 2018) passed into history[4]. The M5 competition reconfirmed this major finding of M4 for the case of retail sales forecasting, both in terms of point and probabilistic forecasts: all top-performing methods were based on regression trees, NNs, or combinations of those, providing forecasts that exceeded the performance of the benchmarks by more than 20%. Moreover, even when simple methods were considered by the top-performing teams, these were mostly used to calibrate the results of the sophisticated methods utilized, or to generate base forecasts among other methods and construct ensembles. The only minor exceptions where simple methods were proven to be relatively competitive, were probably the forecasts at the lowest aggregation levels and the tails of the uncertainty distributions. Table 1 summarizes the improvements reported for the top 3 performing methods of each forecasting challenge over the respective top-performing benchmark.

Table 1: Percentage improvements (%) reported between the top 3 performing submissions of the "Accuracy" and "Uncertainty" challenges and the top-performing benchmarks, i.e., ES_bu and ARIMA, respectively. The improvements refer to the official measures used for evaluating forecasting performance, i.e., the Weighted Root Mean Squared Scaled Error (WRMSSE) and Weighted Scaled Pinball Loss (WSPL), and are reported both for each aggregation level separately and in total. More detailed results on the performance of the submissions made in the "Accuracy" and "Uncertainty" challenges can be found in Makridakis et al. (2020c) and Makridakis et al. (2020d), respectively.

Rank	Team	Aggregation level												Average
		1	2	3	4	5	6	7	8	9	10	11	12	
		Accuracy												
1	YJ_STU	53.3	39.7	31.0	42.1	34.5	32.4	27.5	25.3	21.3	4.5	4.1	3.4	22.5
2	Matthias	56.3	42.8	28.3	48.5	37.3	34.0	26.5	22.7	18.4	-1.1	0.5	0.9	21.3
3	mf	44.6	37.9	27.4	35.6	28.7	29.8	24.2	21.5	17.6	6.1	5.4	4.4	20.1
		Uncertainty												
1	Everyday Low SPLices	53.2	35.1	30.7	38.1	32.9	31.8	32.7	24.2	20.9	11.8	9.4	12.3	24.9
2	GoodsForecast - Nick Mamonov	57.6	34.5	27.6	42.2	32.9	31.2	30.2	21.9	18.4	6.2	4.7	7.3	22.4
3	Ouranos	48.1	32.4	30.7	36.7	32.9	31.2	33.2	26.4	22.9	5.9	3.7	7.3	22.4

VI. Methods that will be accurate in predicting the bottom aggregation level,

[4]We should clarify that sophisticated methods have been successfully used before 2018, i.e., the year M4 took place, in particular forecasting applications, such as energy and finance. However, these methods were built on regression approaches that exploited various explanatory/exogenous variables, being less successful for generic time series forecasting tasks, like the ones examined in M, M3, and M4.

i.e., product-store series, will not necessarily be similarly accurate in predicting the rest of the levels, especially the top one (*this prediction refers to point forecasts*).

George Box famously said that "all models are wrong, but some are useful". Given that different methods may be appropriate for capturing different data patterns, this becomes particularly relevant in forecasting tasks that require the prediction of multiple series of diverse characteristics (Spiliotis et al., 2020a). Although the M5 competition consisted of aligned and mostly correlated series, it did include series of diverse characteristics that change across the 12 aggregation levels considered according to the corresponding level of granularity. Specifically, the series of the M5 competition can be roughly categorized (Syntetos & Boylan, 2005; Syntetos et al., 2005) into smooth, which are found at the top aggregation levels (1-9), and intermittent and erratic ones, which are found at the bottom aggregation levels (10-12). The former series display strong seasonal patterns and trend, while the latter notable variations in demand that in many periods is zero (for more details on the demand classification of the M5 series and an overview of their summary statistics, see section 4 in Makridakis et al., 2021). It can be stated, therefore, that the development of forecasting methods that provide accurate results for all aggregation levels will remain to be a challenging task.

