Journal of Business Research 68 (2015) 1692–1701

Contents lists available at ScienceDirect

Journal of Business Research

Simple versus complex selection rules for forecasting many time series

Robert Fildes [b], Fotios Petropoulos [a,b,*]

[a] Logistics and Operations Management Section, Cardiff Business School, Cardiff University, UK
[b] Lancaster Centre for Forecasting, Department of Management Science, Lancaster University, UK

ARTICLE INFO

Article history:
Received 1 June 2014
Received in revised form 1 January 2015
Accepted 1 January 2015
Available online 4 April 2015

Keywords:
Automatic model selection
Comparative methods
Extrapolative methods
Combination
Stability

ABSTRACT

Selecting the appropriate forecasting method for a large number of time series is a major problem for many organizational forecaster. Researchers propose various selection rules in order to enhance forecasting accuracy. The simpler approach for model selection involves the identification of a single method, which is applied to all data series in an aggregate manner, without taking into account the specific characteristics of a single series. On the other hand, individual selection includes the identification of the best method for each series, though it is more computationally intensive. A simple combination of methods also provides an operational benchmark. The current study explores the circumstances under which individual model selection is beneficial and when this approach should be preferred to aggregate selection or combination. The superiority of each approach is analyzed in terms of data characteristics, existence or not of a dominant method and stability of the competing methods' comparative performance. In addition, the size and composition of the pools of methods under consideration are examined. In order to assess the efficacy of individual model selection in the cases considered, simple selection rules are proposed, based on within-sample best fit or best forecasting performance for different forecast horizons. The analysis shows that individual selection works best when specific sub-populations of data are considered (e.g., trended or seasonal series), but also when the alternative methods' comparative performance is stable over time. A case study demonstrates the efficiency of the recommended selection strategy.

1. Introduction and literature review

Forecasters regularly face the question of choosing from a set of alternative forecasting methods. Where the task the forecaster faces is one of forecasting many series repetitively automatic approaches to selecting the appropriate method are needed — the forecaster has insufficient time to devote to selection for each time series in any one time period. The forecasting methods usually considered are simple, one of a limited range of extrapolative methods including such standbys as exponential smoothing. Two distinct approaches address this problem: aggregate selection where the totality of data series is analyzed and a method chosen and then applied subsequently to all the time series and individual selection, where, for a particular series, each method is compared and the best chosen to produce forecasts for that series (Fildes, 1989). Aggregate selection has the benefit of simplicity but in principle each different time series with its different characteristics (e.g. trend and seasonality, stability) would be better forecast by an individual model that matches those characteristics. Does individual selection generate these expected benefits in terms of improved accuracy? Fildes (2001) shows that if selection could be done perfectly then the gains would be substantial. So the question is worth asking — can practical model selection algorithms, that will lead to forecasting accuracy gains, be implemented? Is the additional effort and added complexity of adopting an individual selection process worthwhile? Additionally, the question is important because simple selection algorithms are implemented in commercial software such as SAP APO-DP.

The task of selecting an appropriate forecasting method is first conditioned by the problem context and the data available. Armstrong (2001), and Ord and Fildes (2013) providing a simplified version, propose selection trees that aim to guide the forecaster to an appropriate set of methods. The current study considers the more limited case of choosing between extrapolative forecasting methods where substantial data are available on which to base the choice. This problem has a long history of research, primarily by statisticians. Broadly, the approach adopted is to assume a particular class of model where selection is to take place within that class, for example within the class of ARIMA models. Accuracy measures based on within-sample fit to the available data are used in the selection, modified in various ways to take into account the number of estimated parameters in each of the models, penalizing more complex models. AIC and BIC are two widely used information criteria. Both are based on the likelihood, including a penalty depending on the number of model parameters. As such, AIC and BIC deal with the trade-off between complexity and goodness of fit of the model. Assuming normal errors, minimizing the AIC is asymptotically equivalent to minimizing the one-step-ahead forecast Mean Squared Error.

* Corresponding author at: Cardiff Business School, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK.
E-mail address: petropoulosf@cardiff.ac.uk (F. Petropoulos).

http://dx.doi.org/10.1016/j.jbusres.2015.03.028
0148-2963/

From the early days of forecasting comparisons, the issue of the strength of the relationship between out-of-sample forecasting accuracy (on the test data) and in-sample fit has been controversial with first Makridakis and Winkler (1989) and then Pant and Starbuck (1990) arguing that little if any relationship exists. Pant and Starbuck examine three different measures of fit and corresponding measures of forecasting performance with mean squared fitted error proving a particularly inadequate guide. But the other measures were not much better. If in-sample fit is inadequate as these authors have argued, then an alternative approach to selection is clearly needed. In response, the forecasting literature became increasingly satisfied with the naïve principle that what has forecast the most accurately, will forecast the most accurately on the out-of-sample data. To operationalize selection based on out-of-sample performance, the available data must be broken into the data used to fit the model (often called the training data), the data used to provide an estimate of out-of-sample fit (the validation data) and the test data where various selection approaches can then be compared.

