An evaluation of simple forecasting model selection rules

Fildes, Robert and Petropoulos, Fotios

Lancaster University Management School

April 2013

Please cite this paper as:

Fildes, R. & Petropoulos F. (2013). *An evaluation of simple forecasting model selection rules* (LUMS Working Paper 2013:2). Lancaster University: The Department of Management Science.

Lancaster University Management School
Working Paper 2013:2

An evaluation of simple forecasting model selection rules

Robert Fildes and Fotios Petropoulos

The Department of Management Science
Lancaster University Management School
Lancaster LA1 4YX
UK

An evaluation of simple forecasting model selection rules

Robert Fildes[1] and Fotios Petropoulos[1,*]

April, 2013

Abstract

A major problem for many organisational forecasters is to choose the appropriate forecasting method for a large number of data series. Model selection aims to identify the best method of forecasting for an individual series within the data set. Various selection rules have been proposed in order to enhance forecasting accuracy. In theory, model selection is appealing, as no single extrapolation method is better than all others for all series in an organizational data set. However, empirical results have demonstrated limited effectiveness of these often complex rules. The current study explores the circumstances under which model selection is beneficial. Three measures are examined for characterising the data series, namely predictability (in terms of the relative performance of the random walk but also a method, theta, that performs well), trend and seasonality in the series. In addition, the attributes of the data set and the methods also affect selection performance, including the size of the pools of methods under consideration, the stability of methods' performance and the correlation between methods. In order to assess the efficacy of 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. Individual (per series) selection is contrasted against the simpler approach (aggregate selection), where one method is applied to all data series. Moreover, simple combination of methods also provides an operational benchmark. The analysis shows that individual selection works best when specific sub-populations of data are considered (trended or seasonal series), but also when methods' relative performance is stable over time or no method is dominant across the data series.

Keywords: automatic model selection, comparative methods, extrapolative methods, combination, stability.

[1] Lancaster Centre for Forecasting, Department of Management Science, Lancaster University Management School, Lancaster, LA1 4YX, UK

[*] corresponding author: f.petropoulos@lancaster.ac.uk

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 stand-byes as exponential smoothing. Two distinct approaches have been proposed for dealing with this problem: *aggregate selection* where the totality of data series are analysed 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). Both methods can be updated period by period. Aggregate selection has the benefit of simplicity but in principle different time series with their 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 we implement practical model selection algorithms that lead to forecasting accuracy gains? Is the complexity of individual selection worthwhile? Additionally, the question is important because simple selection algorithms are implemented in commercial software such as SAP APO.

The task of selecting an appropriate forecasting method is first conditioned by the problem context and the data available. Armstrong (2001), and Ord & Fildes (2013) offering a simplified version, have proposed selection trees that aim to guide the forecaster to an appropriate set of methods. Here we consider the more limited case of choosing between extrapolative forecasting methods where there are substantial data 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, e.g. 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, penalising more complex models. The two

best known criteria are the AIC and BIC based on mean square error with a penalty function for the number of parameters in the model, the latter penalising a larger number of parameters more.

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 & Winkler (1989) and then Pant & Starbuck (1990) arguing that there was little if any relationship at all. Pant & 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 it is necessary to break the available data 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. One distinct approach has been to use the data characteristics of the series to predict performance with Shah (1997) and Meade (2000) demonstrating some success, but such selection rules are complex. Collopy & 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 Billah, King, Snyder & Koehler (2006) consider selection within the class of exponential smoothing models where there is an overarching general model. Their results for a subset of the M3 data demonstrated 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 extrapolative forecasting methods they considered. However, with the M3 data, automatic individual selection based on Forecast Pro's algorithm (Goodrich, 2000) proved effective, beating most aggregate selection approaches post hoc. Further work has been reported by Crone & Kourentzes (2011) who, 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 conclusions leads to the following observations: individual selection can never be worthwhile if there is a dominant aggregate forecasting method in the data set and similarly, it will not provide benefits if the methods under consideration produce similar results.

This paper aims to provide evidence on the effectiveness of the various selection criteria we have introduced. Following on from the above argument, we will need to vary the methods considered for selection and also the data sets on which selection algorithms are tested. In section 2 we introduce the forecasting methods included in the selection comparisons and the error measures we will use to assess their accuracy. 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 data sets we analyse. Section 4 contains the results and discussion, with the conclusions drawn out in section 5. The key question we address is whether we can we identify individual selection rules which generate accuracy benefits beyond various simple benchmarks.

2. Forecasting methods and accuracy metrics

2.1 Extrapolative forecasting methods

In this evaluation of selection methods we wish to emulate typical practice such as that embedded in forecasting software. The forecasting methods we consider 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 larger numbers of series have been analysed and a large number of extrapolation methods have been compared. All

are practical alternatives in commercial applications. Computer intensive methods such as neural networks have been excluded. Therefore, we focus on simple extrapolation methods, methods widely used in practice, and also including some that have demonstrated significant performance in past forecasting exercises. In that sense, we consider the simplest forecasting technique, Random Walk or Naive, 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. Moreover, despite its limited use in practice, ARIMA models have been included as they remain a standard statistical benchmark.

We estimate the exponential smoothing methods using the forecasting package for *R* statistical software (Hyndman & Khandakar, 2008). We also use the Automatic ARIMA function (*auto.arima*) implemented in the same package to identify and estimate the ARIMA models. The *auto.arima* function conducts a search over possible models and returns the best ARIMA model.

In all cases mentioned above, the methods are applied directly to the raw data. However, in previous large forecasting exercises, such as the M3-Competition (Makridakis & Hibon, 2000), the non-seasonal methods were applied to the deseasonalized data. Deseasonalisation of the data is usually conducted with multiplicative classical decomposition, where the seasonal indices calculated are used for the reseasonalization of the final forecasts. In order to be in line with the results of this research, we also consider simple and widely used models (Naive, SES, Holt and Damped) applied to the seasonally adjusted data instead of the raw data.

Lastly, the Theta model (Assimakopoulos & Nikolopoulos, 2000), which was the top performer in M3-Competition, is considered. The Theta model proposes the decomposition of the data in two or more "Theta lines", defined by the short term curvature and any long term trend. These "Theta lines" are extrapolated separately, while final forecasts are derived from the combination. Again, the Theta model's procedure requires, prior to forecasting, seasonal adjustment of the data.

The full set of methods considered in this paper, along with the respective short names, is presented in Table 1.

Table 1. Forecasting methods

#	**Method**	**Short Name**	**Applied to**
1	Naive	Naive1	Raw data
2	Naive 2	Naive2	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

2.2 Measuring Forecast Error

Measurement of each method's forecasting performance is needed in two distinct phases of this research. Firstly, the forecasting performance of each method can be calculated over the validation data set (which we define rigorously in Section 3.1), and these measures can then be used in the selection of an appropriate method. The fit of the models in-sample can also be calculated. Secondly, metrics for measuring the performance of the methods and selection rules are necessary in order to assess the efficacy of the latter. Towards this direction, we consider measures for bias, accuracy and variance of the forecasting errors.

