ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

The M4 forecasting competition – A practitioner's view

Chris Fry *, Michael Brundage

Google, Inc., 1600 Amphitheater Parkway, Mountain View, CA 94043, the United States of America

1. Introduction

The M4 Forecasting Competition (Makridakis, Spiliotis, & Assimakopoulos, 2018) was a huge success along many dimensions. It enabled comparisons of 60 different forecasting methods across 100,000 real-world time series. It also created the opportunity for participants and observers to learn from each other through the open sharing of both the full set of test data and the solutions applied.

As practitioners who have worked on a wide range of forecasting problems, we were excited about the competition and its results, as well as about the innovative ideas that were shared as a result of the competition. At the same time, we also saw some disparities between the nature of the competition and the types of problems that we work on and have seen others working on. We share our suggestions for five themes which we observe in the forecasting community and would like to see reflected in future competitions to tie them more closely to today's real-world forecasting and prediction challenges. We also outline some attributes of forecasting approaches that we expect will be key success factors in building prediction models for such problems, and discuss how they resonate with the competition findings. Lastly, we compare the attributes of the M4 competition data set with those of a recent Kaggle competition on web traffic time series forecasting, which was hosted during the same time period.

2. Future directions: five themes to consider when designing the M5 and future forecasting competitions

Having been members of the forecasting community for many years, we have observed various changes over time in the types of forecasting challenges that we have been asked to solve and have seen others solve. These changes have been driven in part by a changing world with faster decision-making, a greater availability of information, and a sheer increase in the number of time series being forecasted, as well as by the availability of better tools with which to predict (dramatic improvements in computing power, artificial intelligence, and machine learning). With such changes in mind, we are highlighting five themes that we believe are present in the forecasting space today and which we would like to see reflected in the design of future forecasting competitions.

Theme 1: Time intervals are changing. The M4 competition included 100,000 time series spanning a mix of data frequencies: 23% yearly, 24% quarterly, 48% monthly, 0.4% weekly, 4.2% daily, and 0.4% hourly. We are curious about how this mix was selected and why the mix was weighted heavily toward longer-duration (especially monthly) time intervals. At Google, like many other firms, we more often deal with time series of daily, hourly, and even shorter (e.g., 5-min) intervals, as well as time series with irregular time stamps (e.g. raw event timestamps). We expect that as business cycles speed up and automation enables faster and more agile planning, forecasters will contend increasingly with shorter-interval time series. Often, these shorter-interval time series are noisier, with a higher spectral entropy, a weaker trend and/or seasonality and weaker autocorrelation, and/or are intermittent in nature. Seasonality effects can also become more complex and harder to model in shorter-interval time series, due to the compounding effects of multiple seasonality (e.g. hour-of-day, day-of-week, week-of-year, etc.), evolving seasonality (e.g., weekend dips becoming shallower over time), moving holidays, and other challenges. Lastly, we observe that shorter-interval time series can often be of short duration, compounding the complexity associated with the generation of data-driven forecasts.

As an example, consider the 145,000 time series of Wikipedia traffic volume that were the subject of the

* Corresponding author.
E-mail addresses: chrisfry@google.com (C. Fry), brundagem@google.com (M. Brundage).

https://doi.org/10.1016/j.ijforecast.2019.02.013

Please cite this article as: C. Fry and M. Brundage, The M4 forecasting competition – A practitioner's view. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.02.013.

2017 Kaggle competition on web traffic time series forecasting which was hosted by Google.[1] These were entirely daily series (also not a representative mix), but reflect quite a different profile and probably a very different set of challenges to those presented by M4. We suggest that future M competitions could be strengthened by including a more diverse set of time series. We also note that sparse, shorter-interval time series provide a greater opportunity to use temporal aggregation techniques to improve forecast accuracy, which could potentially alter the landscape of winning solutions.

Theme 2: Time series are often hierarchical. Many forecasting applications require multiple time series to be forecast simultaneously. These are often hierarchical in nature and often represent sets of time series which can be highly correlated. Consider for example a retailer tasked with forecasting the sales of thousands or even millions of product SKUs across hundreds or thousands of store locations, with each forecasted item being a member of both a geographic and a product hierarchy. At Google, we forecast growth in our product areas both globally and across geographies, with geographic and product hierarchies being embedded in the structure of the problem. Such structures present another set of challenges and feature opportunities which we would like to see included in future forecasting competitions to improve the realism. We note that the number of downloads of the R package "hts" (Hyndman, Lee, Wang, & Wickramasuriya, 2018) has reached over 7000 per month as of January 2019,[2] further exemplifying the prevalence of such problems and the strong interest in tools for working with hierarchical series.

