International Journal of Forecasting 36 (2020) 135–141

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Forecasting the M4 competition weekly data: Forecast Pro's winning approach

Sarah Goodrich Darin [*], Eric Stellwagen

Business Forecast Systems, Inc., United States of America

ABSTRACT

Forecast Pro forecasted the weekly series in the M4 competition more accurately than all other entrants. Our approach was to follow the same forecasting process that we recommend to our users. This approach involves determining the Key Performance Metric (KPI), establishing baseline forecasts using our automated expert selection algorithm, reviewing those baseline forecasts and customizing forecasts where needed. This article explores why this approach worked well for weekly data, discusses the applicability of the M4 competition to business forecasting and proposes some potential improvements for future competitions to make them more relevant to business forecasting.

1. Introduction

Forecast Pro is off-the shelf business forecasting software developed by Business Forecast Systems, Inc. (BFS), a company founded by Robert L. Goodrich and Eric A. Stellwagen in 1986. The core forecasting methodology of our software, Expert Selection, has been evolving for the last 32 years and was the driver of our success in the M3 competition (Makridakis & Hibon, 2000), which focused solely on automatic time series methods. The M3 competition, in turn, was an important driver of Forecast Pro's early success, as it established the accuracy of our proprietary algorithms.

Despite the M3 competition's many important contributions, analytics has changed tremendously over the last 20 years, and the M4 competition provided an excellent platform for testing new methodologies and approaches for forecasting "big data". In addition, allowing the competitors to go beyond automated approaches and develop custom approaches that were tailored to the data being forecast made the M4 competition a good test of the way in which practitioners and data scientists forecast data in today's world. However, like the M3 competition, the M4 competition data series were not representative of the types of data businesses typically forecast (i.e., the demand for products and services), and there was no opportunity to integrate domain knowledge. Thus, the findings from the competition may not be fully relevant for business applications. The weekly data are perhaps the most similar to business data, and this is probably one of the factors that contributed to our first place showing in this category.

Our M4 entry was developed using Forecast Pro TRAC. This product is designed to help business forecasters establish and maintain a comprehensive, accurate and well documented demand forecasting process. Our M4 entry used Expert Selection to forecast a strong majority of series, but we also utilized Forecast Pro's process improvement capabilities to identify problematic series, provide custom models where appropriate and perform out-of-sample analyses.

2. Our approach to business forecasting

Forecast Pro is designed to help business people forecast large amounts of data accurately and efficiently. Our approach for the M4 competition was to make use of the key Forecast Pro functionality pertaining to large-scale business forecasting. Our objective was to determine how well a Forecast Pro user (rather than a data scientist who has the option of writing fully customized code) could forecast the M4 data.

BFS's M4 submission applied the forecasting process framework that we recommend to our customers.

[*] Corresponding author.
E-mail address: sdarin@forecastpro.com (S.G. Darin).

https://doi.org/10.1016/j.ijforecast.2019.03.018

Step 1: Determine and understand your forecasting goals and key performance indicators (KPIs).

Step 2: Establish baseline models, typically through an automated approach.

Step 3: Review baseline models using exception reports, outlier reports, and graphs to identify the time series for which the baseline model may be inadequate.

Step 4: Review the time series identified in Step 3 and, based on domain knowledge, either accept the baseline model, substitute an alternative customized statistical model or judgmentally override the forecast, and document all changes.

Our KPI was the overall weighted average (OWA) of two accuracy measures: the relative mean absolute scaled error (MASE) and the relative symmetric mean absolute percentage error (sMAPE), as described in the *M4 Competitor's Guide*.[1] Our baseline model was determined primarily using Expert Selection, although we did consider some alternative baseline models. The forecast review was performed by using Forecast Pro's filtered reports (primarily exception reports) and graphing abilities to efficiently identify series that showed unusual patterns and therefore might require customized approaches. While we consider applying domain knowledge to be a critical part of the forecasting process, our ability to apply domain knowledge in this case was extremely limited, given that there was no indication of what the series represented.