If we let $y_t(i)$ be the actual value of series *i* for time period *t* and $\hat{y}_t^m(i|h)$ be the point forecast of the same series for method *m* at forecast origin *t* for lead time *h*, then $EM_t^m(i|h)$ is the error measure of series *i* for method *m* at origin *t* for lead time *h*. Error Measure (*EM*) may be one of the following:

- Signed Error (E): $E_t^m(i|h) = y_{t+h}(i) - \hat{y}_t^m(i|h)$
- Squared Error (SE): $SE_t^m(i|h) = E_t^m(i|h)^2$
- Absolute Error (AE): $AE_t^m(i|h) = |E_t^m(i|h)|$
- Absolute Percentage Error (APE): $APE_t^m(i|h) = \left|\frac{E_t^m(i|h)}{y_{t+h}(i)}\right| \cdot 100\ (\%)$

Signed Error is used to demonstrate any consistent differences in terms of the bias of different approaches, whereas the other three are basic measures of the error deviations arising from the various forecasting methods.

We will rely on the mean out-of-sample performance (of series *i* for a method *m*) averaged over forecast origins (and potentially over forecast horizons). We give here examples used later in the paper: the general formulae are given in the Appendix.

A wide range of different error measures are available. A summary of the arguments surrounding their differences has recently been given by Davydenko & Fildes (2013). Based on this, we therefore present results for two measures, Mean Absolute Percentage Error (*MAPE*), and a relative error measure. *MAPE* is the Mean *APE* summarized across all *N* time series, as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} Mean\ APE_t^m(i|1)$$

that is the mean absolute percentage error averaged over series, for forecast horizon 1, and all available forecast origins, *t*. It can be easily extended to include multiple horizons as defined in the appendix. MAPE has been included here as it is the most widely adopted in practice (Fildes and Goodwin, 2007).

Relative error measures have the advantage of negating the effects of outliers somewhat showing how a forecasting method compares to a benchmark (such as the random walk). Summarizing relative errors across series using geometric mean has proved to be robust and more normally distributed than alternative measures. Define the relative Mean Absolute Error (over forecast horizon *h* for a series *i*) as:

$$r_i = \frac{MAE_i}{MAE_i^b}$$

where *b* is a benchmark method. Then

$$AvgRelMAE = \left(\prod_{i=1}^{m} r_i\right)^{1/m}.$$

It is also easily interpretable as showing the average percentage improvement (as measured by the MAE) from using one method (*m*) compared to the benchmark method (*b*).

We, also, need to calculate the past forecasting performance, *PFP*, which is the average of the *MAPE* or in the case of our relative measure, the *MAE*, over the available

validation data and across single or multiple lead times (full definitions are provided in the Appendix).

3. Experimental design

3.1 Forecasting procedure and database

Let T denotes the number of observations of an individual time series. Each series considered in this paper is divided in three time intervals. The first interval contains all observation from origin 1 to origin T_1, having a length of T_1, and acts as an initialisation interval, that is the training data. Observations from origins T_1+1 to T_2 are included in the second interval, the validation data, while the third interval contains observations between origins T_2+1 to T. The second and the third intervals have respectively length (T_2- T_1) and (T-T_2). Both corresponding sets of data are used as hold-out samples, meaning that forecasts are produced without prior knowledge of these values. Once the first set of forecasts is produced, using just the T_1 in-sample observations, one additional observation, the first observation of the validation data, is added to the in-sample data and new forecasts are calculated. This procedure is repeated until every single observation of the validation and test intervals is embodied into the in-sample vector. In other words, we employ rolling forecasting where the forecasts (and selected models) are updated at every single origin. As a result, (T_2- T_1) + (T-T_2) = T-T_1 sets of forecasts are calculated, each one containing h point forecasts, where h denotes the forecasting horizon considered.

The second interval is used only as validation data, in terms of evaluating single extrapolation methods and selecting the most appropriate one for forecasting each series (individual model selection) or a method to apply to all series (aggregate model selection). The third interval is used as both test data for the final evaluation of the selection rules proposed later in this paper, and as the forecast origin is rolled forward it also extends the associated validation data. Multiple lead-time forecasting enables the set-up of simple selection rules that apply to the various forecasting horizon.

The data series selected for this study are a sub-set of the monthly M3-Competition data set, where the total length of available observations is equal to or greater than T=126, giving a total of 998 series. Data series longer than the desired 126 observations are truncated. We set T_1=48 and T_2=90. Thus, the first set of forecasts is calculated from

time origin T_1 (=48). The forecasting horizon was set to h=18 periods ahead, to correspond earlier analyses of the same data (Makridakis & Hibon, 2000).

Upon the calculation of the point forecasts for each method using the first 48 data points, an additional point is added and a new set of points forecasts are calculated for each approach. This procedure is repeated until the origin 108, where the last 1-18 steps-ahead forecasts are produced. The remaining data points (observations 109 to 126) are only used in the evaluation of the last origin's forecasts. So, in total, we are producing 18 point forecasts for each origin (origins 48-108, 61 origins in total) and for each method (12), while the out-of-sample performance of all methods plus the model selection rules are evaluated through observations T_2+1 to T, the test data. The selection of the most appropriate method, based on past forecast performance of the methods in hand as calibrated over the validation data set, takes place at observations T_1+1 (=49) through T_2+k (periods 90 to 108 with k =0 to 18 indexing the test data). These 998 series provide the meta data set within which we will examine subsets of the data.

3.2 Choosing a best method

The objective of any selection rule is to choose the method at time t with the most promising performance. For the purposes of the current research, we examine various simple selection rules, based on the past forecasting performance (*PFP*) of each method as well as for the different lead time. Assuming that we want to perform model selection at the T_2+k origin, the *PFP* is measured between origins T_1 to T_2+k as is the fitted performance The method to be selected is the one with the most promising past performance. The four simple rules implemented and examined in this research are defined as follows:

Rule 1. Use the method with best fit as measured by the minimum one-step ahead in-sample Mean Squared Error (using all the data up to the forecast origin).

Rule 2. Use the method with the best 1-step-ahead forecast error, in terms of Mean Absolute Percentage Error, and apply that method to forecast for all lead times.

Rule 3. Use the method with best h-step-ahead forecast, in terms of Mean Absolute Percentage Error, and apply it to forecast for the same lead time.

Rule 4. Use the best method to forecast for all lead times as measured by Mean Absolute Percentage Error averaged over forecast horizons, 1 through h.

While the mathematical expressions are complex (as shown in the appendix) the ideas behind them are simple. Rule 1 selects the method that has fitted the data best and applies it to forecasting from forecast origin t over the next forecast horizons. Rules 2 to 4 select the best method depending on how the methods performed as measured on past forecast performance over the validation data set up to the forecast origin. Rules 1, 2 and 4 ignore any horizon effects while only Rule 3 attempts to match selection to the forecast horizon. The selections derived from these rules are updated over the test data set (i.e. as k increases), including all available errors in origins T_1 to T_2+k. Moreover, the proposed rules can be applied to aggregate selection where the error measures are summarized over all series and the best method is applied to all series, or to individual selection where a particular method is chosen for each series.