The results of the M5 competition re-confirm this statement and indicate that the concept of "horses for courses" (Petropoulos et al., 2014) is important for improving overall forecasting performance and effectively supporting decisions across different aggregation levels, even when the series involved are highly correlated to each other and organized in a grouped fashion. Notably, none of the top-performing methods provided the "best" forecasts across all aggregation levels, with the winner of the "Accuracy" challenge being the most accurate method at levels 3, 7, 8, and 9, while the winner of the "Uncertainty" challenge at none of the levels. Moreover, most of the methods that performed well at the lowest aggregation level produced sub-optimal results at the top level, and vice versa. For instance, from the top 50 performing methods of the "Accuracy" competition that were ranked in the top five at level 1, their median rank at level 12 was 44. Similarly, from the top 50 performing methods of the "Accuracy" competition that were ranked in the top five at level 12, their median rank at level 1 was 37. In the "Uncertainty" competition, the corresponding ranks were 44 and 45, respectively. To further stress the necessity of identifying "horses for courses" in retail sales forecasting, we note that if the "best" method of the top 50 performing ones was effectively selected at each aggregation level, the improvements over the winners of the 'Accuracy" and "Uncertainty" challenges would be 2.3% and 4.4%, respectively.

This finding can also be linked to the "no free lunch theorem" according to which, given a particular measure for evaluating performance, choosing an appropriate forecasting method requires making specific assumptions about the types of series the method will be used for in order for it to perform significantly better than other methods. In the M5 competition, the measures used for assessing the performance of the submissions assigned equal weights to the 12 aggregation levels considered, meaning that producing the most accurate forecasts at the bottom level would not have guaranteed top performance in total (the bottom level accounts for 1/12 of the overall score). This property was widely considered by the participants of M5 who used various CV strategies to optimally train their methods accordingly. Undoubtedly, when the series of different levels are highly correlated and share similar patterns, optimizing

for the bottom level could imply adequate forecasting performance for the higher levels as well (Nenova & May, 2016). Nevertheless, the literature in hierarchical forecasting research suggests that this is not always the case and that different forecasting approaches are required depending both on the structure of the hierarchy and the characteristics of the series involved (Abolghasemi et al., 2020).

VII. Seasonality, observed at higher aggregation levels, will be the most critical factor for improving overall forecasting accuracy. This would include holidays and special days. However, at the bottom level, intermittency will be the most critical factor influencing forecasting accuracy (*this prediction refers to both point and probabilistic forecasts*).

As noted before, the majority of the aggregation levels considered in the M5 competition consist of series that display strong seasonal patterns, observed both at a weekly and yearly level (for more details on the seasonal patterns observed in the M5 series at different cross-sectional and temporal levels, see section 4 in Makridakis et al., 2021). Moreover, promotions, holidays, and special days significantly affect sales, especially of the "Foods" product-category which is highly influenced by SNAP activities (by about 15%; Makridakis et al., 2021). In order to effectively take these effects into account, all winning submissions of the M5 competition employed methods that, along with time series data, considered various external variables as input that capture calendar-related information, special days, and promotions. Moreover, by analyzing the CV performance of their methods, many teams reported that special days and information about the month, week, and day of week, were critical for enhancing their forecasting performance. This becomes also evident in the M5 "Accuracy" challenge by comparing the performance of the Naive method to that of the seasonal Naive (sNaive), or the performance of SES and ES_bu[5] benchmarks; sNaive and ES_bu that account for seasonality were 52% and 31% more accurate than their non-seasonal counterparts, respectively. Similarly, ESX, which accounted for SNAP activities, was 6% more accurate on average than the ES_td benchmark that employed the same exponential smoothing models but without using explanatory variables. On the contrary, at the bottom aggregation level where intermittency becomes a dominant feature, these improvements decline significantly, with sNaive being 6% more accurate than Naive, ES_bu 1.8% more accurate than SES, while ESX slightly less accurate than ES_td. Therefore, we conclude that although the benefits of effectively capturing seasonality and accounting for intermittency may depend on the base forecasting method used, these are important for improving forecasting performance, both overall and at specific aggregation levels.

