# The M5 competition: Background, organization and implementation

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 M5 competition follows the previous four M competitions, whose purpose is to learn from empirical evidence how to improve forecasting performance and advance the theory and practice of forecasting. M5 focused on a retail sales forecasting application with the objective to produce the most accurate point forecasts for 42,840 time series that represent the hierarchical unit sales of the largest retail company in the world, Walmart, as well as to provide the most accurate estimates of the uncertainty of these forecasts. Hence, the competition consisted of two parallel challenges, namely the "Accuracy" and "Uncertainty" forecasting competitions. M5 extended the results of the previous M competitions by: (a) significantly expanding the number of participating methods, especially those in the category of machine learning, (b) evaluating the performance of the uncertainty distribution along with point forecast accuracy, (c) including exogenous/explanatory variables in addition to the time series data, (d) using grouped, correlated time series, and (e) focusing on series that display intermittency. This paper describes the background, organization, and implementations of the competition, also presenting the data used and its characteristics. Consequently, it serves as introductory material to the results of the two forecasting challenges to facilitate their understanding.

*Keywords:* Forecasting Competitions, M Competitions, Accuracy, Uncertainty, Time Series, Retail Sales Forecasting

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

## 1. Introduction

The aim of forecasting competitions is to empirically evaluate the performance of existing and new forecasting methods, allowing the equivalent of experimentation widely used in hard sciences. (Hyndman, 2020; Makridakis et al., 2021). From these, the M competitions (Makridakis et al., 1982, 1993; Makridakis & Hibon, 2000; Makridakis et al., 2020d) are probably the most influential and widely cited in the field of forecasting, the most recent being the M5 competition that took place in the period of March-June, 2020.

The objective of the M5 competition was to produce the most accurate point forecasts for 42,840 time series that represent the hierarchical unit sales of the largest retail company in the world by revenue, Walmart, as well as to provide the most accurate estimates of the uncertainty around these forecasts. Thus, the M5 competition consisted of two parallel challenges, namely the "Accuracy" and the "Uncertainty" ones, whose results, findings, and conclusions are described in detail in Makridakis et al. (2020e) and Makridakis et al. (2020f), respectively. Given that both forecasting challenges share the same background, organization, implementation, and data, this paper provides a description of them, serving as introductory material to these two papers. Following the various comments and discussions made about the M4 competition (Fry & Brundage, 2020; Önkal, 2020; Gilliland, 2020; Fildes, 2020; Hyndman, 2020; Makridakis et al., 2020c), the M5 was designed and conducted with the aim of addressing the concerns raised and expanding its objectives in several directions, as follows:.

- The competition was hosted by Kaggle[1], the largest online community of data scientists and machine learning (ML) practitioners who compete and provide solutions for various tasks, including forecasting (Bojer & Meldgaard, 2020). As such, the number of participating teams[2] increased significantly. These teams also focused on state-of-the-art methods that can be classified as ML or "unstructured" (Januschowski et al., 2020; Barker, 2020).

- The competition required the estimation of the uncertainty distribution of the predicted series by considering nine different quantiles. Although the M4 competition included the estimation of 95% prediction intervals, this is the first M competition to evaluate forecast distribution performance along with point forecast accuracy, focusing on both the middle of the distribution and its tails.

- In contrast to the previous M competitions, teams were provided with exogenous/explanatory variables, besides time series data, that could be used to improve forecasting performance (Huang et al., 2014; Ma & Fildes, 2017; Ma et al., 2016).

[1] https://www.kaggle.com/competitions

[2] A team may consist of multiple members, or a single participant. The maximum team size in M5 was five members.

- Instead of forecasting mostly unrelated series (Spiliotis et al., 2020a), M5 consisted of grouped, highly correlated ones, organized in a hierarchical, cross-sectional structure, thus representing the forecasting set-up of a typical retail company.

- The competition involved series that display intermittency, i.e. sporadic demand including many zeros (Syntetos & Boylan, 2005; Syntetos et al., 2005). This kind of series, although difficult to predict with conventional forecasting methods like the ones utilized in the previous M competitions, are typical when forecasting unit retail sales at a store or product level (Spiliotis et al., 2020b). Therefore, identifying accurate methods to predict such series can be highly beneficial for numerous retail companies. (Ghobbar & Friend, 2003; Syntetos et al., 2010; Pooya et al., 2019).

We should clarify that although some of the design attributes and principles introduced in M5 have been previously considered in past forecasting competitions (Makridakis et al., 2021), this is the first time they are brought together in retail sales forecasting settings. For instance, although GEFCom2017 (Hong et al., 2019) was probably the first competition to involve a probabilistic hierarchical demand forecasting challenge, its focus has been on the energy sector. Similarly, although Kaggle has previously hosted some other retail sales forecasting competitions (Bojer & Meldgaard, 2020), their focus has been on point forecasts, produced for particular cross-sectional levels of interest (e.g. store or product-store level).

Petropoulos et al. (2020) provide a non-systematic review on the forecasting methods used for intermittent demand and count data, including parametric, non-parametric, and classification approaches (section 2.7), and present their application in indicative operations and supply chain management settings (section 3.2). Fildes et al. (2019) review the research done on forecasting retail demand, present the key forecasting problems that retailers face, discuss the factors that influence demand, and identify best practices by evaluating evidence on comparative forecasting accuracy. The authors conclude that although causal models outperform simple benchmarks widely used in the field, there is still little evidence that machine learning methods can provide superior results. Similarly, they find that inadequate information is available for assessing the effectiveness of the approaches used for forecasting new products. Thus, they motivate future research and suggest practitioners and researchers to focus on the design of dynamic, scalable, and intuitive approaches that can effectively deal with rapid structural changes in the development of online competition, shopping habits, and other market characteristics, as well as on the advancement of probabilistic forecasting methods which remain under-investigated. The results of the M5 competition contribute in these directions, providing new evidence on the performance of existing and novel forecasting methods in retail sales forecasting settings, thus motivating future research in the field.

The rest of the paper consists of four sections. Section 2 describes the background of the M5 competition. Section 3 presents the organization and implementation of the competition, including its running dates and two phases. Section 4 presents and describes the data used in the competition. The last section concludes the paper and makes some recommendations about how the M5 data set should be used.

## 2. Background

Forecasting competitions are vital for the objective evaluation of existing forecasting methods, judging the accuracy of new ones, as well as providing empirical evidence on how to advance the theory and practice of forecasting. Each M competition has introduced some new features or data sets that can aid future research, addressing some limitations of the previous ones while focusing on different applications and aspects of forecasting performance, both in terms of point forecasts and uncertainty estimation.