2.1. Forecast Pro's expert selection methodology

The basic premise of Forecast Pro's Expert Selection methodology is simple: fit the appropriate forecasting model to the data at hand. To accomplish this, Forecast Pro has three logical layers.

The top layer consists of a master control program for selecting the family of models to be considered, e.g. exponential smoothing, Box-Jenkins, intermittent demand models, etc.

The second layer identifies a particular model from the family, e.g. ARIMA(1,1,0)*(0,1,1), multiplicative Winters, etc. Of course, the identification protocol is specialized to the method.

The third layer optimizes the parameters via unconditional least squares and prepares the actual point forecasts. The methodologies considered in step 1 include:

1. 12 variations of Holt/Winters exponential smoothing (Gardner, 2006).
2. Three variations of Winters exponential smoothing utilizing seasonal simplification.[2]
3. NA-CL exponential smoothing, a non-trended constant level model with additive seasonality that is useful for data with strong selling seasons (Stellwagen & Goodrich, 2017).
4. ARIMA (Box & Jenkins, 1976).
5. Croston's intermittent demand model.[3]
6. Simple moving averages.
7. Discrete data models.[4]

The Expert Selection protocol starts with an analysis of the data structure and properties of the individual time series to be forecasted. The data properties are then analysed using a rule-based expert system incorporating upward of 200 rules. The output from this *inference engine* is either direct selection of the methodology to use or a winnowed-down list of potential methodologies to consider. For example, ARIMA may be ruled out of contention if the data are of inadequate length, or Croston's may be selected based on the frequency and distribution of zeros in the data. In cases where the inference engine results in a winnowed-down list, the algorithm uses a rolling out-of-sample testing procedure to select the appropriate model family. The expert system used in our algorithm is an application of artificial intelligence and was prototyped originally in Prolog and then re-implemented in C++ for improved performance.

The accuracy of the Forecast Pro software also depends on the implementations of exponential smoothing and ARIMA identification. The BFS implementation of exponential smoothing uses the simplex algorithm to minimize the historic sum of squares. The Holt/Winters model is identified via the BIC, supplemented with some additional logical rules. If seasonally simplified models and/or the NA-CL model are to be considered, an out-of-sample test is used to determine whether or not they are superior to the selected Holt/Winters model. In addition, Forecast Pro monitors all multiplicative models for signs of instability. Parameter initialization is done via backcasting. As Dr. Goodrich wrote after the M3 competition, "We believe that this method is superior to the out-of-sample identification method used in certain other commercial packages and the results of the competition seem to support this contention." (Goodrich, 2000)

BFS uses a proprietary approach for ARIMA identification. The general approach is to start by overfitting a *state space* model; then, that model is used to obtain parameter estimates for a large number of ARIMA models. The BIC and other proprietary logical rules are used to select the specific ARIMA model, after which the ARIMA parameter estimates are refined via unconditional least squares as described by Box and Jenkins (1976).

Expert Selection provides very good forecasts for most data series, but there are times when the forecast can

[1] Retrieved from https://www.m4.unic.ac.cy/wp-content/uploads/2018/03/M4-Competitors-Guide.pdf.

[2] Seasonal simplification reduces the number of seasonal indices used in a Winters model by grouping multiple consecutive weeks into a single seasonal index (see https://theforecastpro.com/2014/11/10/seasonal-simplification/#more-833).

[3] Forecast Pro implements both the Croston model as originally implemented (Croston, 1972) and the Willemain variant (Willemain, Smart, Shockor, & DeSautels, 1994). The logic for determining model selection is described in the *Forecast Pro Statistical Reference Manual* (Stellwagen & Goodrich, 2017).