VIII. The vast majority of the forecasting methods will continue to underestimate uncertainty, especially for the bottom aggregation level. This underestimation will increase for longer forecasting horizons but will not apply to the top-performing methods of the competition (*this prediction only refers to probabilistic forecasts*).

For many years, forecasting methods have been considerably underestimating uncer-

[5]The same algorithm used in ES_td is employed for forecasting the product-store series of the data set (level 12). Then, the rest of the series (levels 1-11) are predicted using the bottom-up method.

tainty, i.e., providing narrower prediction intervals than required based on the confidence level assumed, especially for longer forecasting horizons (Makridakis et al., 1987). This has mainly been due to the strong assumption made by the methods, estimating uncertainty by assuming that residual errors are independent of each other and normally distributed. This assumption rarely holds true in practice, especially in retail sales forecasting applications where demand may display notable variations through time or be zero for a number of periods (Spiliotis et al., 2020b). The Global Energy Forecasting Competitions 2014 and 2017 (GEFCom2014 & GEFCom2017; Hong et al., 2016, 2019) were the first competitions where the top performing teams provided highly skilled probabilistic forecast for various quantiles in the field of electricity load, wind, solar, and energy price forecasting, followed by M4 in the area of generic time series forecasting, where the two winning methods achieved phenomenal performance in specifying the 95% prediction intervals, indicating that precise estimation of uncertainty is possible when sophisticated forecasting methods are used. The M5 continued the forecasting spring, with the winning teams of the M5 "Uncertainty" challenge outperforming, on average, the top benchmark (ARIMA) by more than 20% according to Weighted Scaled Pinball Loss (WSPL) and achieving a much better calibration.

However, the impact of the aggregation level on the precision of uncertainty estimation should not be overlooked by these summary results. For instance, although all six winning teams managed to outperform the ARIMA method at the bottom level by about 8%, from the remaining 44 methods of the top 50 performing ones, only 32 (73%) managed to outperform ARIMA, having an average improvement over the benchmark of just 2%. Considering that these methods improved overall forecasting performance across all aggregation levels by more than 15%, it becomes evident that intermittency and erraticness have a significant, negative impact on the precision of uncertainty estimation that even advanced forecasting methods may find challenging to tackle. The same is true for the impact of the forecasting horizon since, according to the results of the competition, the performance of the participating methods deteriorated for longer lead times. However, this was only the case for the three lowest aggregation levels, i.e., product, product-state, and product-store (Figure 6; Makridakis et al., 2020d). Based on the above, it is not surprising that the vast majority of the participating teams underestimated uncertainty considerably; from the 892 teams that participated in the "Uncertainty" competition, only 202 (22.6%) managed to perform better than the ARIMA benchmark (Figure 1; Makridakis et al., 2020d).

IX. Methods of innovative features will outperform the Croston's one at the lowest, product-store aggregation level (*this prediction only applies to point forecasts*).

One of the main innovations of the M5 competition was that it involved series that display intermittency and erraticness (Syntetos & Boylan, 2005; Syntetos et al., 2005). This was especially true for the lowest, product-store aggregation level of the data set that included 22,339 intermittent (73%), 5,206 lumpy (17%), 883 erratic (3%), and 2,062 smooth (7%) series (for more details on the demand classification of the M5 series, which is made based on the average inter-demand interval and demand size erraticness of the series, see section 4 in Makridakis et al., 2021). When dealing with intermittent demand data, the standard method of choice of forecasters is the Croston's method (Croston, 1972) and its variants (e.g., SBA and TSB; Spiliotis et al., 2020b). This is because, in contrast to other time series

methods, like Simple exponential smoothing, the Croston's method is less affected by zero demand occurrences and, therefore, provides more accurate estimates of future sales. As such, although simple in nature, the Croston's method has proven to be very competitive in the task of retail sales forecasting at the level of SKU.