3.3 Research questions and preliminary analysis

The main objective of the current research is to investigate the conditions under which model selection may be beneficial. In order to achieve this objective, we consider two primary segmentations of the available time series. Firstly, data are classified as trended or not trended and seasonal or not seasonal. These categorisations have been chosen a priori based on the fact that some of the models are designed to incorporate trend and seasonal (e.g. ARIMA, Holt-Winters) whilst others (e.g. Random walk, Simple Exponential Smoothing) will introduce unnecessary error when applied to series with these characteristics. A further feature believed to affect relative performance is the predictability of the time series. We propose to define a specific time series as unpredictable if the performance of the non-seasonal Random Walk forecasting method (method 1) is better than the median performance of all other methods under investigation as defined by Mean Absolute Error in the validation data (from origins T_1 to T_2) for all forecasting horizons. Note that this classification is available to us ex ante and does not use the test data.

In terms of trend, we perform the robust Cox-Stuart test for trend on the 12-period centred moving average,to remove any contamination from seasonality. Lastly, the potential seasonal behaviour of the monthly series considered is tested by Friedman's non-parametric test. As a result, we consider six segments of the time series data set, namely "predictable", "unpredictable", "trended", "non-trended", "seasonal" and "non-seasonal" and this suggests the first research question..

RQ1. *Is individual model selection more effective when applied to groups of time series with specific characteristics?*

A second factor that may limit the value of individual selection is the number of models included in the pool of alternatives. Effectively a variant of over-fitting, the more models included, the higher the probability that the wrong model is chosen due to the randomness in the data. Given that the largest pool can be structured with all methods introduced in Section 2.1 (twelve in total), we also examine every possible combination of smaller pools of two (2) up to twelve (12) methods. For example, in the case of a pool of methods equal to four (4), all 495 possible pools of methods are checked, the number of 4-combination in a set of 12 or $\binom{12}{4}$. This leads to our second research question:

RQ2. *What are the effects on individual selection of including more methods in the pool under consideration?*

Many of the methods included in typical extrapolative selection competitions are similar which may be difficult to distinguish using a selection rule. The methods themselves produce correlated errors and these are shown in Table 2. Values under 0.5 are presented in bold. Many methods are highly correlated, most obviously those with similar seasonality components.

Table 2. Correlation of methods in terms of signed errors (all series)

Methods	Naive1	Naive2	Expsmoo	DExpsmoo	Holt	DHolt	HoltWint	Damp	DDamp	DampMult	Theta	ARIMA
Naive2	0.65	1.00										
Expsmoo	0.94	0.59	1.00									
DExpsmoo	0.60	0.91	0.64	1.00								
Holt	0.70	**0.42**	0.72	**0.43**	1.00							
DHolt	**0.46**	0.75	**0.46**	0.80	**0.46**	1.00						
HoltWint	**0.16**	**0.49**	**0.15**	**0.51**	**0.17**	0.53	1.00					
Damp	0.88	0.54	0.93	0.59	0.85	0.51	**0.17**	1.00				
DDamp	0.54	0.84	0.57	0.92	**0.47**	0.92	0.52	0.59	1.00			
DampMult	**0.40**	0.70	**0.41**	0.74	**0.33**	0.70	0.74	**0.41**	0.74	1.00		
Theta	0.59	0.89	0.63	0.98	**0.45**	0.83	0.52	0.59	0.92	0.75	1.00	
ARIMA	0.64	0.69	0.69	0.78	**0.50**	0.66	**0.34**	0.66	0.75	0.58	0.79	1.00

We see that Holt and Holt Winters have the fewest high correlations with the remaining methods. More promisingly for selection, we also note that ARIMA is the method least correlated with any of the others. Selection between methods that are similar cannot prove valuable. The average error correlation from the various methods participating in a specific pool is therefore examined. In order to measure the effect of correlation between the methods taking part in a specific combination, the combinations in each pool size are separated into high and low correlated; a certain combination is considered as highly correlated if the average correlation of the methods' outputs is equal to or greater than 0.7. Thus, the following research question deals with the effect of correlation among methods.

RQ3. *Do pools of methods with low correlation, in terms of forecast error, provide better forecasting performance when individual selection rules are considered compared to more highly correlated pools?*

Individual selection would be unlikely to be beneficial when a single method is dominant for the obvious reason that if a single method was appropriate for all series, selection rules would be dominated by the effects of noise. Table 3 presents summary one-step-ahead error statistics over the validation and test data sets for MAPE and AvgRelMAE. Also, the percentage of times where each method was ranked among the three top methods for each origin and each series is presented (column "Top 3").

Table 3. Performance of all methods measured by the one-step-ahead performance, averaged across series and origins.

Method	MAPE (%)			AvgRelMAE			Top 3 (%)		
	Origins								
	48-108 (All)	48-89 (initial selection sample)	90-108 (test sample)	48-108 (All)	48-89 (initial selection sample)	90-108 (test sample)	48-108 (All)	48-89 (initial selection sample)	90-108 (test sample)
Naive1	11.76	11.73	12.52	1.000	1.000	1.000	25.4	25.1	26.3
Naive2	10.93	10.98	11.74	0.899	0.908	0.939	24.5	23.8	26.1
Expsmoo	10.82	10.87	11.72	0.934	0.934	0.944	25.7	25.5	26.2
DExpsmoo	9.48	9.53	10.45	0.821	0.826	0.864	21.5	21.2	22.3
Holt	10.77	10.80	11.52	0.912	0.903	0.897	27.0	27.4	26.0
DHolt	9.45	9.53	10.37	0.811	0.811	0.842	24.7	24.7	24.5
HoltWint	9.65	9.59	10.20	0.844	0.837	0.862	25.5	26.0	24.5
Damp	10.77	10.84	11.67	0.903	0.897	0.900	27.4	27.7	26.7
DDamp	9.47	9.56	10.43	0.803	0.804	0.836	22.7	22.8	22.4
DampMult	9.49	9.48	10.13	0.804	0.801	0.830	24.2	24.6	23.3
Theta	9.34	9.39	10.27	0.808	0.806	0.838	20.0	19.9	20.27

ARIMA	9.97	10.05	11.32	0.795	0.792	0.816	31.1	31.0	31.4

Interestingly the three error measures show different performance rankings, particularly for the best performing methods with the largest discrepancies in the relative performance of ARIMA, Theta and DExspsmoo. Figure 1 illustrates the results in a different way showing how the relative performance of the different methods measured by one-step ahead *MAPE* and *AvgRelMAE* (relative to the Naïve model) changes over time. It is immediately apparent that the relative performance of Naïve 2 improves relative to Holt and that *AvgrelMAE*, with its more robust attribute, points to the better relative performance of ARIMA. For all series, Theta performs best as measured by *MAPE* though with *AvgRelMAE* ARIMA is the overall best performer.