Theme 3: Forecasts require more information than just the time series. In our experience, time series forecasting rarely works with time series data in isolation. We almost always have a much richer set of information about the time series, including: categorical attributes, exogenous time series (such as weather or other external events), hierarchical relationships among the time series (as noted above), and logical constraints (e.g., some forecasts should always be non-negative). This information can be used successfully to increase the forecast accuracy. In some cases, these other features are more valuable to the model than the historical time series values. For example, consider Google Trends data, which shows the scaled search volumes for trending query terms. Obviously the search terms and the features extracted from them are important covariates for any model that wishes to attempt to forecast these time series. The exclusion of feature information from past M competitions may help to explain the discrepancy between Kaggle forecasting contests, where ML-based models have performed very well, and the M competitions, where pure ML models have been less successful. We recommend that some collections of time series with relationships among them, as well as descriptive attributes of some time series, be added to future M competitions.

Theme 4: Prediction intervals are important. We have worked on numerous supply chain planning and capacity planning problems at Google and elsewhere. We set capacity in Google's data centers to support spikes in compute workloads that may occur at certain times of day, on certain days of the week, or in conjunction with specific events, as well as to support the growth of those workloads over time. For these purposes, we work extensively with range forecast prediction intervals, building the capacity to support up to the 95th or higher percentiles of our forecasts. We place great emphasis on getting the tails right. In fact, for some capacity planning problems we find that it is often more important to get the upper quantiles correct than to actually get the point forecasts correct. We believe that the same holds true in other applications as well. We were glad to see prediction intervals introduced into the M4 competition this year, and were impressed with the performances of the top competitors. We hope that this trend continues.

Theme 5: One size doesn't fit all. In industry, our time series are almost never neat and tidy. Some time series are sparse, some are all-constant, some are erratic with no discernible pattern, and some have dynamic ranges over many orders of magnitude. The Wikipedia dataset mentioned in theme 1 includes examples of all of these types of challenges. In addition, we are often tasked with forecasting cold start time series with little or no direct history, in which case it is critical to be able to rely on related data. Our forecasting solutions must work well in these environments across a wide diversity of data, and this naturally pushes end-to-end solutions towards a hybrid approach that combines the algorithms that perform best on each "kind" of time series. We recommend that future M competitions be extended to include more varied and messier datasets, including some that might ordinarily be considered unforecastable. We recommend that some of the different types of time series datasets be identified and algorithm performances be reported on each type separately, in addition to on the full dataset.

3. What it takes to succeed in the new environment

As the nature of forecasting problems and challenges evolves, we offer our suggestions for strategies that we expect to be differentiators for building successful forecasting models in the new landscape.

1. Global models > local models. We were delighted to see our friends from Amazon highlight the distinction between local and global forecasting models in their 2018 *Foresight* article "Deep learning for forecasting: current trends and challenges" (Januschowski, Gasthaus, Wang, Rangapuram, & Callot, 2018). We concur with this distinction and believe that it highlights an important consideration when designing a forecasting approach. For forecasting problems that involve large numbers of related time series, hierarchical time series, and series with categorical attributes or exogenous features, we expect that models trained *across* time series will have greater

[1] www.kaggle.com/c/web-traffic-time-series-forecasting, July 13 – September 10, 2017.

[2] The hts package recorded 7358 downloads per month as of 2019-01-30; https://cranlogs.r-pkg.org/badges/hts.

Please cite this article as: C. Fry and M. Brundage, The M4 forecasting competition – A practitioner's view. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.02.013.