However, the results of the M5 "Accuracy" competition indicate that methods of innovative features can be used to produce more accurate results and better support decisions related with supply-chain management. Such innovations may be realized in terms of optimization criteria considered, like the Tweedie loss function that many of the winning submissions considered to effectively account for zero sales, and approaches that mitigate uncertainty, such as ensembles of different methods trained using different pools of series or data augmentation techniques.

At the product-store level, the Croston's method reported a Weighted Root Mean Squared Scaled Error (WRMSSE) of 0.926. On the contrary, the winner of the competition (*YJ_STU*) reported a score of 0.884, while the best-performing method of that particular level (*wyz-Jack_STU*) a WRMSSE of 0.875. This is an improvement of 4.5% and 5.5%, respectively. Although the performance differences are limited in absolute values, it is widely accepted that small gains in accuracy can lead to considerable stock holding reductions and service levels improvements (Syntetos et al., 2010; Ghobbar & Friend, 2003; Pooya et al., 2019). Therefore, the potential value of such innovative methods should not be overlooked. On the other hand, given that the winning method displayed much greater improvements for the rest of the aggregation levels (an average improvement of about 50% over the Croston's method), it becomes clear that more research is required in the area of retail sales forecasting to significantly improve accuracy at low granularity levels, as well as to properly translate such improvements to monetary or better customer satisfaction benefits. For instance, as shown in Table 2, although the SBA (Syntetos & Boylan, 2005) and TSB (Teunter et al., 2011) methods, as well as temporal aggregation approaches, such as the ADIDA (Nikolopoulos et al., 2011) and iMAPA (Petropoulos & Kourentzes, 2015), are considered more appropriate than the Croston's method for dealing with intermittent demand data, the results of the M5 competition indicate that the latter can perform equally well or even better.

X. Methods generating probabilistic forecasts empirically, e.g., by utilizing bootstrapping and custom-made loss functions, will achieve significantly better results than those based on theoretically derived formulas or normality assumptions (*this prediction refers only to probabilistic forecasts*).

As noted, forecasting methods have been, for many years, underestimating uncertainty considerably by assuming that the residual errors are independent and normally distributed; a hypothesis that rarely holds in practice (Makridakis et al., 1987). On the other hand, recent algorithmic advances can be used to improve the precision of probabilistic forecasts (Salinas et al., 2020; Hewamalage et al., 2021). Therefore, forecasting approaches that build on ML/DL methods or empirically estimate uncertainty through bootstrapping techniques and simulations are expected to perform significantly better than statistical ones.

The top 50 performing methods of the M5 "Uncertainty" competition, each of which outperformed the top-performing benchmark by more than 12.5%, adopted three major approaches to estimating uncertainty: (i) directly produce probabilistic forecasts for the quantile of interest, (ii) produce point forecasts and determine factors that can be used

Table 2: Performance (WRMSSE) reported between the top 3 performing submissions of the "Accuracy" challenge, the Croston's method (CRO), and other popular benchmarks in the field of intermittent demand forecasting. The results are reported for the product-store level (level 12) and the average of the 12 levels considered in the competition. More detailed results on the performance of the submissions made in the "Accuracy" challenge, as well as the benchmarks, can be found in Makridakis et al. (2020c).

Method	Level 12	Average
Benchmarks		
Naive	1.253	1.752
sNaive	1.176	0.847
CRO	0.926	0.957
SBA	0.919	0.957
TSB	0.928	0.961
ADIDA	0.922	0.955
iMAPA	0.925	0.960
Winning submissions		
YJ_STU	0.884	0.520
Matthias	0.907	0.528
mf	0.875	0.536

to convert them into probabilistic ones, and (iii) produce point forecasts and approximate uncertainty by processing the residual errors, either empirically (simulations and empirical distributions) or theoretically (assuming a theoretical distribution). From the six winning teams, the methods ranked 1^{st} and 4^{th} fell in the first category, the method ranked 3^{rd} in the second category, while the methods ranked 2^{nd}, 5^{th}, and 7^{th} in the third one. Moreover, the top-performing teams experimented with various loss functions (pinball loss, both weighted and not, negative binomial, binomial, Tweedie, Poisson regression, and coverage optimization, among others) that in many cases changed at different aggregation levels based on which loss resulted in better forecasts according to the CV strategy considered.