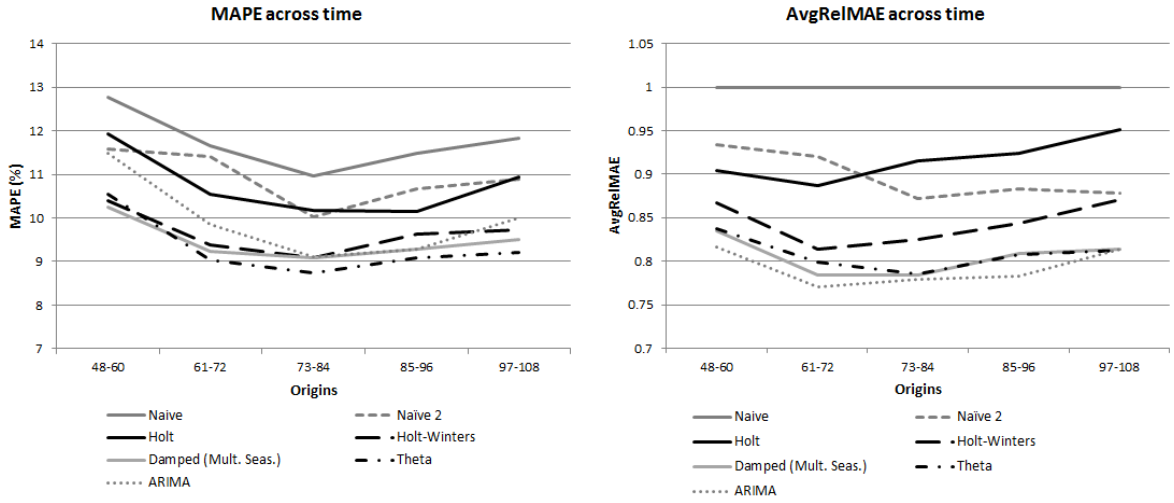

Figure 1. The relative performance of methods over time

A second segmentation is, therefore, considered: to divide the data series into two groups in terms of the performance of one of the best methods. For this purpose, we have chosen Theta. In the M3 Competition the Theta method proved a strong performer and as can be seen in Table 3 it also performs well over the subset of 998 series. The aim here is identify sub-populations where there is (or is not) a dominant method. The threshold for a specific time series to be grouped in one of the two groups will be the Theta model's achievement to be ranked (or not) among the top three (out of twelve) methods. In other words, past forecasting performance, as measured by Mean Absolute Error for the validation data, must be lower (or higher) than the value of the first quartile, that is:

1st Group: Theta's performance in the top three (measured by MAE)

2nd Group: Theta's performance outside the top three

RQ4. *Individual method selection is of most value when there is no dominant method across the population.*

We also analyse the performance of the different methods for their stability. Using the validation and test data and error measures calculated for leads 1-6 for each data point we can measure stability in a specific series by the average (across time origins) Spearman's rank correlation coefficient where the ranked performance of methods at each forecast origin is correlated to the rank of the average performance of the method summarized across all origins. A value of 1 implies the rankings of all methods remain the same over time. The median of our stability measure is 0.45 with range 0.01 to 0.91. Thus, we may segment further the 998 series considered in regards to the stability of methods' performance. We define a series as stable when its Spearman's rho falls in the top quartile of the data set (>0.59). As a result, we suggest the final research question:

RQ5. *Individual selection is only effective compared to aggregate selection when relative performance in the pool of methods under consideration is stable.*

In total, segmentations of data considered in the current research are summarized in Table 4, where the respective populations of the groups are displayed.

Table 4. Segmenting the data set: number of time series per segment

Segment	Number of series
Entire data set	998
Predictable	694
Unpredictable	304
Trended	894
Non-trended	104
Seasonal	608
Non-seasonal	390
Theta Best	428
Theta Worst	570
Stable performance	250
Unstable performance	748

4. Empirical results

4.1 Out-of-sample performance of methods

Firstly, we examine the out-of-sample performance of the forecasting methods considered in this study using error measures averaged across all origins through T_2 to

T_2+18 and across all lead times (1 to 18). Table 5 presents the results when *MAPE* is selected as error measure. Each row refers to a single extrapolation method (please, refer to Table 1 for the abbreviations). At the same time, each column refers to a specific segmentation of the data, as described in Table 4.

Even a quick view of this table unveils some very interesting observations. Firstly, across all segments, the best performance, in terms of accuracy is recorded for Theta followed by SES, when applied on the seasonally adjusted data, and seasonal versions of Damped. Over all series, Holt and Naive demonstrate the worst performance, neglecting as they do seasonality or any deterministic trend. Theta and Deseasonalised exponential smoothing, correlated at 0.98 from Table 2, perform very similarly for all segments and are the top performers. The largest differences across the methods are recorded for predictable and seasonal series, where methods with specific features, such as the ability to handle seasonality, perform better than benchmark methods, like random walk. On the other hand, simpler methods catch up with more complex ones when unpredictable on non-seasonal series are considered. The presence of trend or seasonality naturally favour methods with the ability to capture these features. SES on deseasonalized data (DExpsmoo) performs interestingly well and better than ARIMA.

The segment of data series containing the non-trended series suffers from relatively high levels of inaccuracy (an average MAPE across methods of 24.4% compared with 14.7% overall), almost doubling for some methods. As expected, Holt performs worst, failing to estimate the zero trend. When segmenting on the stability in the methods' performance is recorded, the performance of the non-seasonal methods is uniformly poor suggesting stability in performance is related to the ability of a method to capture persistent seasonality. For the 748 unstable series differences in performance are much smaller. Finally, the last row presents *MAPE* values for perfect information, meaning that the best method is selected for each series (individual selection) in an ex-post manner. It is apparent that possible margins of improvements are between 25 to 30% for all segments, compared to the best method in each segment applied to all series (ex-post aggregate selection). Therefore, individual model selection is worth investigating as Fildes (2001) had previous argued. In addition, the segmentation we are using emphasizes the importance of trend, seasonality and stability.

The out-of-sample performance analysis was performed for two more error measures, *MdAPE* and *AvgRelMAE*. The use of *MdAPE* results in significant lower errors, as expected. Decreases across the different methods on each segment are consistent, resulting in stable ratios of *MdAPE*/*MAPE*. The smallest improvements are recorded for unpredictable series, where *MdAPE* is, on average, 35% lower than *MAPE*, while the largest improvements (57%) are observed for non-seasonal data. SES's performance on deseasonalized data (DExpsmoo) is also enhanced relatively to other methods when *MdAPE* is examined, especially in the case of unstable data. *AvgRelMAE* confirms the superiority of Theta and DExpsmoo across all series and for most of the segments, with Naive1 being among the best methods for unpredictable and non-seasonal series. Lastly, the relative performances of Holt on deseasonalized data (DHolt) and Holt-Winters are, across all series, worse than that of Naive1.

Table 5. MAPE (%) of single methods across all lead times for the test data per segment of series.