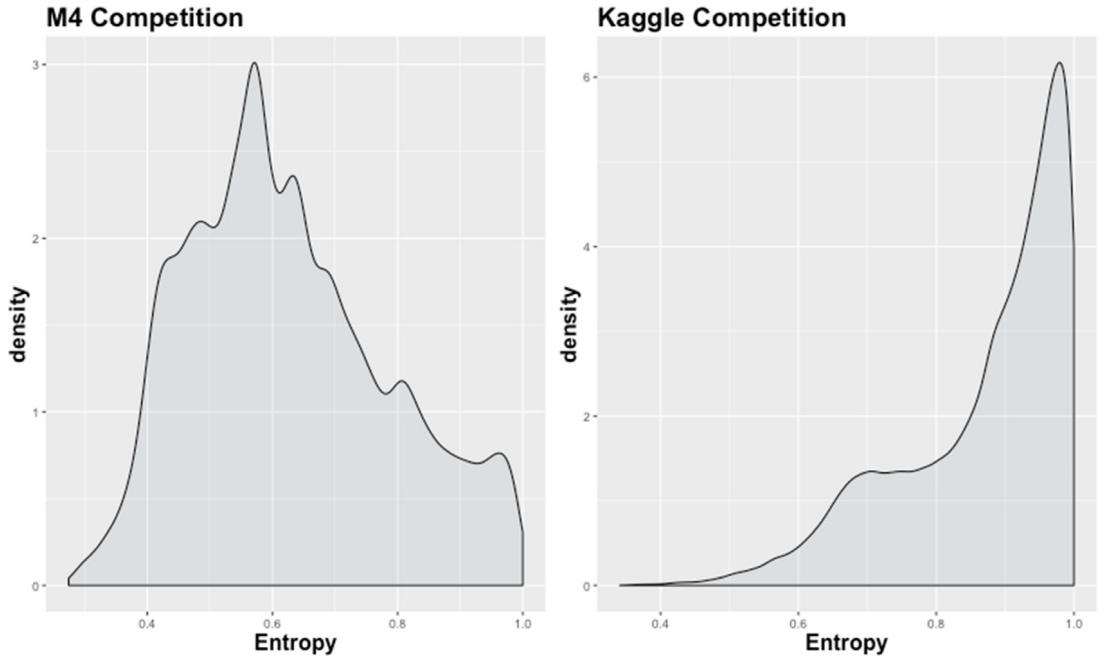

Fig. 1. Comparison of spectral entropy densities for the 100,000 M4 competition time series and the 145,000 series from the 2017 Kaggle competition on web traffic time series forecasting that was hosted by Google.

predictive power. At Google we have had considerable success in moving from local to global models for forecasting problems across multiple domains.

2. Machine learning is a game-changer. At Google, machine learning is becoming a tool of choice, both for us internally in the forecasting domain and for our Cloud customers. Many forecasting applications which previously were purely statistical in nature are being migrated to machine-learning models so that numerous explanatory features can be included in the prediction structure more easily. We have seen significant improvements in forecast accuracy from the use of machine-learning models, and we expect this trend to continue. We are excited that forecasting has become a field to which statisticians, econometricians, operations research professionals, and machine learning experts can each contribute and in which they can collaborate by introducing their own techniques and perspectives.

3. Smart combining > ensembling. Finding 1 in the main M4 competition paper (Makridakis et al., 2018) highlighted "the improved accuracy of combining". We agree with this insight but would further differentiate between raw ensembling (e.g. arithmetic averaging of independent time series forecasts) and "smart combinations", by which we mean hybrid models that bring together multiple modeling approaches in a directed way that is tailored to the structure of the problem at hand. We view both the first- and second-place solutions as examples of such "smart combinations". Slawek Smyl's first-place ES-RNN (exponential smoothing – recurrent neural network) solution (Smyl, Ranganathan, & Pasqua, 2019) was a custom-designed hybrid model that embedded time series solutions directly into a neural network that exploited the seasonal structure of the data. The second-place solution of Montero-Manso, Athanasopoulos, Hyndman, and Talagala (2018) involved the use of machine learning to optimize ensemble weights to best fit the specific characteristics of each time series' residual structure. We expect that the most accurate models in many future applications will require a careful consideration of ways of combining the available tools intelligently in order to build custom solutions that match the unique characteristics of the time series being forecast.