Beyond the examination of in-sample measures of fit and their link to performance on test data, earlier empirical research has been sparse. Using the data characteristics of the series to predict performance is one approach with Shah (1997) and Meade (2000) demonstrating some success, but such selection rules are complex. Collopy and Armstrong (1992) also use series characteristics to develop rules that combine various extrapolative models depending on the data conditions. Rule-based forecasting has shown promising performance in various empirical comparisons. A contrasting approach which benefits from simplicity is to consider past performance as the critical factor predicting future performance. A recent contribution is from Billah, King, Snyder, and Koehler (2006) who consider selection within the class of exponential smoothing models where an overarching general model exists. Their results for a subset of the M3 data demonstrated that information criteria outperform the use of the validation data in selection: but as they remark, the sample of out-of-sample validation forecasts is small, which, they conjecture, might explain their findings. The differences are also small between selection methods so a reasonable conclusion to draw might be that selection is not worthwhile — but that only applies to their particular data set and the extrapolative forecasting methods they considered. However, with the M3 data, automatic individual selection based on Forecast Pro's algorithm (Goodrich, 2000) had earlier proved effective, beating most aggregate selection approaches post hoc. Crone and Kourentzes (2011) report additional work and using a different data set demonstrate the benefits of using out-of-sample error measures compared with in-sample. In short, earlier research has produced conflicting results.

The contradictory findings leads to the following observations: individual selection can never be worthwhile if a dominant aggregate forecasting method is identified in the data set. Moreover, selecting the best method individually will not provide significant benefits if the methods under consideration produce similar forecasts.

The present study provides evidence on the effectiveness of the various selection criteria introduced. Following on from the above argument, there is the need to vary the methods considered for selection and also the data sets on which selection algorithms are tested. Section 2 considers the forecasting methods in the selection comparisons and the error measures being used to assess their accuracy are introduced. Section 3 considers the meta-data set (part of the M3 database), introduces the simple selection rules and also explains the rationale behind the different segments of the data sets examined. Section 4 contains the empirical results and provides a discussion of the results. Section 5 comments on the practical implications and limitations of the current research, including a brief case study. Section concludes. The key question that this study addresses is under which circumstances can individual selection rules generate accuracy benefits.

2. Forecasting methods and accuracy metrics

2.1. Extrapolative forecasting methods

In this evaluation of selection methods typical practice is emulated such as that embedded in forecasting software. The forecasting methods considered are therefore chosen broadly to represent standard approaches but are not themselves nested in an overall model, such as in the exponential smoothing class of Billah et al. (2006). They have been chosen from those considered in the forecasting competitions, in particular the M3 competition (Makridakis & Hibon, 2000) in which a large numbers of series were analyzed and a large number of extrapolation methods have been compared. All are practical alternatives in commercial applications. Computer intensive methods such as neural networks are excluded. Therefore, the focus is on simple extrapolation methods, methods widely used in practice, and also including some that have demonstrated significant performance in past forecasting exercises. The simplest forecasting technique, random walk or naïve, where the forecast is the latest observation, is therefore included along with widely used models from the exponential smoothing family (ETS, Hyndman, Koehler, Snyder, & Grose, 2002), namely Simple Exponential Smoothing (SES), Holt, Holt–Winters, damped trend and damped with multiplicative seasonality. Despite their limited use in practice, ARIMA models have been included as they remain a standard statistical benchmark.

The exponential smoothing methods are estimated using the forecast package for R statistical software (Hyndman & Khandakar, 2008). The automatic ARIMA function (auto.arima) implemented in the same package is used to identify and estimate the ARIMA models. The auto.arima function conducts a stepwise selection over possible models and returns the best ARIMA model. One could argue that this advantages ARIMA over other methods (such as SES or Holt), as the automatic ARIMA function already aims to choose the best model from within a broad class of models. Hyndman and Athanasopoulos (2012) describe how the auto.arima function works, with details on the options allowed in the ARIMA modeling.

These methods are applied directly to the raw data. However, in prior large forecasting exercises, such as the M3-Competition (Makridakis & Hibon, 2000), the non-seasonal methods were applied to the deseasonalized data. Deseasonalization of the data is usually conducted with multiplicative classical decomposition, where the seasonal indices calculated are used for the reseasonalization of the final forecasts. More details on how the deseasonalization is applied in the current research can be found in the Appendix. In order to be in line with the results of this research, simple and widely used models (naive, SES, Holt, and damped) applied to the seasonally adjusted data, instead of the raw data, are considered. Lastly, the Theta model (Assimakopoulos & Nikolopoulos, 2000), which was the top performer in M3-Competition, is considered. More details on the Theta model appear in the Appendix. The full set of methods that this paper considers, along with the respective short names, appears in Table 1.

Table 1
Forecasting methods included in the experiment.

#	Method	Short name	Applied to
1	Naive	Naive	Raw data
2	Naive 2	DNaive	Deseasonalized data
3	SES	Expsmoo	Raw data
4	SES 2	DExpsmoo	Deseasonalized data
5	Holt	Holt	Raw data
6	Holt 2	DHolt	Deseasonalized data
7	Holt–Winters	HoltWint	Raw data
8	Damped	Damp	Raw data
9	Damped 2	DDamp	Deseasonalized data
10	Damped with multiplicative seasonality	DampMult	Raw data
11	Theta	Theta	Deseasonalized data
12	ARIMA	ARIMA	Raw data