	Entire data set	Predictable	Unpredictable	Trended	Non-trended	Seasonal	Non-seasonal	Theta Best	Theta Worst	Stable	Unstable
Naive1	17.1	18.2	14.7	15.6	29.8	21.1	10.9	19.1	15.7	20.9	15.8
Naive2	14.2	14.1	14.3	13.3	21.6	16.1	11.2	14.6	13.9	13.1	14.6
Expsmoo	15.7	16.4	14.3	14.3	28.2	19.3	10.2	17.2	14.6	19.3	14.5
DExpsmoo	12.8	12.4	13.6	12.0	19.6	14.3	10.4	12.8	12.8	11.5	13.2
Holt	18.1	19.0	16.1	16.3	33.2	22.5	11.2	20.5	16.3	22.9	16.5
DHolt	14.6	13.8	16.3	13.8	21.3	16.4	11.7	15.0	14.2	12.3	15.3
HoltWint	15.3	14.5	17.3	13.8	28.4	17.6	11.8	15.1	15.5	13.7	15.9
Damp	16.0	16.5	14.9	14.5	28.8	19.7	10.2	17.8	14.7	19.5	14.8
DDamp	13.0	12.3	14.7	12.2	19.9	14.6	10.6	13.0	13.1	11.1	13.7
DampMult	13.2	12.6	14.4	12.2	21.6	14.8	10.6	13.3	13.1	11.6	13.7
Theta	12.6	12.3	13.3	11.8	19.3	14.2	10.2	12.7	12.6	11.6	13.0
ARIMA	14.2	13.5	15.7	13.3	21.8	16.3	10.9	14.6	13.9	12.4	14.7
Perfect Information	9.2	9.1	9.5	8.7	14.1	10.7	7.0	9.6	9.0	8.4	9.5

4.2 Performance of selection rules

Having analysed the performance of single extrapolation methods, we will now present the performance results of the simple selection rules presented in section 3.2. We focus on the cases improved by performing individual selection versus the two simple benchmarks:

(i) aggregate selection. This uses the single best method based on the one-step-ahead out-of-sample performance on the validation sample

(ii) the combination of methods using equal weights to each method included in the selection pool. Thus, the accuracy gains (or losses) through using simple individual selection rules is examined through the percentage of cases where individual selection rules performed better than the above benchmarks.

Accuracy performance across all series in each case is calculated through Median Absolute Percentage Error (MdAPE). The results are presented for each segment of the data (Table 4) separately, and are segmented by the sizes of the pools of methods under consideration (e.g. a pool size of two when just ARIMA and Expsmoo are being considered) and the correlation of methods in a specific pool (e.g. ARIMA and Expsmoo have a low correlation). In each table, the first column displays the number of methods considered in a selection pool (i.e. the size of the pool) grouped into three classes (2-4, 5-8 and 9-12 methods). Columns 2 and 3 break down the selection pool into those groups of methods which are highly correlated on average and which are low, showing the number of combinations in each group. Each of the remaining columns refers to the improvements in performance for individual selection (as a percentage of comparisons) compared to the two benchmarks.. Improvements in more than 50% of the cases are presented with bold interface. “NA” means that no combinations for this specific group of pools sizes and correlation is available.

Table 6 presents the percentage of cases improved in terms of accuracy when all series are considered. Recall that Rule 1 uses in-sample fit, Rule 2, 1 period ahead past forecast performance, Rule 3 matches the lead time in individual selection while Rule 4 takes a more aggregate approach with selection based on average performance over all lead times.

Overall, there are fewer cases with methods characterized as highly correlated. The first observation is that the relative number of cases improved by individual selection is

higher when the rules are applied to low correlated or uncorrelated methods, especially when small pools are considered. Also as the pools' size increases, we observe that individual selection generally becomes better. At the same time, these improvements are only achieved for Rules 2 to 4, with Rule 1 (best in-sample fit) having the worst performance. Moreover, individual selection always outperforms both aggregate selection and combination in more than 80% of the cases when Rule 4 is applied to methods identified as low correlated.

Table 6. Percentage (%) of cases with on average improved forecasting accuracy when all series (998) are considered.

Methods in selection pool	Average Correlation	Number of cases	% of cases Individual Selection performed better							
			vs. Aggregate				vs. Combination			
			Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
2-4	Low	611	31.3	**65.0**	44.4	**86.4**	33.4	**75.0**	**83.1**	**90.0**
	High	170	13.5	**65.3**	**55.3**	**75.3**	18.8	**60.6**	**70.6**	**80.6**
5-8	Low	2712	17.4	**51.5**	47.7	**92.7**	38.6	**80.7**	**89.5**	**97.5**
	High	291	4.8	**78.7**	**62.2**	**90.7**	9.6	**84.5**	**82.1**	**95.9**
9-12	Low	295	3.7	46.4	41.7	**97.6**	45.1	**91.2**	**95.3**	**100.0**
	High	4	0.0	**100.0**	25.0	**100.0**	25.0	**100.0**	**100.0**	**100.0**

The cases displaying improvements for the 694 predictable and the 304 unpredictable series are presented in Table 7. For predictable series, the majority of the pools are characterized as low correlated while for the unpredictable series they are highly correlated. For predictable series, aggregate selection works best for almost all pools. Recall the definition of a predictable series is made on the validation data and therefore offers guidance on whether to use individual selection. In addition, for predictable series, aggregate selection works better than combination since in most combinations, consistently poorer methods are included.

Moreover, individual selection works efficiently against combination only for "predictable" series, with a much limited improvements for "unpredictable" series and just for large pools of methods. The exact opposite is observed against aggregate selection. So, model selection (in an individual or, especially, an aggregate manner) is more successful when investigating series identified as "predictable", with combination

being more robust in "unpredictable" series. Once again, Rules 2 to 4 perform best, while Rule 1 has very limited value, especially against combination. The results also prove robust across the range of error measures we have considered (*MAPE* and *AvgRelMAE*), with gains from using model selection being more apparent in predictable series, while individual selection outperforms aggregate selection when unpredictable series are investigated.

Table 7. Percentage (%) of cases with on average improved forecasting accuracy when predictability is examined.

	Methods in selection pool	Average Correlation	Number of cases	% of cases Individual Selection performed better							
				vs. Aggregate				vs. Combination			
				Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
Predictable	2-4	Low	698	23.8	44.6	28.2	**50.1**	45.8	**78.2**	**75.8**	**84.5**
		High	83	19.3	**53.0**	22.9	36.1	32.5	**56.6**	41.0	49.4
	5-8	Low	2948	11.2	24.3	14.0	34.7	40.3	**76.3**	**74.7**	**87.7**
		High	55	3.6	41.8	1.8	9.1	7.3	32.7	20.0	21.8
	9-12	Low	299	1.0	10.0	3.7	21.1	37.8	**80.9**	**87.0**	**96.3**
		High	0	NA	NA	NA	NA	NA	NA	NA	NA
Unpredictable	2-4	Low	47	**83.0**	**76.6**	**93.6**	**89.4**	12.8	19.2	**53.2**	**57.4**
		High	734	43.9	**60.6**	**78.8**	**74.9**	10.6	27.5	40.5	45.9
	5-8	Low	4	**100.0**	**100.0**	**100.0**	**100.0**	0.0	0.0	25.0	50.0
		High	2999	23.6	**61.5**	**90.0**	**84.4**	1.7	23.4	45.7	**51.8**
	9-12	Low	0	NA	NA	NA	NA	NA	NA	NA	NA
		High	299	8.0	**71.2**	**98.0**	**97.0**	0.0	18.1	**60.2**	**69.6**

Table 8 presents the percentage of cases improved in terms of accuracy when trended and non-trended series are considered, respectively. When trended data are examined, the majority of pools are identified as highly correlated. Individual selection versus aggregate selection seems to work reliably only for Rules 2 and 4. At the same time, Rules 2, 3 and 4 results in significant improvements when contrasting individual selection to simple combination of methods.