In summary, all top-performing methods generated probabilistic forecasts empirically by utilizing bootstrapping, simulations, and appropriate loss functions, and achieved significantly better performance than the top-performing benchmarks. However, we will still have to classify this prediction/hypothesis as "partially correct" given that *Ka Ho_STU*, the top-performing student and the 7^{th} best submission in total, generated probabilistic forecasts by collecting the residual errors of the produced point forecasts and constructing the respective i.i.d. normal distributions, i.e., by making normality assumptions. Nevertheless, we should clarify that the literature involves some successful examples of ML/DL methods that assume particular types of distributions and rely on maximum likelihood estimations for producing probabilistic forecasts, indicating that such assumptions may be of secondary importance in practice, at least when robust, highly parameterized methods that learn from multiple series are used for their implementation (Salinas et al., 2020).

3. Overall evaluation of predictions/hypotheses

Our objective for making the ten predictions/hypotheses described in this paper was to predict the actual findings of the M5 competition as accurately as possible before it

started, rather than rationalizing its results after its completion and falling into the practice of HARKing. In this regard, current best practices and beliefs about factors that affect forecasting accuracy or are important for improving forecasting performance were evaluated for the case of retail sales forecasting and unexpected findings were identified along with interesting innovations.

We should admit that our predictions were strongly influenced by our knowledge, experience, and lessons learned from past forecasting competitions and especially the M4. Moreover, given the findings of the M4 competition, we were more confident about the potential value of ML methods, that of cross-learning and CV, and combinations, as well as the abilities and skills of the forecasting community to outperform existing methods in the field of retail sales forecasting.

Overall, our ten predictions/hypotheses managed to effectively capture the main findings of the M5 competition. Also, apart from prediction X, which was partially correct, all hypotheses were in line with the actual results of both the "Accuracy" and "Uncertainty" challenges. This conclusion indicates that approaches widely used in other fields of forecasting (e.g., combining forecasts, considering horses-for-courses, and using explanatory/exogenous variables) are also appropriate in retail sales forecasting and that recent methodological and algorithmic advances that have shown great potential in other applications (e.g., utilizing cross-learning or ML/DL methods) are also valid in demand forecasting settings. At the same time, there is evidence that current approaches, even the state-of-the-art ones that reported excellent results in M5 overall, may have still difficulties to accurately forecast high granularity data and precisely estimate the tails of the uncertainty distribution. Thus, although methods of innovative features may improve forecasting performance over existing approaches, there are still areas with room for improvement.

The only two points that we failed to predict were the extended use of external adjustments, aiming at calibrating the base forecasts and removing their bias, and the exploitation of augmentation techniques, mitigating modeling uncertainty and dealing with overfitting. Moreover, the M5 competition demonstrated that constructing diverse, homogeneous pools of time series for training global models is highly beneficial, a point that was not predicted explicitly. Similarly, our predictions referred to the superiority of ML and DL methods in general, failing to capture the extended exploitation of decision-tree-based models that dominated both forecasting challenges. Undoubtedly, these unexpected results and welcome innovations open interesting paths for future research.

Would the ten hypotheses/predictions have been as accurate if we were not familiar with the previous M competitions? The answer is a definite no. The knowledge gained from the M competitions such as the beneficial value of combining has become an indispensable best practices, while others, such as the effectiveness of simple versus sophisticated methods, has been challenged and replaced by the superiority of ML methods. The M4 competition ended the long forecasting winter, starting a new era in the time series forecasting field dominated by ML/DL methods coupled with advances in computer performance enabling the computationally hungry methods to be implemented. The M5 continued the forecasting spring by introducing methods of innovative features and practices focusing in the area of retail sales forecasting. We hope that this spring will continue and that the M6 will keep the tradition of the previous five ones to continue advancing the theory and practice of forecasting.