Hyndman (2020) highlights the benefits and contributions of forecasting competitions, including the M ones, and illustrates the desirable features that future competitions should have. He concludes that: (i) a wider range of benchmarks and data sets that are regularly updated is desired in order to mitigate overfitting in published data used to evaluate forecasting methods, (ii) future competitions should clearly define the domain to which they apply, (iii) objective measures that are based on well-recognized attributes of the forecast distribution should be used, (iv) forecast distribution performance should be assessed along with point forecast accuracy, (v) large-scale multivariate time series forecasting should be considered to exploit possible cross-correlations between the series, (vi) high frequency data, such as hourly, daily, and weekly should be introduced to investigate how multiple seasonal patterns and irregularly spaced observations could be properly handled, as well as how data collected from sensors could be optimally used, and finally, (vii) exogenous/explanatory variables should be provided along with time series data to determine whether they contribute to more accurate forecasts. Similar remarks were made by the discussants and commentators of the M4 competition, stressing the need for more representative, higher frequency, hierarchically structured data that are accompanied by exogenous/explanatory variables (Fry & Brundage, 2020; Önkal, 2020; Gilliland, 2020; Fildes, 2020). More recently, Makridakis et al. (2021) have provided suggestions about the principles and design attributes to be included in future competitions in terms of scope, diversity and representativeness, data structure, granularity, and availability, forecasting horizon, evaluation setup, performance measurement, and benchmarks, putting a particular emphasis on learning as much as possible from their implementation.

The M5 competition tried to address these concerns and suggestions by introducing the following innovative features:

- A large data set of 42,840 series was used, along with several benchmarks. In this way, existing and new forecasting methods could be objectively evaluated and the results of previous studies effectively tested for replicability.

- The competition focused on a specific forecasting application which was to accurately predict the daily unit sales of retail stores across various locations and product categories, as well as precisely estimate the uncertainty distribution of the predicted value.

- Objective measures were utilized to evaluate forecasting performance. In the "Accuracy" competition, the measure (Weighted Root Mean Squared Scaled Error, WRMSSE) evaluated the deviation of the point forecasts around the mean of the realized values of the series being predicted. In the "Uncertainty" competition the measure (Weighted Scaled Pinball Loss, WSPL) evaluated the deviation of the probabilistic forecast around the realized value, taking into account the respective probability level.

- Forecast distribution performance was assessed along with point forecast accuracy by considering nine different quantiles that can sufficiently approximate the complete distribution of future sales, namely the median and the 50%, 67%, 95%, and 99% prediction intervals.

- The series of the data set were grouped and highly correlated, thus enabling the utilization of multivariate and "cross-learning" methods.

- The data set involved daily data which requires accounting for multiple seasonal patterns, special days, and holidays.

- The data set included exogenous/explanatory variables, such as product prices, promotions, and special days.

The M5 competition was initially announced at the end of December, 2019, first on the Makridakis Open Forecasting Center (MOFC) website[3] and then on the International Institute of Forecasters (IIF) blog[4]. In addition, just as was done in the M4, invitation emails were sent to all those who had participated in previous forecasting competitions and forecasting conferences, as well as to those who had published articles in respected journals in the field of forecasting (for more information about the invitations, please see Makridakis et al., 2020d). Social media (LinkedIn, Twitter, and Facebook) were also used to promote the competition.

The competition started on March 3rd, 2020, when the initial train set became available, and ended on June 30th, 2020, when the final leaderboard was announced by Kaggle. The rules of the competition, prizes, and additional details were all made available on Kaggle and the M5 competition's website[5]. The preliminary results of the competition were presented virtually on October 28th, 2020, at the 40th *International Symposium on Forecasting*, with the final results and winning methods to be presented at the M5 conference in New York on December 6th and 7th, 2021.

[3] https://mofc.unic.ac.cy/

[4] www.forecasters.org

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

Like all the previous M competitions, M5 was completely open, encouraging the participation of both academics and practitioners and ensuring fairness and objectivity, while also emphasizing that each team was free to utilize its own preferred method. Moreover, like the M4, teams were encouraged to submit the code used for producing their forecasts, as well as a description of their methods, thus promoting reproducibility and replicability (Boylan et al., 2015; Makridakis et al., 2018a). In fact, reproducing the results of the submissions was a prerequisite for collecting any prize and, therefore, all the winning teams of the competition provided their code to the organizers, which were verified for their correctness and then uploaded to the M5 GitHub repository[6] for anyone interested to use. The public discussions[7] (695 topics; 602 in the "Accuracy" and 93 in the "Uncertainty") and notebooks (754 scripts; 677 in the "Accuracy" and 77 in the "Uncertainty") published on Kaggle, further facilitated replicability and exchange of information, provided fruitful insights, and encouraged the exchange of ideas among the participants and forecasting community in general. Nevertheless, we should note that although the organizers of the competition sent several emails to the top performing teams (beyond those who won the prizes) of the competition to request information for their submissions, including the code used for producing their forecasts and a description of their methods, only few replied to our request. Specifically, in the "Accuracy" and "Uncertainty" challenges, only 17 and 15 of the 50 top performing methods shared information about their forecasting approaches, respectively. Similarly, although some of the teams ranked below the 50th position decided to demonstrate their forecasting approaches in public notebooks on Kaggle, they did not always provide their complete solutions which, in many cases, missed a detailed description.

In addition, the M5 competition continued the major innovation of the M4, which was to predict/hypothesize its findings prior to its completion (Makridakis et al., 2020a). Rather than rationalizing its results in a post-hoc reasoning, our ten predictions/hypotheses, submitted to the EiC of *International Journal of Forecasting* five days before launching the competition, demonstrated our expectations of its findings. We have now evaluated our predictions/hypotheses in a separate paper (Makridakis et al., 2020b), reflecting on our successes and errors, as well as findings that our predictions/hypotheses missed or did not specify explicitly.

The competition also placed particular emphasis on benchmarking, considering a variety of methods, both traditional and state-of-the-art, that can be classified as statistical, ML, and combinations. In the first three M competitions (Makridakis et al., 1982, 1993; Makridakis & Hibon, 2000), and for many years in the forecasting literature (Bates & Granger, 1969; Claeskens et al., 2016), combinations and relatively simple methods were regarded as being at least as accurate as sophisticated ones. The M4 competition (Makridakis et al., 2020d), however, despite confirming the value of combining, indicated that more sophisticated, ML methods could provide significantly more accurate results. These findings indicate that comparing the methods submitted in the M5 competition with various benchmarks would allow us to validate the findings

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

[7] The numbers reported refer to the topics and notebooks published at the end of the competition

of the previous competitions and identify possible areas for improvements.

The benchmarks of the competition were selected based on their popularity, availability, ease of use, and computational requirements, which are relatively low. A description of the benchmarks used in the "Accuracy" and "Uncertainty" forecasting competitions is provided in the Appendices of the Makridakis et al. (2020f) and Makridakis et al. (2020e) studies respectively, along with their forecasting performance. Moreover, the code used for implementing the benchmarks is publicly available in the M5 GitHub repository. Note that most of the benchmarks considered in the competition were previously tested in a different data set, involving the daily product sales of a large Greek retail company (Spiliotis et al., 2020b). The main conclusion of this study was that some ML methods provided less biased and more accurate forecasts than well-established, statistical methods, like the Croston's method and its variants, especially when "cross-learning" approaches were considered, thus confirming the main finding of M4.