On the other hand, when non-trended data are examined, almost all pools are classified as low correlated. Improvements in more than 50% of the cases are now limited to only low correlated pools of methods. Individual is better against aggregate

selection for small pools of methods. In contrast, the bigger the pool of methods, the larger the percentage of cases improved against combination. For all cases, low correlated pools, in terms of methods' outputs, produce higher improvements in contrast to higher correlated ones. One plausible explanation is the ability of selection to identify methods that include trend. We therefore investigated the cases improved when trended (or non-trended) series are extrapolated only with methods with the potential for incorporating (or excluding) trend. As expected, excluding non-trended methods when extrapolating trended series, results in better forecasts for simple combinations, though individual selection is still best for smaller pools of methods. No significant differences are recorded in the case of non-trended series.

Table 8. Percentage (%) of cases with on average improved forecasting accuracy when trend is examined.

	Methods in selection pool	Average Correlation	Number of cases	% of cases Individual Selection performed better							
				vs. Aggregate				vs. Combination			
				Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
Trended	2-4	Low	201	46.3	**63.2**	**57.2**	**80.1**	19.9	**67.7**	**67.2**	**83.1**
		High	580	13.1	**53.6**	41.4	**60.2**	22.1	**67.9**	**70.7**	**79.0**
	5-8	Low	263	36.9	48.7	46.4	**75.7**	9.9	**76.8**	**66.2**	**87.8**
		High	2740	6.7	39.7	37.2	**66.7**	16.1	**71.0**	**70.0**	**84.2**
	9-12	Low	0	NA	NA	NA	NA	NA	NA	NA	NA
		High	299	0.7	20.1	27.8	**77.3**	10.7	**81.3**	**84.3**	**98.7**
Non-Trended	2-4	Low	711	28.6	47.0	**66.5**	**61.6**	26.7	**75.5**	**87.1**	**85.5**
		High	70	31.4	24.3	30.0	27.1	8.6	21.4	34.3	38.6
	5-8	Low	2954	21.0	23.8	45.3	35.0	35.7	**68.0**	**84.4**	**82.4**
		High	49	10.2	8.2	8.2	6.1	4.1	32.7	28.6	36.7
	9-12	Low	299	14.0	7.0	20.4	18.7	**56.5**	**60.9**	**90.6**	**88.3**
		High	0	NA	NA	NA	NA	NA	NA	NA	NA

According to Table 9, improvements for seasonal data are substantial, especially when Rules 3 and 4 are applied, suggesting reliance on 1-step ahead forecasts for monthly seasonal forecasting is unwise. Both against aggregate selection and combination, individual selection effectiveness increases as we consider more methods in a selection pool. Essentially, selection is capturing the persistent seasonality in the

series. Improvements are higher against aggregate selection in the case of high correlated pools of methods, while, as expected, the reverse is true against the combination of methods. Individual selection does not usually work when non-seasonal series are examined.

The examination of only seasonal methods applied to seasonal series did not demonstrate any significant differences in the percentage of cases improved. On the other hand, percentage of cases improved using individual instead of aggregate selection raised when non-seasonal series are extrapolated with only non-seasonal methods, but the small number of cases (26 combinations) does not allow any strong generalisations.

Table 9. Percentage (%) of cases with on average improved forecasting accuracy when seasonal component is examined.

	Methods in selection pool	Average Correlation	Number of cases	% of cases Individual Selection performed better							
				vs. Aggregate				vs. Combination			
				Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
Seasonal	2-4	Low	678	24.3	46.5	**58.6**	**72.7**	33.8	**75.5**	**97.2**	**97.8**
		High	103	20.4	48.5	**74.8**	**77.7**	21.4	46.6	**77.7**	**74.8**
	5-8	Low	2893	7.9	29.2	**75.0**	**83.8**	40.1	**73.1**	**98.9**	**99.4**
		High	110	0.0	30.0	**90.9**	**93.6**	0.9	18.2	**86.4**	**90.9**
	9-12	Low	299	0.3	10.7	**95.7**	**96.3**	**59.9**	**79.3**	**100.0**	**100.0**
		High	0	NA	NA	NA	NA	NA	NA	NA	NA
Non-Seasonal	2-4	Low	3	**100.0**	**66.7**	**100.0**	**100.0**	0.0	33.3	0.0	0.0
		High	778	36.0	49.1	47.6	**51.5**	6.0	16.1	9.0	12.5
	5-8	Low	0	NA	NA	NA	NA	NA	NA	NA	NA
		High	3003	22.3	25.0	18.4	37.7	0.5	4.5	0.8	2.7
	9-12	Low	0	NA	NA	NA	NA	NA	NA	NA	NA
		High	299	17.7	4.0	1.0	22.4	0.0	0.0	0.0	0.0

We have also analysed the results from segmenting the series using the performance of the Theta method (Table 10). We first examine the segment containing the series that Theta fell into the top three performers. Unsurprisingly aggregate selection is the best option (although recall Theta is not necessarily included in each case). The advantages of using individual selection are way more promising when applied to Theta Worst segment where no other method proves dominant. Rules 2 and 4 work very well for

individual selection versus aggregate selection and combination with some cases where the percentage of times performing better is greater than 90%. Once more, the gains are significantly higher when pools of methods with low correlated outputs are examined.

Table 10. Percentage (%) of cases with on average improved forecasting accuracy when Theta's performance is considered.

	Methods in selection pool	Average Correlation	Number of cases	% of cases Individual Selection performed better							
				vs. Aggregate				vs. Combination			
				Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
Theta Best	2-4	Low	717	11.3	27.3	20.2	28.3	39.6	**68.1**	**73.6**	**80.9**
		High	64	17.2	28.1	23.4	21.9	20.3	**51.6**	45.3	**53.1**
	5-8	Low	2986	3.9	11.8	7.0	20.0	**57.0**	**69.4**	**81.7**	**88.5**
		High	17	0.0	0.0	0.0	5.9	11.8	29.4	41.2	**52.9**
	9-12	Low	299	0.3	0.7	0.7	15.7	**76.6**	**80.9**	**93.3**	**96.3**
		High	0	NA	NA	NA	NA	NA	NA	NA	NA
Theta Worst	2-4	Low	261	46.7	**59.0**	**59.8**	**75.9**	42.9	**74.0**	**70.9**	**85.1**
		High	520	26.2	**56.9**	45.6	**54.4**	23.3	**72.3**	**61.9**	**73.8**
	5-8	Low	681	30.4	**55.8**	**52.4**	**65.5**	26.3	**83.4**	**74.6**	**88.5**
		High	2322	13.9	36.4	30.6	39.3	9.8	**62.5**	46.1	**67.0**
	9-12	Low	11	9.1	**72.7**	**63.6**	**63.6**	0.0	**100.0**	**100.0**	**100.0**
		High	288	3.5	19.8	17.4	20.1	4.5	67.4	35.1	72.2

Finally, Table 11 contrasts any differences when series are segmented with regards to the stability of methods' ranked performance. This classification results in pool sizes with low average correlation for methods with stable performance. With stability in a method's performance it is of course easier to identify the best individual and also aggregate selection. Improvements against aggregate selection are therefore limited in contrast to combination, with Rules 2 and 4 being effective. Moreover, the improvements in both cases get bigger as we consider larger pools of methods.