References

Abolghasemi, M., Hyndman, R. J., Spiliotis, E., & Bergmeir, C. (2020). Model selection in reconciling hierarchical time series. Available at: https://arxiv.org/abs/2010.10742.

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

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*, 60 – 74.

Barker, J. (2020). Machine learning in M4: What makes a good unstructured model? *International Journal of Forecasting*, *36*, 150–155.

Barrow, D., Kourentzes, N., Sandberg, R., & Niklewski, J. (2020). Automatic robust estimation for exponential smoothing: Perspectives from statistics and machine learning. *Expert Systems with Applications*, *160*, 113637.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*, 451–468.

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.

Croston, J. D. (1972). Forecasting and Stock Control for Intermittent Demands. *Journal of the Operational Research Society*, *23*, 289–303.

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

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.

Godahewa, R., Bandara, K., Webb, G. I., Smyl, S., & Bergmeir, C. (2021). Ensembles of localised models for time series forecasting. *Knowledge-Based Systems*, (p. 107518).

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, *37*, 388–427.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, *32*, 896–913.

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, *35*, 1389–1399.

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

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting*, *36*, 167–177.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. *International Journal of Production Economics*, *143*, 198–206.

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

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., & Yan, X. (2020). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. Available at: https://arxiv.org/abs/1907.00235.

Lishner, D. A. (2021). HARKing: Conceptualizations, harms, and two fundamental remedies. *Journal of Theoretical and Philosophical Psychology*, .

Ma, S., & Fildes, R. (2020). Forecasting third-party mobile payments with implications for customer flow prediction. *International Journal of Forecasting*, *36*, 739–760.

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*, 111–128.

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., 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., 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., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the m-competition. *International Journal of Forecasting*, *3*, 489–508.

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. (2020a). Predicting/hypothesizing the findings of the M4 Competition. *International Journal of Forecasting*, *36*, 29–36.

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

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020c). *The M5 Accuracy competition: Results, findings and conclusions*. Working paper available at: https://drive.google.com/drive/u/1/folders/1S6IaHDohF4qalWsx9AABsIG1filB5V0q.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021). The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, .

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

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.

Montero-Manso, P., & Hyndman, R. J. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, *37*, 1632–1653.

Nenova, Z. D., & May, J. H. (2016). Determining an optimal hierarchical forecasting model based on the characteristics of the data set. *Journal of Operations Management*, *44*, 62–68.

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.

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: neural basis expansion analysis for interpretable time series forecasting. *CoRR*, *abs/1905.10437*.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, *37*, 343–359.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*, 545–554.

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

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.

Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc

Hypothesizing Harm Scientific Progress. *Review of General Psychology*, *21*, 308–320.
Rubin, M. (2019). The Costs of HARKing. *The British Journal for the Philosophy of Science*, .
Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*, 1181–1191.
Seeger, M. W., Salinas, D., & Flunkert, V. (2016). Bayesian intermittent demand forecasting for large inventories. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 29.
Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2020). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, . Accepted.
Sen, R., Yu, H.-F., & Dhillon, I. S. (2019). Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 32.
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., Abolghasemi, M., Hyndman, R. J., Petropoulos, F., & Assimakopoulos, V. (2021). Hierarchical forecast reconciliation with machine learning. *Applied Soft Computing*, *112*, 107756.
Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020a). Are forecasting competitions data representative of the reality? *International Journal of Forecasting*, *36*, 37–53.
Spiliotis, E., Makridakis, S., Semenoglou, A.-A., & Assimakopoulos, V. (2020b). Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research: An International Journal*, (pp. 1–25).
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. (2020c). Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Applied Energy*, *261*, 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.
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.
Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*, 804–819.