4. Invest in prediction interval improvements. As was mentioned in Theme 4 above, we believe that prediction intervals are a critical element of forecasts in many applications. Building models that do a good job of generating realistic prediction intervals is crucial, and we expect the importance of this to increase over time. Given that many off-the-shelf forecasting tools routinely underestimate prediction intervals (Hyndman, 2014; Hyndman, Koehler, Snyder, & Grose, 2002), we expect that research effort that is spent in improving the quality of prediction intervals (e.g., as measured by the difference between expected and observed coverage) will be another differentiator going forward. At Google, we often use statistical backtesting to validate the quality of prediction intervals, and even to create prediction intervals in the first place. Backtesting has the benefit of comparing forecasts to reality, and of not being reliant on assumptions of independence or heteroscedasticity that can confound auto-generated prediction intervals in practice. We are also investing in the building of models of uncertainty that better reflect fat-tailed errors.

Please cite this article as: C. Fry and M. Brundage, The M4 forecasting competition – A practitioner's view. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.02.013.

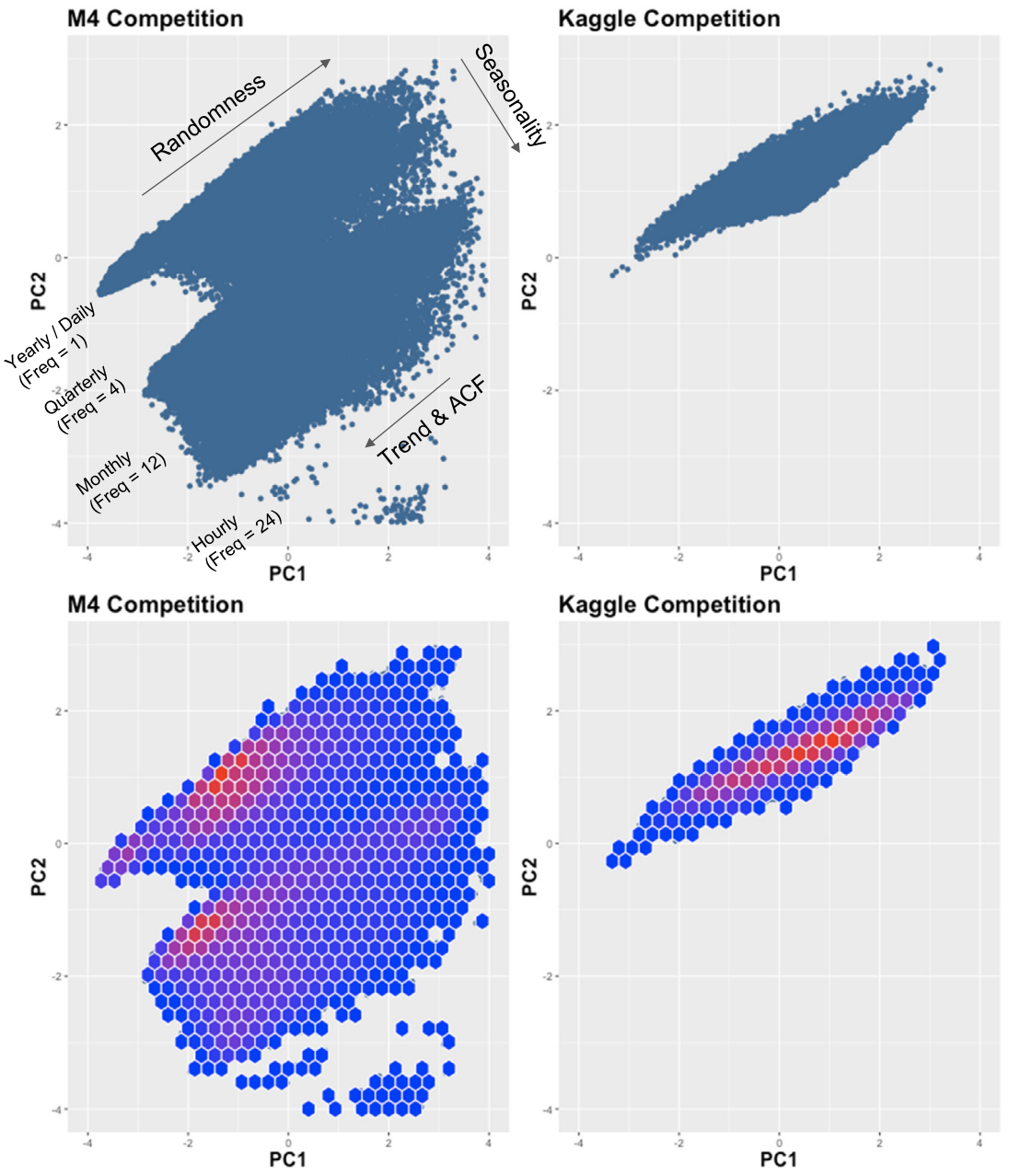

Fig. 2. Comparison of the data for the 100,000 M4 competition time series with the data from the 2017 Kaggle competition on web traffic time series forecasting hosted by Google using the 2-dimensional instance space described by Kang, Hyndman, and Smith-Miles (2017). The top plots are unweighted scatter plots that are similar to those presented in the original paper. The bottom plots show the density of the data in each hexbin region.

4. Case study – comparing the M4 data to the kaggle web traffic competition data

Fig. 1 plots the spectral entropy of the M4 data against that of the 145,000 Wikipedia time series that were the subject of the 2017 Kaggle competition on web traffic time series forecasting which was hosted by Google. The stark difference shows a much higher degree of spectral entropy in the Kaggle competition data, making accurate forecasting arguably "harder".

In addition, Fig. 2 compares the M4 data signature, as profiled using the 2D visualization of time series in the feature space of Kang et al. (2017), against the same data signature for the Wikipedia time series. Kang et al.'s approach plots the top two principal components of a feature space, reflecting the frequency, seasonality, trend, randomness, ACF1 and the Box–Cox λ for each time series.

Shown side by side, we see that the Kaggle data profile is again quite different from that of the M4 data, showing a higher degree of randomness, a lower seasonality, and

Please cite this article as: C. Fry and M. Brundage, The M4 forecasting competition – A practitioner's view. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.02.013.