On the other hand, series with unstable methods' performance are characterized mostly by highly correlated point forecasts. The differences in cases improved are also substantial with combination typically outperforming selection apart from the situation of highly correlated methods. Individual selection does however improve over aggregate selection.

Table 11. Percentage (%) of cases with on average improved forecasting accuracy when stability of methods' performance is examined.

	Methods in selection pool	Average Correlation	Number of cases	% of cases Individual Selection performed better							
				vs. Aggregate				vs. Combination			
				Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
Stable	2-4	Low	722	31.3	**65.0**	37.4	**55.8**	**62.5**	**87.8**	**89.6**	**92.4**
		High	59	18.6	**61.0**	42.4	**50.8**	23.7	**67.8**	**52.5**	**52.5**
	5-8	Low	2987	17.3	49.1	32.4	**68.6**	**76.9**	**95.4**	**95.6**	**98.1**
		High	16	6.2	**50.0**	18.7	**50.0**	12.5	**56.3**	43.7	**50.0**
	9-12	Low	299	3.7	17.4	26.1	**83.3**	**91.0**	**99.0**	**99.0**	**100.0**
		High	0	NA	NA	NA	NA	NA	NA	NA	NA
Unstable	2-4	Low	83	**72.3**	**71.1**	**74.7**	**80.7**	7.2	6.0	25.3	39.8
		High	698	33.5	45.8	**52.7**	**63.8**	13.3	19.8	38.7	**51.7**
	5-8	Low	26	**73.1**	**73.1**	**73.1**	**76.9**	0.0	0.0	7.7	30.8
		High	2977	22.0	30.6	40.2	**69.6**	4.6	12.7	27.8	49.9
	9-12	Low	0	NA	NA	NA	NA	NA	NA	NA	NA
		High	299	13.4	12.0	15.0	**82.9**	1.3	7.0	12.4	**51.8**

Table 12 presents the summarised *MdAPE*s for aggregate selection, simple combination and individual selection (applied by Rule 4) when the selection pools contain 2, 6 or 10 methods. For all cases, the formation of larger pools of methods results in better accuracy results with lower variance, thus better forecasting performance. Moreover, when comparing aggregate selection and combination versus individual selection, the latter results in lower median *MdAPE* (especially for large pools), while the variance is also lower in the small pools. Essentially, individual selection is on average slightly more accurate than the two simple benchmarks but more importantly, it is more reliable.

The same comparisons have been applied to the various segmentations. The results show relatively little gains in terms of accuracy, with largest differences arising in the cases of seasonal and stable series. At the same time, individual selection results in the lowest interquartile ranges for most of the times. This observation proves robust across the range of error measures we have considered, in particularly *MAPE* and the robust *AvgRelMAE*.

Table 12. Summarized *MdAPE*s across series for aggregate selection, simple combination and individual selection.

Methods in a selection pool	Aggregate Selection			Combination			Individual Selection (Rule 4)		
	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
2	7.36	7.51	8.45	7.55	7.85	8.10	7.33	7.50	7.81
6	7.36	7.36	7.51	7.48	7.61	7.82	7.16	7.24	7.32
10	7.36	7.36	7.36	7.50	7.57	7.72	7.13	7.15	7.20

4.3 Discussion

The empirical findings of this study provide some interesting evidence on the efficiency of selection rules. First, segmenting the series helps us to identify suitable sub-populations of data with specific characteristics, where the application of individual selection is more effective (RQ1) compared to the simples rules of aggregate selection or combination. Individual selection is particularly effective against our two simple benchmarks for seasonal and trending series as well as those series with no dominant method (theta worst). Individual selection works well against combination for all segments, except for unpredictable, non-seasonal and unstable segments. These are all segments where the risk averaging aspect of combinations was expected to work well. Aggregate selection works most effectively where there is a dominant stable method, as Fildes (1989) shows in analysing a method, robust trend, designed for the specific data set, or when data are identified as predictable.

RQ2 questioned the effects of including more methods in the pool under consideration. In most cases improvements from individual selection over aggregate selection are recorded when small pools of methods are considered. However, when comparing against combination, more methods in the selection pool generally results (except in the case of non-seasonal series) in a larger numbers of cases improved. This is because of the incorporation of methods not performing particularly well in the combination while the selection pool benefits is not disadvantaged by the poorer methods. In the case of non-seasonal segmentation, the differences among methods are relatively small (Table 5) so there is little or no loss and the usual benefits of combination are available.

With regard to the correlation of the methods in the selection pool, it is obvious from almost all segments of data that segments identified as low correlated offer the greatest

opportunity for individual selection. Specifically, when individual selection is contrasted against combination, the cases improved from selecting pools containing methods identified as low correlated are in some cases almost double (for example, Table 12, Rules 3 and 4 with median improvements of over 9% for pools with up to four methods). In answer to RQ3 therefore, pools of methods with low correlation generally provide a better foundation for individual model selection. However, some exceptions apply, specifically seasonal series and series characterized as unstable when comparing with aggregate selection and combination respectively.

Aggregate selection is expected to produce better results than individual selection, when a single method displays dominant performance across a specific sample of series (RQ4). The hypotheses is verified through segmenting the data in series where Theta achieved (or not) a ranking among the top three methods (out of twelve in total). Aggregate selection works better than individual selection in the Theta Best sample of series, with median improvements of 1.6%. On the other hand in the Theta Worst series where there is no dominant method, aggregate selection produces lower improvements (up to 0.5%). In addition, individual selection displays significant gains over combination for both segments (whether a specific method is dominant or not). Combination has, on average, 4.5% higher *MdAPE* than aggregate selection.

Lastly, as expected, when stability in methods' ranked performance is used as a basis of segmentation, individual selection produces more accurate forecasts for most of the pools of methods examined and especially for Rules 2 and 4 (RQ5). Stability in performance of methods enables the accurate selection of the most appropriate method individually, with average performance improvements of 2.8% and 16% against aggregate selection and combination respectively. This is a direct result from the great differences in the performance of methods over these series (Table 5). This difference also allows individual selection to work effectively with larger pools of methods without any negative consequences. On the other hand, for the unstable segment, the combination of methods is the most robust choice, displaying the smallest *AvgRelMAE*. At the same time, individual selection outperforms aggregate selection for the smaller pools examined.

In our introduction we did not speculate on the effectiveness of the different selection rules, merely noting the existing evidence was conflicting. In the results we have presented Rule 1 based on a measure of fit is uniformly ineffective compared to rules based on the validation sample and in particular Rule 4 which looks at aggregate performance averaged over lead times.

Rule 3 which attempts to match selection for a specific forecast horizon to its corresponding past accuracy on the validation sample performed poorly.

5. Conclusions

When forecasting a population of time series, individual selection of the most appropriate method is intuitively appealing and may result in substantial gains. In the current research we analysed the circumstances under which selection of an individualised method per series should be preferred to selecting a single method (aggregate selection) for the whole population of series or by a combination of methods. To explore the conditions when individual selection is most likely to be of benefit, the entire data set was segmented into sub-populations with regard to basic series characteristics (predictability, trend and seasonality). Moreover, we examined the efficacy of individual selection when a specific method is dominant and when the methods' performance are stable across forecast origins. Lastly, we considered the effect of the number of methods taking part in selection (pool size) and the correlation between methods.