[4] Discrete data models consider the Poisson distribution and the negative binomial distribution (Stellwagen & Goodrich, 2017).

be improved by applying domain knowledge. Even without domain knowledge, there are times when an analyst can identify Expert Selection forecasts that can clearly be improved. The M4 data contain many incidences of unusual cyclical patterns and large structural shifts in the data. Because Forecast Pro is designed for business forecasting, our automated methods are not calibrated to very unusual data. If a business series exhibits unusual structural changes, the reason is virtually always known (e.g., a patent expired, the company merged with another firm, etc.), and this domain knowledge will guide how the series should be forecast. Forecast Pro provides a wide range of models that can be applied in such situations.

2.2. Forecast Pro's procedure for forecasting weekly data in the M4 competition

BFS prepared our M4 entry using the four-step procedure described earlier.

The first step was to select our KPI. We used an out-of-sample overall weighted average, as described in the *M4 Competitor's Guide*. The overall weighted average is defined relative to the Naive2[5] model, so methods with an OWA above 1 perform worse than the Naive2 model, while methods with an OWA below 1 perform better.

The second step was to establish baseline models. Given that the M4 forecast horizon for weekly data was 13 weeks, we defined a 13-week holdout sample and then considered the following approaches:

1. Expert Selection.
2. Expert Selection with a log transform.
3. Expert Selection with outlier detection.
4. Forecast Pro's Custom Component Model with the level set to the last observed data point, no trend, and multiplicative seasonality. This is equivalent to a Naive2 model.
5. Forecast Pro's Same as Last Year model. This is equivalent to a seasonal naïve model.
6. A dynamic regression model with automated dynamics (Forecast Pro optimizes the Cochrane-Orcutt terms and lagged dependent variables to use) and a shift variable for the last year of the data series.
7. A dynamic regression model with automated dynamics, a shift variable for the last year of the data series and monthly indicator variables.

We considered approaches other than Expert Selection because our goal in entering the M4 competition was to forecast the M4 dataset accurately, not to test our expert selection algorithm. Expert Selection was developed and optimized for forecasting business data, and the vast majority of the M4 data clearly were not business data. Based on our review of the data, it was obvious that a consideration of additional methodologies was likely to improve the accuracy of our M4 entry. The additional methods were selected primarily to accommodate some of the unusual outliers, level shifts and cyclical structures that were apparent across the M4 dataset.

We applied each of the above approaches and calculated the overall weighted average (OWA) for the holdout period. We selected the model that had the lowest OWA for the static holdout period, provided that it was at least 0.03 lower than Expert Selection. If the OWA for Expert Selection was no more than 0.03 higher than the minimum, we chose Expert Selection. Given that the Expert Selection algorithm is more robust and proven than a simple out-of-sample OWA comparison, it seemed appropriate that a non-Expert Selection alternative should be used only if it outperforms Expert Selection by a non-trivial margin. We selected 0.03 as our threshold after testing various alternatives.

The third step was to identify the time series for which the baseline model might be inadequate. Using Forecast Pro's exception reports and filtering tools, we sorted the series in ascending order based on OWA. Series with OWA values above 1 (meaning that the Naive2 model performed better) were flagged as potentially inadequate models.

The fourth step was to review the models flagged in step 3 and consider customizing the model for these time series. Due to lack of domain knowledge, this was an ad hoc process and we did not consider judgmental overrides to the forecast. In some cases, we specified dynamic regression models for individual series in order to capture simplified seasonal patterns or data structures that the automatic dynamics capability does not support. In some cases, there were short data series that appeared to be seasonal but did not have enough data history to support a seasonal model. We customized these series using same-as-last-year-with-percentage-change models. Finally, we corrected outliers on a case-by-case basis and eliminated initial data points for any series that had a large visual shift in data structure at the beginning of the series. We then applied the selected models to the full historical data series and generated forecasts.