## 3. Organization and implementation

The M5 competition began on March 3rd, 2020, when the initial train data set became available to download on the Kaggle platform. The rules and information regarding the competition were likewise posted, on both Kaggle and the M5 competition's website. The competition ended on June 30th, 2020, when the final leaderboard was posted by Kaggle. The deadline for accepting the competition rules (entry deadline) and joining or merging teams (team merger deadline) was June 23rd. All teams were free to participate in both forecasting challenges, as long as their individual members were part of a single team per competition. Note that it was not allowed to use multiple accounts to make multiple submissions. Also, privately sharing code or data outside of teams was not permitted. However, sharing code was allowed if it was made available to all participants on the forums.

Chronologically, the competition was divided into two phases; the "validation" and "test" phase. In the "validation" phase, the participating teams were receiving feedback through a live leaderboard, exchanging ideas and insights with the rest of the community about the hidden 28 day data they were predicting, while at the "test" phase no feedback about their performance was provided, as is the case in real-life forecasting.

The "validation" phase took place from March 3rd, 2020 to May 31st of the same year. During this phase, the teams were allowed to train their forecasting methods with the data initially provided by the organizers (consisting of 1,913 days) and validate the performance of their methods using a hidden sample of 28 days (equal to the forecasting horizon considered by the competition), not made publicly available. This sample corresponded to the four weeks succeeding the initial train set, i.e. days 1,914 to 1,941. By submitting their forecasts to the Kaggle platform (a maximum of five entries per day), the teams were informed about the performance of their submission, which was then published onto Kaggle's real-time leaderboard. Given this feature, teams were allowed to effectively revise and resubmit their forecasts by learning from the received feedback (Athanasopoulos & Hyndman, 2011). Note that, ideally, and in order to avoid overfitting and construct robust, generalized forecasting solutions, the leaderboard should be used for assessing the methods developed in terms of forecasting algorithms and data-handling techniques utilized, and not for indirectly optimizing their settings in regard to hyper-parameters used. For instance, the participants could exploit the leaderboard scores to decide whether a decision-tree-based model provides on average more accurate results than a statistical one, but not for explicitly tuning the learning rate of the former as the "optimal" value of such a hyper-parameter may be subject to the short time period considered for performing the validation. Similarly, the participants could benefit from the live feedback received to determine the explanatory variables to be used by their causal models but not for externally adjusting the level of the forecasts produced in an unstructured or unjustified fashion (e.g. using arbitrarily selected multipliers). Some of the discussions made on Kaggle on this topic[8] highlight the negative impact of misused leaderboard scores and demonstrate that ad-hoc adjustments that may temporarily lead to better leaderboard ranks due to overfitting, can in turn result in poor post-sample performance.

After the end of the "validation" phase, i.e. June 1$^{\text{st}}$, 2020, the teams were provided with the actual values of the 28 days of data used for assessing their performance during the "validation" phase. They were then asked to re-estimate or adjust their forecasting methods (if needed), in order to submit their final forecasts for the following 28 days, i.e. the test data used for the final evaluation of the teams. During this time, there was no leaderboard, meaning that no feedback was given to the teams about their actual performance after submitting their forecasts. Therefore, although the teams were free to (re)submit their forecasts any time they wished (a maximum of five entries per day) during the "test" phase, they were not aware of their absolute, nor their relative, performance.

The final ranks of the teams were only disclosed at the end of competition, when the test data was made available. For their evaluation, each team had to select a single set of forecasts (one submission). If no particular forecasts were selected, the ones with the highest performance during the "validation" phase were automatically selected by the system. This was done in order for the competition to simulate reality as closely as possible, given that in real life forecasters do not know the future and they have to provide a single set of forecasts which they believe will simulate the future as accurately as possible.

At this point, we should note that making submissions during the "validation" phase of the competition was completely optional and teams were free to decide whether they were going to exploit the public leaderboard to validate their methods, or their own, privately constructed cross-validation (CV) tests (Tashman, 2000). However, despite the preferences of each team, assessing the post-sample performance of the developed methods effectively was of critical importance for performing well in the M5 competition. This conclusion is discussed in Finding 6 of the "Accuracy" competition (Makridakis et al., 2020e) and Finding

[8] https://www.kaggle.com/c/m5-forecasting-accuracy/discussion/163190

4 of the "Uncertainty" one (Makridakis et al., 2020f), suggesting that CV can provide useful insights about the models that are expected to derive more accurate forecasts, the optimal values of their parameters, and the exogenous/explanatory variables that they should be provided with as input. This is particularly true when dealing with flexible ML methods where the number of features that can potentially be used as input, the training approaches that can possibly be adopted, and the hyper-parameters that can be accordingly adjusted, are huge (Ma et al., 2016; Makridakis et al., 2018b). In addition, CV strategies, as well as other modeling features like information criteria (Sakamoto et al., 1986), can become useful for avoiding overfitting and mitigating data, parameter, and model uncertainty (Petropoulos et al., 2018).

## 4. Data

The M5 data set, generously made available by Walmart, involves the unit sales of various products sold in the USA, organized in the form of grouped time series. More specifically, the data set involves the unit sales of 3,049 products, classified into three product categories (*Hobbies*, *Foods*, and *Household*), and seven product departments in which the above mentioned categories are disaggregated. The products are sold across 10 stores, located in three states (California - *CA*, Texas - *TX*, and Wisconsin - *WI*). In this respect, the most disaggregated data, i.e. product-store unit sales, can be grouped based on either location (store and state) or product-related information (department and category), as shown in Figure 1. We should clarify that the states and stores were carefully selected by Walmart with the objective to represent selling locations of different characteristics, shopping habits, and dynamics. Similarly, product categories and departments were selected with the aim to represent both consumables and durable goods, as well as fast-moving and slow-moving products.

Given that multiple meaningful hierarchies can be constructed from the M5 data, the organizers decided to consider all possible cross-sectional levels of aggregation for the evaluation, as shown in Table 1. Although the identifiers of the various levels (level id) do not indicate an actual hierarchical structure, they facilitate reference, also highlighting the extent of aggregation that takes place: high levels of aggregation generally correspond to low identification numbers (e.g. levels 1 to 5), while low levels of aggregation to higher identification numbers (e.g. levels 10 to 12).

The data is daily and covers the period from 2011-01-29 to 2016-06-19 (1,969 days or approximately 5.4 years). As described in the previous subsection, the first 1,913 days of data (2011-01-29 to 2016-04-24) were initially provided to the participating teams as a train set, days 1,914 to 1941 (2016-04-25 to 2016-05-22) served as a validation set, while the remaining 28 days, i.e. 1,942 to 1,969 (2016-05-23 to 2016-06-19) were used as a test set.

