International Journal of Forecasting 36 (2020) 195–196

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Data adjustments, overfitting and representativeness

Keith Ord

McDonough School of Business, Georgetown University, the United States of America

I would like to offer my thanks to the organizing team and to all who contributed their analyses to the project. The forecasting literature has been enhanced by this study. The wealth of results and the introduction of new approaches together mean that a lot of further research is needed in order to interpret and understand the implications of the project. Further, I would like to salute the team for undertaking the unglamorous but vital job of validating all of the results. I will restrict my comments to three aspects of the study: data adjustments, overfitting, and the representative nature of the data.

1. Data adjustments

It is apparent from Table 2 of the main paper by Makridakis, Spiliotis, and Assimakopoulos (2020, hereafter MSA2019) that the series were seasonally adjusted for most of the methods when needed, with the exception of ETS and ARIMA. It would be useful if this adjustment process was described, so that interested researchers could carry out such analyses on other data series. Also, applying ETS and ARIMA to the seasonally adjusted series as well as the unadjusted versions would provide some insights into the effectiveness of the adjustment procedure. Some previous studies have suggested that deseasonalizing series before fitting models may produce results that are superior, or at least as good. Furthermore, the number of methods that need to be considered is reduced considerably. Combinations could be created effectively using SES, Holt and damped ES.

Of course, the grandparent of all seasonal adjustment procedures is the program developed over many years by the U.S. Census Bureau, known in its current version as X-13ARIMA-SEATS; for details, see https://www.census.gov/srd/www/x13as/.

It is not reasonable to suggest that this program be run for the present study, not least because the inputs required include series-identifying elements such as the number of trading days (in a month) or the timing of certain holidays. However, such factors can be important for individual series, such as retail sales; failing to allow for such elements means that the inferences that can be drawn from the MSA2019 study may be limited with respect to some classes of time series. The savvy forecaster will want to consider other special features such as sales promotions or extreme weather conditions.

2. Overfitting

Figure 4 in Makridakis, Spiliotis, and Assimakopoulos (2018) is revealing because it shows an increase in sMAPE with the computational complexity. This suggests the presence of over-fitting. That is, random noise and occasional outliers are being treated erroneously as part of the structural model. An illustration of such problems is provided by Ord, Fildes, and Kourentzes (2017, chapter 9): a regression model for the retail price of gasoline used the lagged price, crude oil prices, and several macroeconomic variables, such as personal disposable income, consumer prices, retail sales, and unemployment. The model looked good until the recession of 2008 came along, when it blew apart and a simpler model based only on the lagged price and crude oil prices became vastly superior. Weak relationships may fall victim to structural change, a lesson that needs to be kept in mind for the M5 Competition.

What about outliers? The presence of outliers (and level shifts) makes it more likely that a model selection procedure will select a non-stationary model. Likewise, a stationary model is more likely to be selected if a series is adjusted for outliers. When forecasts are generated using each model, a non-stationary form will tend to adjust to new circumstances more readily. Of course, the extreme example is the random walk, which ‘walks’ away from the outlier just one period later. Thus, yet again, an ML method is likely to overfit the data, resulting in inferior forecasts.

E-mail address: ordk@georgetown.edu.

https://doi.org/10.1016/j.ijforecast.2019.02.014

How do we protect against overfitting? Traditional time series methods use an information criterion (IC), but the construction of a theory-based IC for ML methods seems out of reach. However, the best should not be the enemy of the good. Given a sample of size n and observed error terms $\{e_t, t = 1, 2, K, n\}$, one may set up a quasi-IC of the form:

Minimize $\sum_{t=1}^{n} e_t^2 + KP$.

Here, P might represent the total number of nodes and K a suitable constant (such as $K = 2$ for AIC or $K = \ln(n)$ for SBC). Better formulations may well exist, but the point is that the results of the M4 Competition indicate that overfitting may be a serious problem for ML methods. It is a question that needs further consideration, perhaps by comparing in-sample fit statistics with forecasting performances.

3. Representativeness

One issue for forecasting competitions is the extent to which the results have meaning for forecasting activities beyond the particular data set examined. Spiliotis, Kouloumos, Assimakopoulos, and Makridakis (2020, hereafter SKAM) note this concern and go to considerable lengths to examine the extent to which broader conclusions may be drawn. The partitions of the M4 series by recording frequency and subject matter provide additional insights into the performance.

However, the lurking question, "Are the data *representative*?", remains. Representative of the mythical set of all time series? Clearly not. As was stated by SKAM, "The main problem in providing such evidence is that obtaining a complete picture of the 'real world' is impossible in practice due to its limitless applications and types of data involved. Nevertheless, if a large collection of real and indicative data was available, it would be reasonable to exploit it to provide evidence for this argument".

From the perspective of the forecaster who is about to embark on a new forecasting exercise (for multiple series), how can the huge database provided by M4 be most useful? Petropoulos, Makridakis, Assimakopoulos, and Nikolopoulos (2014) conducted an extensive simulation study that showed how series attributes affected the performances of different forecasting methods. Their study indicates the need to select forecasting procedures according to circumstances, or 'horses for courses'.[1]

The analysis in SKAM follows the approach of Kang, Hyndman, and Smith-Miles (2017), who used key attributes such as the *strength of trend* and *strength of seasonality* (six in all) to describe each series. These attributes are then reduced to a two-dimensional (or sometimes three-dimensional) *instant space* using principal components analysis. Thus, each series in the M4 database has a profile based upon these characteristics. I suggest that our forecaster should analyze (a subset of) her series and record these attributes for those series. A sample of (several hundred or a thousand) series that mirror the patterns in the new series to be forecast may then be drawn from the M4 database. The summary results for the selected sample of series with similar attributes would provide insights to guide the choice of methods.

In summary, representativeness, like good art, is in the eye of the beholder (that is, the person with the new forecasting task). Nevertheless, she will certainly benefit from the insights provided by the M4 study.

References

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

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, *13*(3), e0194889.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74.

Ord, K., Fildes, R., & Kourentzes, N. (2017). *Principles of business forecasting* (2nd ed.). New York: Wessex Press.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). Horses for courses in demand forecasting. *European Journal of Operational Research*, *237*, 152–163.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality?. *International Journal of Forecasting*, *36*(1), 37–53.

Keith Ord is Professor Emeritus at the McDonough School of Business, Georgetown University. He has co-authored a number of papers on forecasting using state space methods, and on spatial statistics. He is co-author of the research monograph *Forecasting with Exponential Smoothing* and of the textbook *Principles of Business Forecasting*.

[1] In 2008, in the first edition of *Principles of business forecasting*, the publisher insisted on including a footnote explaining this term!