Due to the structure of the M4 competition, we added an additional step to our standard approach, a final "sanity check". We used Forecast Pro's exception reports to compare the cumulative forecasts to the equivalent period in the prior year. Sorting by the size of the absolute value of the change, we reviewed series with large changes and adjusted them as needed (for example, if there was a large percentage increase and a log transform was applied, we removed the transform to dampen the increase a bit). If the forecast appeared unstable, we often changed the model to a naïve model or a simple moving average. We also adjusted models that had periods with zero forecasts, typically by applying a log transform. In short, we went through a final review process to make sure that our submission made sense. Ultimately, we rejected the original baseline forecasts for 10% of the weekly series using the review process described. Given the efficiency of the exception report approach for identifying problematic series, this process took only a few hours. The frequency of our final model selection was:

[5] As is described in the *M4 Competitor's Guide*, the Naïve2 model applies a Naïve or random walk model to seasonally adjusted data.

Model applied	Frequency
Expert Selection	69.0%
Expert Selection with log transform	14.5%
Dynamic regression with shift	5.8%
Simple moving average	2.8%
Same as last year	2.5%
Dynamic regression with shift and months	2.2%
Customized dynamic regression	1.9%
All other	1.1%

2.3. Why our approach worked for weekly data

The Forecast Pro software is designed for business forecasting. While the bulk of the M4 data series clearly were not business data, our approach worked well for weekly data primarily because the weekly data were more similar to business data than those of other frequencies. Business forecasters rarely use quarterly or annual data (and we consider that extrapolation methods are not the best approach for annual data), so Forecast Pro is not optimized to these frequencies. Most of our clients use monthly or weekly data, but an increasing number are using daily and hourly data as well. Our approach also worked relatively well for monthly and hourly data, where our submissions came in at tenth and ninth places respectively.

It is worth noting that, unlike the weekly data, both the monthly and daily data exhibited very strong cross-series correlations, which the top performers in these categories leveraged. While it is to be expected that seasonality and industry trends will drive contemporaneous correlation, non-contemporaneous relationships are less common and are expected only when there is a true dependency of one product on another. As ProLogistica Soft observed, "in the M4 data set some series are very highly correlated to fragments of other series".[6] As a result, there is an advantage in informing the forecast for a given series using data from a different series. The daily data structure is particularly unusual, with only two competitors beating the naïve forecast by more than 0.03; however, the correlations across series are so strong that making use of the cross-series correlations allowed ProLogistica Soft to outperform the naïve model significantly. While we observed the strong correlations across series, we were committed to using Forecast Pro in the manner for which it was designed. Given that we would not recommend using one data series essentially as a leading indicator for another without having some sort of domain knowledge to confirm that the relationship makes sense, we did not make use of any cross-series correlations. Like most of the business data for which we design our product, the weekly data did not exhibit any unexpected cross-series correlations, thus allowing Forecast Pro to deliver the most accurate weekly forecasts of all of the entrants.

Another possible reason for our strong performance on weekly data is that, like many business series, the weekly series are relatively short. More than half of the weekly series have less than two years of data. Forecast Pro considers series length in the Expert Selection process and is programmed to select the most accurate and robust method for short series. In contrast, other data frequencies in the M4 competition do not include short series (relative to the cycle length).

It is also worth noting a few other ways in which the M4 data are not representative of business data. For example, most of the daily series are non-seasonal random walks. For the monthly data, many series have shocks or level shifts in the last year of the data, and some of these structural changes occur right near the end of the series (see Appendix). Of course, shocks are common in business data, but the analyst almost always knows what is driving the shocks and whether a data point is an outlier, an event or a permanent level shift. Forecasting an edge condition without knowing what happened is like flipping a coin, and good forecasting processes include determining the drivers of any observed anomalies. Finally, none of the series show an intermittent demand pattern. Dealing with low-volume data is very common in business, and therefore forecasting low-volume data accurately is more important to business forecasting than forecasting random walks. While the weekly data have some outliers, they show fewer shocks than other frequencies, and stronger cyclical patterns than the daily data.