The M5 competition data set also involved exogenous/explanatory variables, including calendar-related information, selling prices, and promotional activities. Thus, apart from the past unit sales of the products

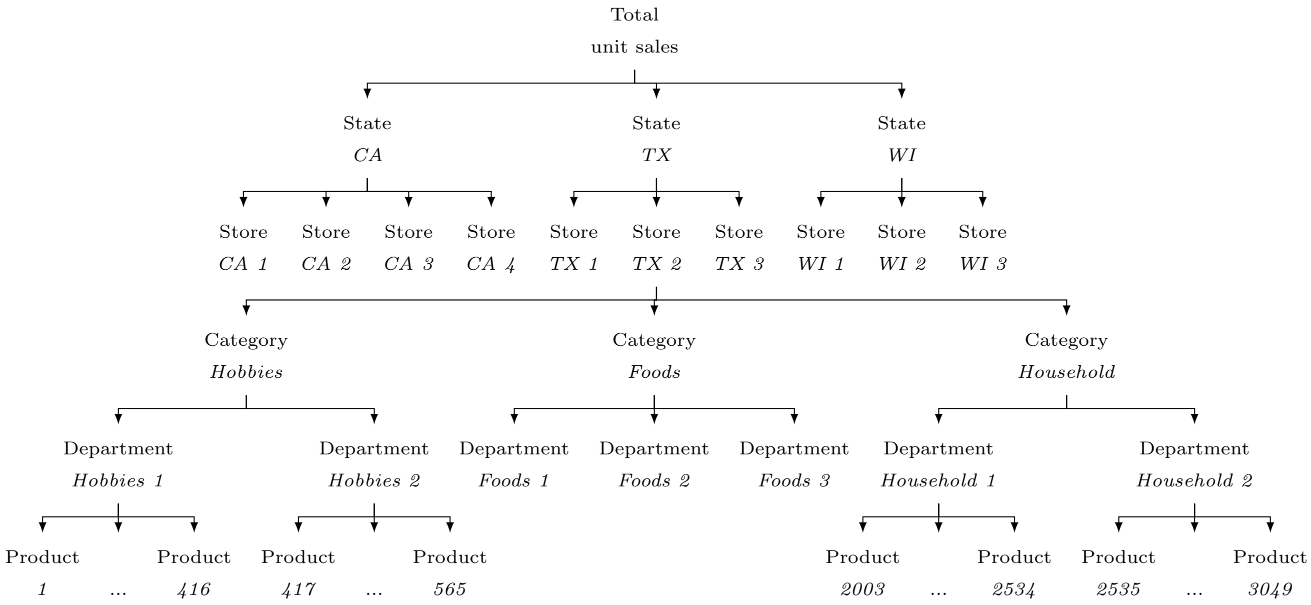

Figure 1: Grouped time series used in the M5 competition. The data can be aggregated in 12 different levels using either location (state and store) or product-related information (category and department).

and the corresponding timestamps (e.g. date, weekday, week number, month, and year), there was also information available about:

- Special days and holidays (e.g. Super Bowl, Valentine's Day, and Orthodox Easter), organized into four classes, namely "Sporting", "Cultural", "National", and "Religious". Special days account for about 8% of the days included in the data set. From these days, "Sporting", "Cultural", "National", and "Religious" events account for about 11%, 23%, 32%, and 34% of the cases, respectively.
- Selling prices, provided on a week-store level (average across seven days). If not available, this means that the product was not sold during the week examined. Although prices are constant on a weekly basis, they may change with time.
- SNAP [9] activities that serve as promotions. This is a binary variable (0 or 1) indicating whether the stores of *CA*, *TX* or *WI* allow SNAP purchases on the date examined; 1 indicates that SNAP

[9] The United States federal government provides a nutrition assistance benefit called the Supplement Nutrition Assistance Program (SNAP). SNAP provides low income families and individuals with an Electronic Benefits Transfer debit card to purchase food products. In many states, the monetary benefits are dispersed to people across 10 days of the month and on each of these days 1/10 of the people will receive the benefit on their card. More information about the SNAP program can be found here: https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program

Table 1: Number of M5 series per aggregation level.

| Level id | Level description | Aggregation level | Number of series |
|---|---|---|---|
| 1 | Unit sales of all products, aggregated for all stores/states | Total | 1 |
| 2 | Unit sales of all products, aggregated for each state | State | 3 |
| 3 | Unit sales of all products, aggregated for each store | Store | 10 |
| 4 | Unit sales of all products, aggregated for each category | Category | 3 |
| 5 | Unit sales of all products, aggregated for each department | Department | 7 |
| 6 | Unit sales of all products, aggregated for each state and category | State-Category | 9 |
| 7 | Unit sales of all products, aggregated for each state and department | State-Department | 21 |
| 8 | Unit sales of all products, aggregated for each store and category | Store-Category | 30 |
| 9 | Unit sales of all products, aggregated for each store and department | Store-Department | 70 |
| 10 | Unit sales of product $i$, aggregated for all stores/States | Product | 3,049 |
| 11 | Unit sales of product $i$, aggregated for each state | Product-State | 9,147 |
| 12 | Unit sales of product $i$, aggregated for each store | Product-Store | 30,490 |
| Total | | | 42,840 |

purchases are allowed. Although the dates that the SNAP activities take place are different between the three states, all of them have 10 days per month for which SNAP purchases are allowed, meaning that about 33% of the days are affected by these activities.

Other external data were not provided but teams were free to use additional, publicly available data if they wished, provided that (i) this data was made available to use by all participants of the competition at no cost, (ii) the existence of this data was made publicly known in the official competition forum prior to the entry deadline, and (iii) this data did not leak any true information about the forecast period, i.e. referred only to information that would actually be known at the time the required forecasts would have been originally produced. Although some teams provided such external data[10], their use was limited and none of the top-performing teams considered them in their solutions. This included game dates for the NBA finals, natural disaster declarations, gasoline prices, as well as population, unemployment rate, and per capita income information.

The forecasting horizon (forecasts 28 days ahead) was determined based on the nature of the decisions that companies typically support when forecasting data similar to that of the M5, i.e. daily series disaggregated in various locations and product categories. Consequently, the test set was randomly chosen by the organizers from the original data set provided by Walmart (around six years of data), with the only restrictions being that (i) more than five years of data should be available for training, and (ii) at least two special days should be included in the validation and test set to account for possible deviations in sales.

[10] Available at: https://www.kaggle.com/c/m5-forecasting-accuracy/discussion/133468#888822

Therefore, the test set involved three special days, namely Memorial day, part of Ramadan, and the NBA finals, while the validation set involved Pesach, Orthodox Easter, Cinco De Mayo, and Mother's day.

Figure 2 presents the series that refer to the unit sales reported overall (level 1), by state (level 2), and by product category (level 4). As seen, all series are characterized by trend and strong seasonal patterns, identified both at daily and monthly levels. Saturdays and Sundays account for approximately 20% more sales than the average weekday, followed by Fridays and Mondays. The average deviations of that seasonal pattern are relatively small and mainly refer to the sales of *WI* and *Hobbies* which are lower than the rest of the series on Sundays. With the exception of the *Hobbies* category, monthly seasonality is also strong and consistent across the series. July and August account for about 6% more sales than the average weekday, followed by March, whereas February and November are the worst-selling months of the year. Therefore, seasonality is expected to be the most critical factor for improving overall forecasting performance, especially if effectively captured at different aggregation levels.