a lower periodicity. We have also plotted the density for each data set, which shows that the M4 time series were concentrated more in the lower left of the region, whereas the Kaggle data were concentrated more toward the upper right (less seasonality, higher randomness). We suggest that these differences may have been key drivers in the different outcomes of the two contests, with time series based methods faring better in the M4 competition while machine learning methods fared better in the Kaggle competition, and the winning model in the Kaggle competition being a tensor-flow-based recurrent neural network (RNN).

5. Concluding remarks

We believe that the M4 forecasting competition was revolutionary in its attempt to create a forum for competition and sharing based on a wide range of real-world time series examples. The competition has led to significant discoveries, as well as to advances in the field. We hope that our commentary can help to build on these successes to create future competitions that continue to achieve the same goals, while also evolving with the changing forecasting landscape.

References

Hyndman, R. J. (2014). Prediction intervals too narrow. Hyndsight blog, 22 October 2018. https://robjhyndman.com/hyndsight/narrow-pi/. Accessed October 2018.

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*(3), 439–454.

Hyndman, R. J., Lee, A., Wang, E., & Wickramasuriya, S. (2018). hts: Hierarchical and grouped time series. https://cran.r-project.org/web/packages/hts/index.html.

Januschowski, T., Gasthaus, J., Wang, Y., Rangapuram, S. S., & Callot, L. (2018). Deep learning for forecasting. *Foresight*, *50*, 35–41.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2018). FFORMA: feature-based forecast model averaging. Working paper, Monash Business School Department of Economics and Business Statistics.

Smyl, S., Ranganathan, J., & Pasqua, A. (2019). M4 forecasting competition: introducing a new hybrid ES-RNN model. Uber Engineering blog, June 25, 2018. eng.uber.com/m4-forecasting-competition/. Accessed October 2018.

Chris Fry leads the Resource Efficiency Data Science team within Google's Technical Infrastructure division. His team provides data science support for compute and storage resource efficiency initiatives, resource load forecasting and capacity planning, as well as tools and metrics to support the efficiency initiatives. Prior to joining Google he was Managing Director of Strategic Management Solutions, an analytics and data science consulting firm specializing in forecasting, pricing, and supply chain optimization.

Michael Brundage is a Data Scientist at Google, working on Google's internal supply chain. Prior to Google, he worked on data science at Microsoft and search at Amazon, where he helped create the Data Scientist job ladder. Before that, Michael has been a software engineer at Amazon, Yahoo, Microsoft, and Caltech/JPL.

Please cite this article as: C. Fry and M. Brundage, The M4 forecasting competition – A practitioner's view. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.02.013.