3. Business applicability of the M4 competition

The M4 competition is a good measure of what methods work for single series forecasting in the absence of domain knowledge. However, business forecasting is not that simple. Choosing the right method for a single series is clearly important for good business forecasting, but understanding how to organize data into a hierarchy, determining the right level of aggregation for forecasting, developing custom approaches and integrating domain knowledge into a forecast (through dynamic regression, event models or judgmental adjustments) are all critical for the forecast accuracy. Some of these actions can be automated via software, and future software solutions are likely to feature greater automation. For example, automated hierarchy optimization could help forecasters to determine the best aggregation level for forecasting and hierarchy reconciliation. Similarly, business rules and knowledge could be integrated into automated model selection algorithms in order to allow automated models to be customized for a specific business or industry. There has been significant and useful research on these topics, and we hope to see continued efforts in these areas. It is our opinion that focusing on these areas is critical for delivering real improvement in business forecasting for practitioners.

While we understand that the goal of the M competitions is to advance forecasting as a whole, not business forecasting specifically, we would like to see future M competitions focus more on business applicability. Business forecasting is different enough from general forecasting that applying the M4 findings broadly to business

[6] https://github.com/M4Competition/M4-methods/tree/master/237%20-%20prologistica.

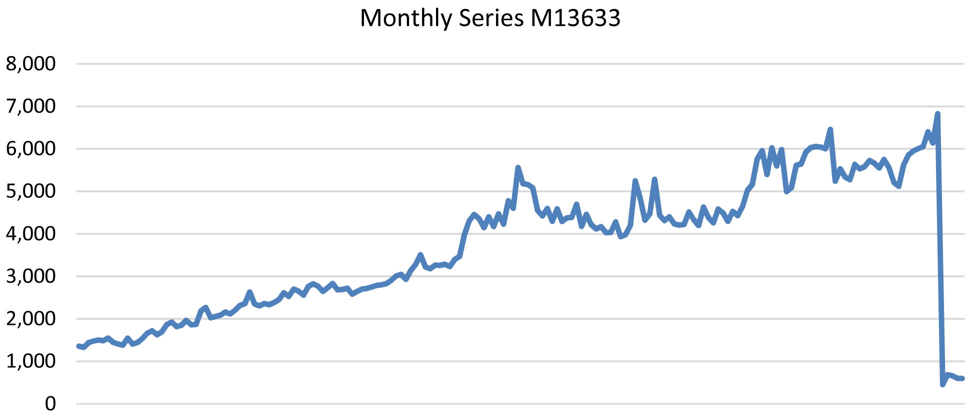

Fig. A.1. The data decline by 90% towards the end of the series.

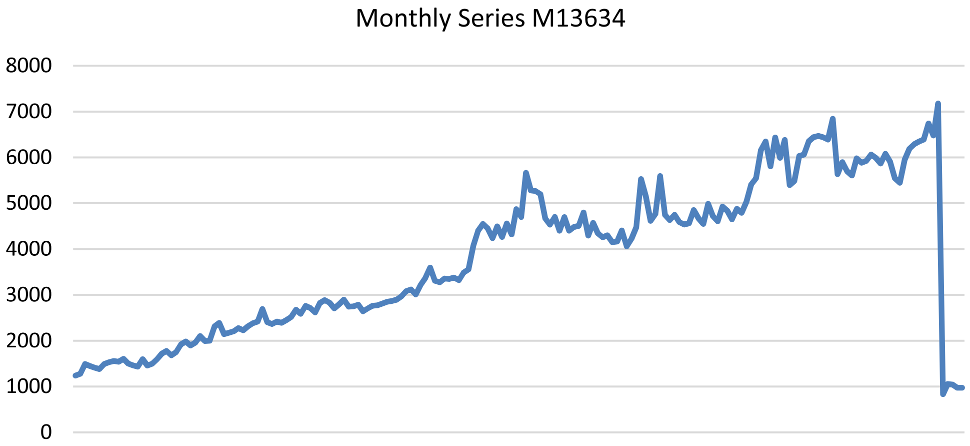

Fig. A.2. The data decline towards the end of the series, showing similar pattern to M13633.