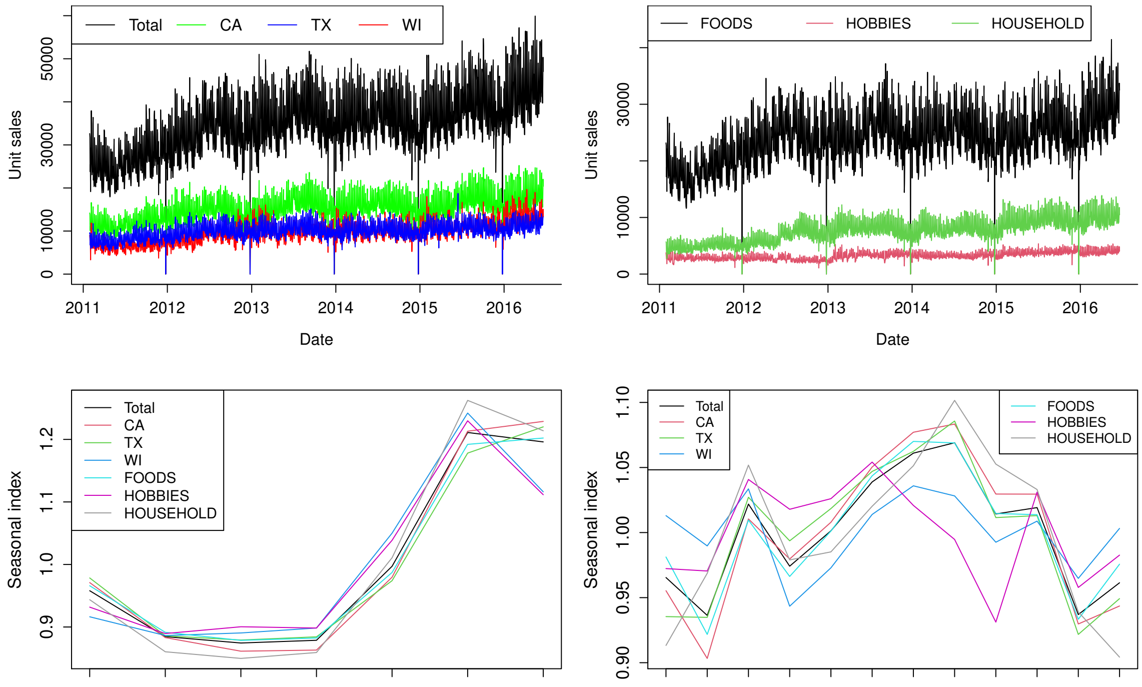

Figure 2: Unit sales series in the M5 competition. Top left: Daily unit sales reported overall (level 1) and by state (level 2); Top right: Daily unit sales reported by product category (level 4); Bottom left: Weekly seasonality (seasonal indexes) estimated for the abovementioned series using the classical multiplicative decomposition (Makridakis et al., 2020d); Bottom right: Monthly seasonality estimated for the abovementioned series.

Figure 3 presents the series that refer to the unit sales reported by store (level 3) and by product department (level 5). Observe that some stores account for significantly lower or higher proportions of the overall sales, with some of them also displaying structural changes (e.g. *WI 1* and *WI 2*). The same is true for the case of the departments. As seen, *Foods 3* accounts for the largest proportion of sales, followed by *Household 1*, *Foods 2*, and *Hobbies 1*. Consequently, developing forecasting models that are specialized for predicting different stores, product categories, and product departments should be expected to be an effective strategy for capturing the particularities of these series and enhancing overall forecasting performance.

Since it is impossible to visualize every single series of the rest of the aggregation levels, Figure 4 presents some indicative series of the product-store level which accounts for the vast majority of the series that must be predicted in the competition. As seen, each series displays its own characteristics, some of them being smooth, while others displaying intermittency and erraticness (Syntetos & Boylan, 2005). Moreover, some products started being sold at the beginning of 2011, while others in 2014 or even later. In addition, the series display major differences in terms of demand size, some ranging from one to four units, while others from five to 30. Finally, note that none of the series display strong seasonality or trend, meaning that even if present, it would be challenging to effectively extract in order to enhance forecasting performance. These time series characteristics were expected to significantly complicate the whole forecasting process, requiring specialized models to allow for accurate predictions.

Following the suggestions of Syntetos & Boylan (2005), for each series of the most disaggregated level of the competition we compute the $CV^2$ (squared coefficient of variation of the demand when it occurs) and $ADI$ (average inter-demand interval) values and then use the thresholds proposed by Syntetos et al. (2005) to categorize them (0.5 and 4/3, respectively). $CV^2$ represents demand size erraticness, while $ADI$ intermittency, thus allowing an intuitive categorization of the data. Note that although the thresholds proposed by Syntetos & Boylan (2005) were originally used to facilitate comparisons of specific forecasting methods, they have been applied more generally since then for distinguishing between the four time series categories. Figure 5 presents the 30,490 product-store series of the M5 in a logarithmic scaled $CV^2$-$ADI$ scatter-plot to facilitate visualization, since a few series display extreme $CV^2$ and $ADI$ values. In total, the data set includes 22,339 intermittent (73%), 5,206 lumpy (17%), 883 erratic (3%), and 2,062 smooth (7%) series. Considering that most of the series display strong intermittency and erraticness, we expect the extrapolation of the product-store series of the competition to be a challenging task. A summary of the main characteristics of the product-store series, including information about their length, demand size, intermittency, erraticness, and spread, as well as average price value, variation, and spread, is provided in Table 2. As seen, prices do not change frequently over time, but when they do they vary on average by about 14%, a factor that is expected to significantly affect sales. Note also that, according to Figure 5, a limited number of series is identified either as intermittent or erratic at the product and product-state levels, while all series at the rest of the levels are characterized as smooth.

Regarding the effect of SNAP purchases, we find that, on average, total daily unit sales are approximately

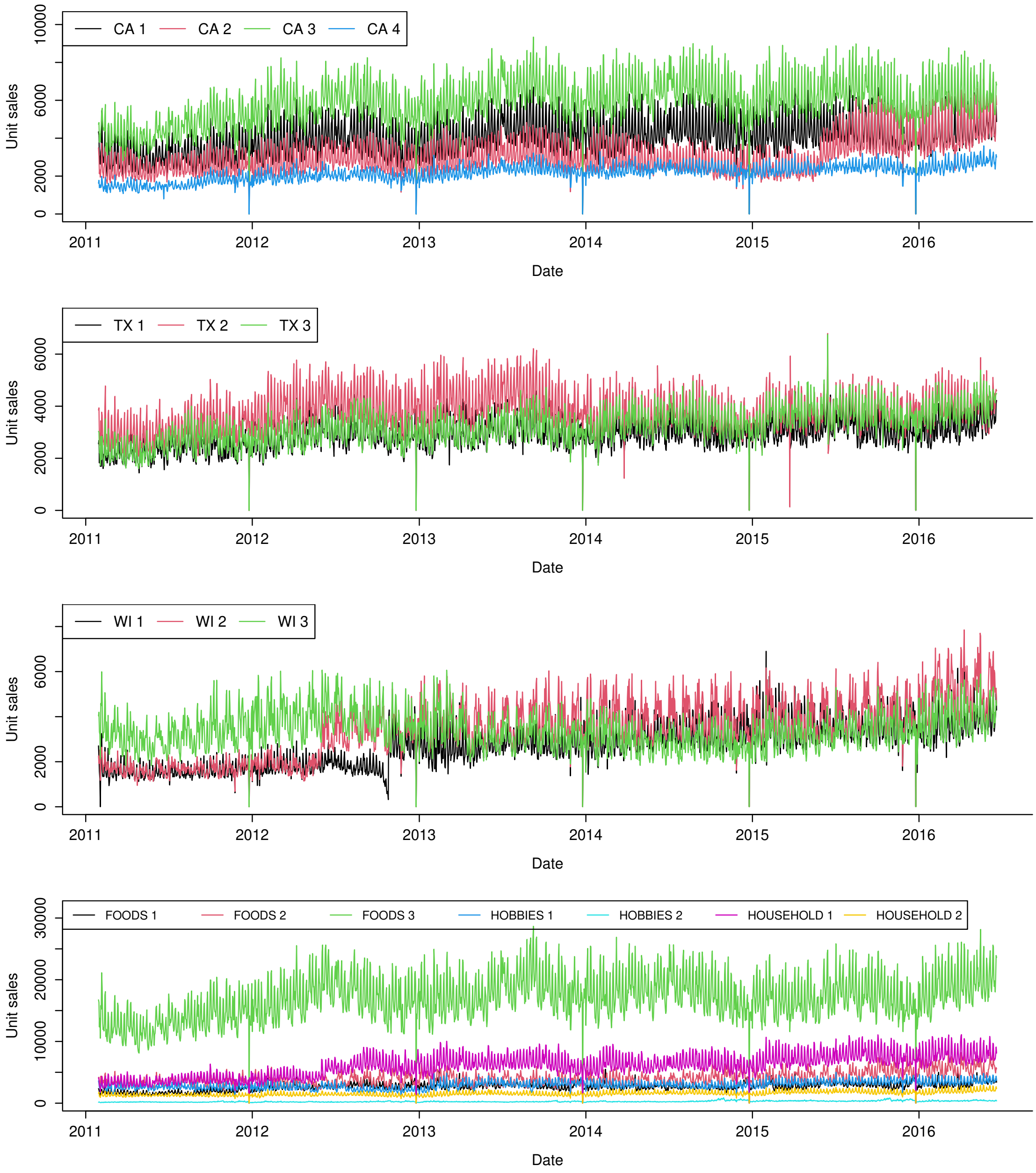

Figure 3: Daily unit sales reported by store (level 3) and department (level 5).

11% higher during SNAP activities, when the sales of *Foods*, *Household*, and *Hobbies* categories increase by about 15%, 4%, and 2%, respectively. Moreover, the number of states that are eligible for SNAP purchases is important for determining total sales: when a single state is eligible the sales increase by 8%, while when all three states are eligible by 14%. *WI* is significantly affected by SNAP purchases, with its average daily unit sales increasing by about 22% when compared to non-SNAP days. Similarly, the units sales of *TX* and

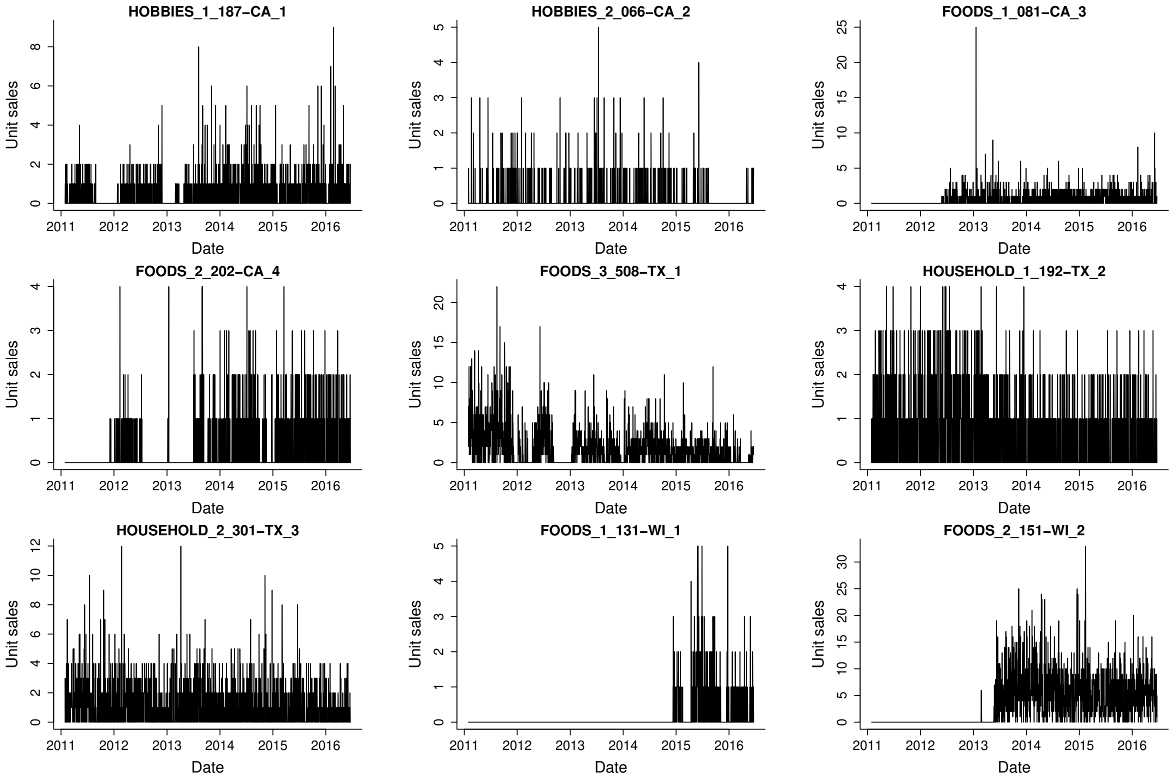

Figure 4: A sample of series reporting the daily unit sales recorded at product-store level. Series "HOBBIES_1_187-CA_1" refers to the sales recorded for the 187th product of the *Hobbies 1* department sold at the first store of *CA*, "HOBBIES_2_066-CA_2" to the sales recorded for the 66th product of the *Hobbies 2* department sold at the second store of *CA*, and so on.

*CA* increase by about 12% and 8%, respectively. Note, however, that not all stores are similarly affected by SNAP purchases, even when located in the same state. For instance, *WI 2* and *WI 3* sell about 33% and 25% more products during these events, but *WI 1* just 5%. This finding indicates that the proposed forecasting methods should take into account state, product category, and store-related information, to accurately predict unit sales.