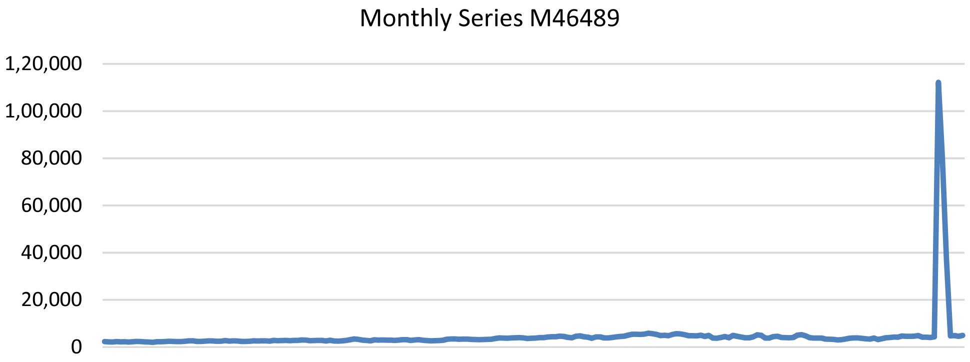

Fig. A.3. The data spike by 2500% towards the end of the series.

forecasting is not advisable. Most importantly, we would like to see forecasting research leverage data that are representative of business data. Businesses typically forecast at the product level for sales and revenue planning, at the SKU level for demand planning and at the SKU-by-location level for supply chain purposes. We saw little evidence that these types of business series were included in the study. Although 19% of the M4 data series were labelled as “Industry” data, the high volume and substantial length (e.g., all of the monthly industry series had more than five years of data and 57% had 15 or more years) of these series suggests that they are not business series collected at a product or SKU level (an educated guess is that they were collected at the industry level). We would also like to see research data being weighted more heavily towards monthly, weekly, daily and hourly

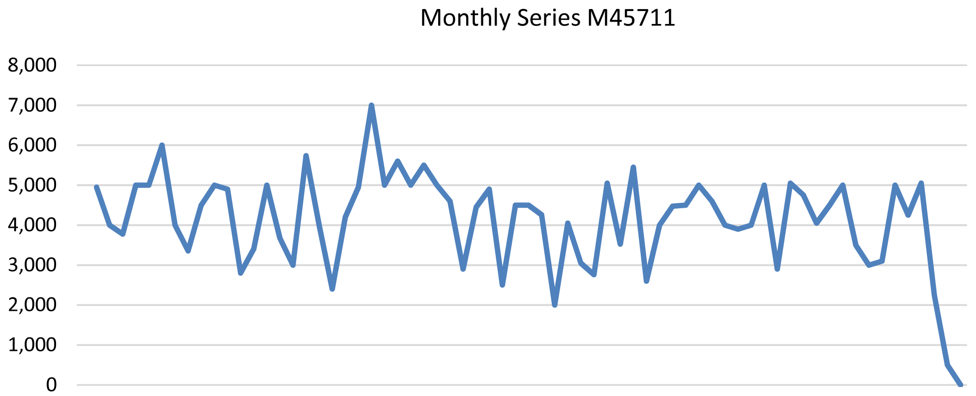

Fig. A.4. The data drop by 90% in the final two months.

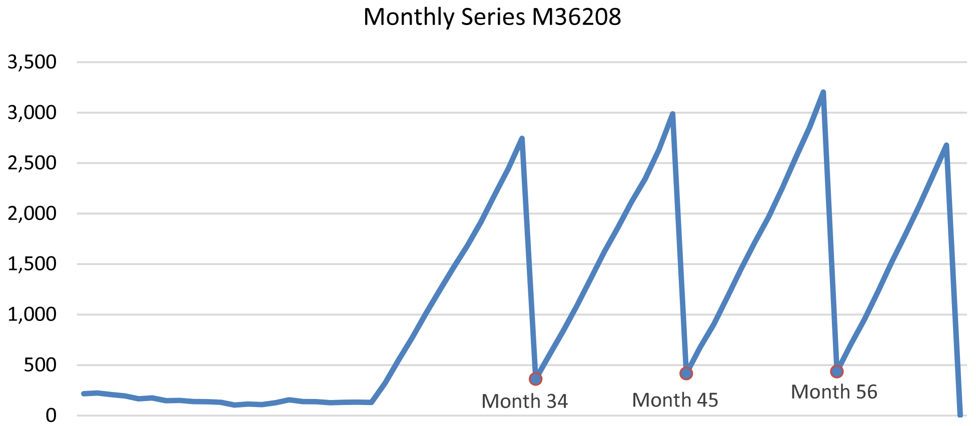

Fig. A.5. The data show an 11-month periodicity; clearly not business data.

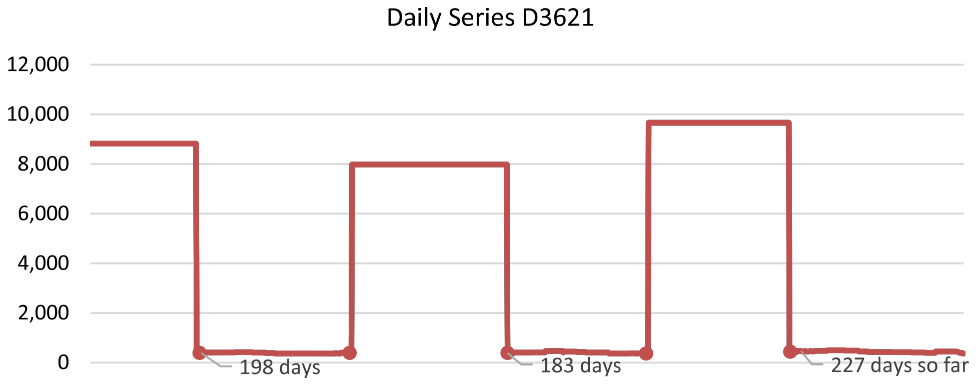

Fig. A.6. A daily series that clearly is not business data; the current "low volume" period has lasted 29 days longer than the next-longest "low volume" period.

business data. The M4 dataset consists of 23,000 annual data series, 24,000 quarterly data series, 48,000 monthly series, and 5000 daily, hourly and weekly series. Thus, annual and quarterly data comprise almost half of the dataset, even though typically they are not forecasted directly by business forecasters. Furthermore, extrapolating annual data forward six years without domain knowledge simply is not realistic. Monthly data are still used commonly by business forecasters (although many business forecasters are migrating to weekly data), but again, most of the monthly data series used in the M4 dataset clearly are not business data. Ideally, a business dataset would include low-volume SKU-level data with a hierarchical structure that could be leveraged in the forecasting process via top-down or middle-out forecasting. It is our hope that we will see more frequent M competitions, and that future M competitions will provide specific learnings that can and should be implemented in business forecasting software.

Appendix. Examples of series from the M4 competition

See Figs. A.1–A.6.

References

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (revised ed.). San Francisco: Holden-Day.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, *23*(3), 289–303.

Gardner (2006). Exponential smoothing: the state of the art—Part II. *International Journal of Forecasting*, *22*, 637–666.

Goodrich, R. (2000). The Forecast Pro methodology. *International Journal of Forecasting*, *16*, 451–476.

Makridakis, S., & Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.

Stellwagen, E., & Goodrich, R. (2017). *Forecast Pro statistical reference manual*. Business Forecast Systems, Inc.

Willemain, T. R., Smart, C. N., Shockor, J. H., & DeSautels, P. A. (1994). Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. *International Journal of Forecasting*, *10*(4), 529–538.