In contrast to SNAP activities, special days and holidays display a less significant effect on average daily unit sales, which is slightly negative (about 4% less sales in total when compared to typical days). This is true for all states and product categories, especially in *WI* and for the *Hobbies* and *Household* products. Note, however, that these findings may be a consequence of lag effects, i.e. instances where customers buy more/less products the day before or after the events, which is likely to happen in practice. Moreover, the findings may be subject to the "constant low prices" strategy adopted by Walmart for attracting customers, meaning that different conclusions on the impact of special events and holidays could be true for other retail firms that drive their sales through promotions and discounts. The effect of "Cultural" and "Religious" events is negligible, while "Sporting" and "National" events lead on average to 5% more and 15% less sales,

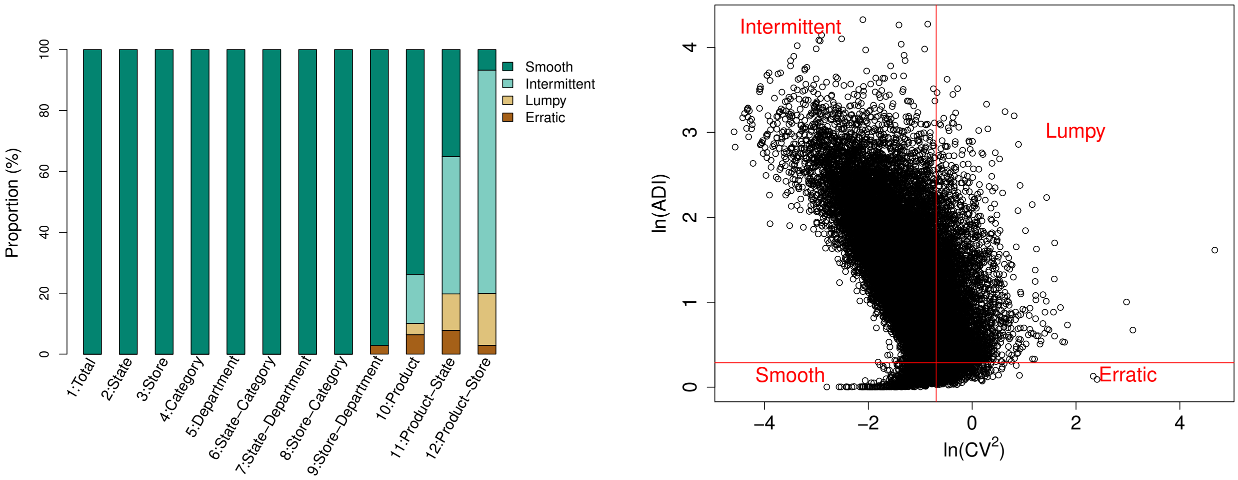

Figure 5: Demand classification of the M5 series based on their intermittency ($ADI$) and erraticness ($CV^2$). Left: The 42,840 series of the data set are separated per aggregation level into intermittent, lumpy, erratic, and smooth. Right: The 30,490 product-store series of the data set are presented in a logarithmic scaled $ADI$-$CV^2$ scatter-plot to facilitate visualization. The complete data set includes 26,956 intermittent, 6,416 lumpy, 1,791 erratic, and 7,677 smooth series. Accordingly, the product-store level includes 22,339 intermittent, 5,206 lumpy, 883 erratic, and 2,062 smooth series.

respectively. In the case of the “Sporting” events, the increase in sales is mostly driven by the *Foods* category, while in the case of the “National” events the decrease in sales is similarly distributed among the product categories. Hence, we expect forecasting methods that effectively take into account the effect of “Sporting” and “National” events to provide more accurate forecasts than the ones that do not consider their impact. On the other hand, since the number of special days is relatively limited across the year, we do not expect them to be a critical factor for improving overall forecasting performance, at least when compared to SNAP events.

As a final step in this exploratory data analysis, we compare the share of some key series of the data set on the overall sales, both unit and dollar ones. Specifically, we estimate the proportion of each series of aggregation levels one to five on the overall sales for the train, validation, and test set separately, thus identifying possible variations in their relative importance. This is critical since, as described in Makridakis et al. (2020e) and Makridakis et al. (2020f), the measures used in the M5 competition for evaluating forecasting performance consider a weighting scheme that puts more emphasis on the series that account for higher monetary sales and for which accurate forecasts are expected to add more value to the retail company. Moreover, if the properties of the test set differ from those of the train and validation set, that could mean that selecting and optimizing forecasting methods based on cross-validation results is a challenging task. However, by observing the results of Table 3 we find that this is not the case. The variations of

Table 2: Summary statistics for the key characteristics of the product-store series of the M5 competition: Length, average unit sales, percentage of days with zero sales, average inter-demand interval ($ADI$), coefficient of variation of the demand when it occurs ($CV^2$), spread of the sales when they occur, average price, coefficient of variation of the price, and spread of the price.

| Characteristic | Min | Mean | Median | St. Deviation | Max |
|---|---|---|---|---|---|
| Length | 124.00 | 1562.81 | 1810.00 | 477.18 | 1969.00 |
| Average unit sales | 0.02 | 1.35 | 0.61 | 2.88 | 130.68 |
| Zero sales (%) | 0.15 | 59.89 | 63.11 | 23.35 | 98.68 |
| ADI | 1.00 | 4.16 | 2.71 | 4.23 | 75.65 |
| CV$^2$ | 0.00 | 0.39 | 0.35 | 0.69 | 107.00 |
| Spread of sales | 0.00 | 12.58 | 7.00 | 22.18 | 762.00 |
| Average price | 0.19 | 4.45 | 3.38 | 3.52 | 30.51 |
| CV$^2$ of price | 0.00 | 0.00 | 0.00 | 0.01 | 0.95 |
| Spread of price | 0.00 | 0.61 | 0.28 | 1.30 | 104.06 |

the proportions of the key series of the competition are minor, both in unit and monetary terms. This is particularly true for the validation set since it refers to the 28 days before the test period and, therefore, it is not influenced by long-term trends and structural changes in the operation of the company and its stores, as well as variations in customer preferences. By comparing the proportions of the units sales to those of the dollar sales, we also find that customers in *CA* tend to buy more expensive products than customers in *TX* and *WI*, and that *Foods* are less expensive than *Hobbies* and *Household* products but still have more value for the company since they represent a greater proportion of total unit sales.

## 5. Conclusion

This paper described the background, organization, and implementations of the M5 competition, demonstrating its design attributes and features. Moreover, it presented the data set used that was analyzed to explore its key characteristics. The aim of this work is to facilitate the understanding of the results of the competition and to serve as introductory material to researchers interested in replicating its findings or exploiting its data set and the forecasting methods introduced by the competition to advance their future research. In addition, it is aimed to provide information to practitioners interested in applying the findings of the competition to improve the accuracy and uncertainty of the forecasting methods they use in their own firms or to benchmark their own results against those of the M5.

The M5 competition, contrary to the previous M ones, focused exclusively on retail sales forecasting, with the objective to empirically evaluate the performance of the methods used to produce point forecasts and estimates of uncertainty. To do so, the competition used a large data set of 42,840 time series, presenting

Table 3: Proportion (percentage share) of key series of the competition on total sales, both unit and dollar ones.

| Aggregation level | Level id | Number of aggregated series | Unit sales (%) *train set* | Dollar sales (%) *train set* | Unit sales (%) *validation set* | Dollar sales (%) *validation set* | Unit sales (%) *test set* | Dollar sales (%) *test set* |
|---|---|---|---|---|---|---|---|---|
| Total | 1 | 30,490 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| CA | 2 | 12,196 | 43.61 | 44.85 | 42.31 | 44.14 | 42.60 | 43.62 |
| TX | 2 | 9,147 | 28.70 | 28.76 | 26.74 | 27.60 | 26.99 | 28.12 |
| WI | 2 | 9,147 | 27.70 | 26.39 | 30.95 | 28.27 | 30.42 | 28.25 |
| CA 1 | 3 | 3,049 | 11.69 | 11.97 | 10.88 | 11.13 | 10.90 | 11.19 |
| CA 2 | 3 | 3,049 | 8.74 | 9.36 | 10.79 | 11.25 | 11.19 | 11.59 |
| CA 3 | 3 | 3,049 | 16.93 | 17.01 | 14.24 | 15.14 | 14.04 | 14.19 |
| CA 4 | 3 | 3,049 | 6.25 | 6.51 | 6.40 | 6.63 | 6.46 | 6.66 |
| TX 1 | 3 | 3,049 | 8.50 | 8.37 | 7.92 | 7.90 | 8.18 | 8.19 |
| TX 2 | 3 | 3,049 | 10.92 | 10.88 | 9.36 | 9.61 | 9.37 | 9.78 |
| TX 3 | 3 | 3,049 | 9.28 | 9.51 | 9.47 | 10.09 | 9.44 | 10.16 |
| WI 1 | 3 | 3,049 | 7.88 | 7.90 | 9.13 | 8.67 | 9.06 | 8.80 |
| WI 2 | 3 | 3,049 | 10.04 | 9.50 | 12.50 | 11.31 | 11.96 | 11.09 |
| WI 3 | 3 | 3,049 | 9.77 | 8.99 | 9.32 | 8.29 | 9.40 | 8.36 |
| Foods | 4 | 14,370 | 68.59 | 57.97 | 67.59 | 57.30 | 67.29 | 56.02 |
| Hobbies | 4 | 5,650 | 9.33 | 12.19 | 9.41 | 12.90 | 9.38 | 13.07 |
| Household | 4 | 10,470 | 22.08 | 29.83 | 23.01 | 29.80 | 23.33 | 30.91 |
| Foods 1 | 5 | 2,160 | 7.76 | 6.90 | 8.31 | 7.26 | 8.03 | 7.23 |
| Foods 2 | 5 | 3,980 | 11.67 | 13.38 | 13.41 | 14.68 | 12.84 | 14.43 |
| Foods 3 | 5 | 8,230 | 49.16 | 37.70 | 45.86 | 35.36 | 46.43 | 34.36 |
| Hobbies 1 | 5 | 4,160 | 8.51 | 11.56 | 8.33 | 12.20 | 8.49 | 12.43 |
| Hobbies 2 | 5 | 1,490 | 0.81 | 0.63 | 1.08 | 0.70 | 0.89 | 0.64 |
| Household 1 | 5 | 5,320 | 17.53 | 22.01 | 18.05 | 22.32 | 18.22 | 22.89 |
| Household 2 | 5 | 5,150 | 4.55 | 7.82 | 4.96 | 7.47 | 5.11 | 8.02 |

the hierarchical unit sales of Walmart, the largest retail company in the world, and including a number of explanatory variables, typically used by retailers to improve the accuracy of their forecasts and estimate uncertainty as precisely as possible to determine appropriate inventory levels for higher levels of customer satisfaction.

Is the M5 data set provided by Walmart representative of the retail sector to enable generalizing its findings and conclusions about hierarchical, unit sales forecasting (for a feature-based investigation on this topic please refer to Theodorou et al., 2021)? What is obvious is that the data set originates from a giant retail firm that has embraced a "constant low prices" strategy for driving its sales, covering ten stores located in three USA states and three different categories of products, as well as a specific period of time. Therefore, we believe that more research would be needed to generalize the findings of the M5 and draw conclusions for

- Smaller retail firms.
- Retailers that rely heavily on promotions instead of constant low prices to attract customers.

- Firms that operate outside the USA, or in regions with significantly different shopping habits and market dynamics than those in Texas, Wisconsin, and California.
- Product categories outside foods, hobbies, and household items.
- Sales of e-commerce retail firms.
- Different time periods than those covered by the M5 data set.

As we have indicated on several occasions in this paper, the findings and conclusions of the M5 competition are significant, guiding the theory and practice of forecasting. At the same time there are limitations to how much they can be generalized beyond the data they represent. After all, forecasting is not "crystal balling" but an objective way of identifying and extrapolating established patterns and relationships, based on accumulated knowledge built over the years, to obtain the best possible results. Slowly but clearly the field of forecasting is advancing over time, providing useful advice on ways to improve performance and offering concrete benefits to those following them.

## References

Athanasopoulos, G., & Hyndman, R. J. (2011). The value of feedback in forecasting competitions. *International Journal of Forecasting*, *27*, 845–849.

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

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

Bojer, C. S., & Meldgaard, J. P. (2020). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, (pp. 1–17).

Boylan, J. E., Goodwin, P., Mohammadipour, M., & Syntetos, A. A. (2015). Reproducibility in forecasting research. *International Journal of Forecasting*, *31*, 79–90.

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.

Fildes, R. (2020). Learning from forecasting competitions. *International Journal of Forecasting*, *36*, 186–188.

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

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

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.

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

Huang, T., Fildes, R., & Soopramanien, D. (2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research*, *237*, 738–748.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*, 7–14.

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.

Ma, S., & Fildes, R. (2017). A retail store SKU promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research*, *260*, 680–692.

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

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., Fry, C., Petropoulos, F., & Spiliotis, E. (2021). The future of forecasting competitions: Design attributes and principles.

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. (2018b). 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). *Predicting/hypothesizing the findings of the M5 competition*. Working paper available at: https://drive.google.com/drive/u/1/folders/1S6IaHDohF4qalWsx9AABsIG1filB5V0q.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020c). Responses to discussions and commentaries. *International Journal of Forecasting*, *36*, 217–223.

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

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

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Bergmeir, C., Bessa, R. J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., Baets, S. D., Dokumentov, A., Fiszeder, P., Franses, P. H., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Arenas, J. R. T., Wang, X., Winkler, R. L., Yusupova, A., & Ziel, F. (2020). *Forecasting: theory and practice*. Working paper available at: https://arxiv.org/abs/2012.03854v2.

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.

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.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.

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

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

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

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 Forecast-*

*ing*, *16*, 437–450.

Theodorou, E., Wang, S., Kang, Y., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Exploring the representativeness of the M5 competition data. Available at: https://arxiv.org/abs/2103.02941.

Önkal, D. (2020). M4 competition: What's next? *International Journal of Forecasting*, *36*, 206–207.