Empirical results, based on the long monthly series of the M3-Competition provided insights with regards to the effectiveness of individual selection versus the simple rules of aggregate selection or combination. When a population of series is divided in sub-populations with specific characteristics, then selection per series is more effective, especially for series identified as seasonal, without a dominant method (Theta worst) or non-trended. In addition, individual selection is superior when methods' ranked performance in each series is stable. On the other hand, aggregate selection is the best choice when one single method is dominant across a sub-population of series, while combination is efficient when there is a lack of stability. Finally, with some exceptions, individual selection works better (especially against aggregate selection) when small pools of uncorrelated methods are selected.

Of the various individual selection rules we considered, Rule 4, which relied on aggregated forecast performance over horizons, proved better than relying on 1-step ahead rules, or even Rule 3 which matched selection to the corresponding horizon. Simply relying on past performance over the fitted data proved inadequate.

The practical implications of the current research are significant. Given that we only considered simple and widely used extrapolating methods, the outcomes are generally applicable. The insights provided can be directly applied to broadly used ERPs and Forecasting Support Systems (e.g. SAP), as to further enhance the integrated automatic selection procedures. SAP uses a selection rule applied across all series using 10 similar base methods. However, a single method approach, which would combine seasonal and trend features (such as DDamp or Theta), continues to work well, with minimum losses when contrasting its median performance against aggregate or individual selection.

A natural path for future research is to extend the range of methods to include ones with distinctive performance characteristics, such as Neural Networks. Moreover, the selection rules used in this study are only based on model fit and past forecast performance of methods across single or multiple lead times. These could be enhanced by a large number of variables proposed in the literature (Shah, 1997; Meade, 2000; Adya, Collopy, Armstrong & Kennedy, 2001). Lastly, the current research does not fully take into account the specific features of each extrapolating method, with all pools of possible methods being handled in the same manner. This is done in an attempt to gain a holistic view on the effectiveness of the individual selection rules over the aggregate selection and simple combination of methods. However, in a managerial set up it would provide additional insight to include only the appropriate methods in the selection pool that match the characteristics of a specific sub-population of data (e.g. trended or seasonal series). As ever with forecasting competitions, an extension of the range of series considered to distinct homogenous populations should prove illuminating, not least to identify the populations where the performance differences between extrapolative methods are large enough to be important.

We conclude by noting that for many applications selection rules are likely to deliver improved forecast accuracy. While for most sub-populations (if the population we have analysed here is informative) the gains are not usually large, the reliability is improved. While aggregate selection, perhaps the standard simple rule in application, can clearly deliver where there is a specific stable structure to the time series population (e.g. the telecoms data of Fildes(1989)), where the data are more heterogeneous as here, individual selection is needed. A final note, the simple rule of combining proved ineffective for most of the segmented data sets.

References

Adya, M.; Collopy, F.; Armstrong, J. & Kennedy, M. (2001), 'Automatic identification of time series features for rule-based forecasting', *International Journal of Forecasting* **17**, 143-157.

Armstrong, J. S., ed. (2001), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Boston and Dordrecht: Kluwer.

Assimakopoulos, V. & Nikolopoulos, K. (2000), 'The Theta model: a decomposition approach to forecasting', *International Journal of Forecasting* **16**(4), 521 - 530.

Billah, B.; King, M. L.; Snyder, R. D. & Koehler, A. B. (2006), 'Exponential smoothing model selection for forecasting', *International Journal of Forecasting* **22**(2), 239 - 247.

Collopy, F. & Armstrong, J. (1992), 'Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations', *Management Science* **38**, 1392-1414.

Crone, S. & Kourentzes, N. (2011), "Automatic Model Selection of Exponential Smoothing - an empirical evaluation of Trace Errors in forecasting for Logistics"31st Annual International Symposium on Forecasting ISF 2011, June 26-29, 2011, Prague, Czech Republic.'.

Davydenko, A. & Fildes, R. (2013), 'Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts', *International Journal of Forecasting*(0), - .

Fildes, R. (2001), 'Beyond forecasting competitions', *International Journal of Forecasting* **17**, 556-560.

Fildes, R. (1989), 'Evaluation of aggregate and individual forecast method selection rules', *Management Science* **39**, 1056-1065.

Goodrich, R. L. (2000), 'The Forecast Pro methodology', *International Journal of Forecasting* **16**(4), 533 - 535.

Hyndman, R. J. & Khandakar, Y. (2008), 'Automatic Time Series Forecasting: The forecast Package for R', *Journal of Statistical Software* **27**(3), 1 - 22.

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.

Makridakis, S. & Hibon, M. (2000), 'The M3-Competition: results, conclusions and implications', *International Journal of Forecasting* **16**(4), 451 - 476.

Makridakis, S. & Winkler, R. L. (1989), 'Sampling distributions of post-sample forecasting errors', *Applied Statistics-Journal of the Royal Statistical Society Series C* **38**, 331-342.

Meade, N. (2000), 'Evidence for the selection of forecasting methods', *Journal of Forecasting* **19**, 515-535.

Ord, K. & Fildes, R. (2013), *Principles of Business Forecasting*, South-Western Cengage Learning.

Pant, P. N. & Starbuck, W. H. (1990), 'Innocents in the Forest - Forecasting and Research Methods', *Journal of Management* **16**, 433-460.

Shah, C. (1997), 'Model selection in univariate time series forecasting using discriminant analysis', *International Journal of Forecasting* **13**, 489-500.

Appendices

A1. Error measures

Let us define the error made in forecasting series from time origins t_2 to t_2 averaged over horizons h_1 to h_2 as

$$Mean\ EM^{m}_{t_1,t_2}(i|h_1,h_2)_i = \frac{1}{h_2-h_1+1}\sum_{h=h_1}^{h_2}\left(\frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2} EM^{m}_{t}(i|h)\right)$$

Mean Absolute Percentage Error (*MAPE*) is the Mean *APE* summarized across all *N* time series, as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} Mean\ APE^{m}_{t_1,t_2}(i|h_1,h_2)$$

that is the mean absolute percentage error averaged over series, forecast horizons and origins.

The relative Mean Absolute Error for a series *i* can be defined as:

$$r_i = \frac{MAE_i}{MAE_i^b}$$

where MAE_i^b – MAE for baseline forecast for series i, MAE_i^m – MAE for method *m* for series i. MAE_i^b and MAE_i^m can be obtained from the arithmetic mean absolute error averaged across all forecast origins and forecast horizons h_1 to h_2 for series *i*:

$$MAE_i^b = Mean\ AE^{b}_{t_1,t_2}(i|h_1,h_2)$$

$$MAE_i^m = Mean\ AE^{m}_{t_1,t_2}(i|h_1,h_2)$$

Davydenko & Fildes (2013) showed that when making comparisons between methods, the use of arithmetic means rather than geometric can lead to misinterpretations. Instead, they proposed the use of a geometric average relative MAE.

$$AvgRelMAE = \left(\prod_{i=1}^{m} r_i\right)^{1/m}.$$

As is standard practice, we also use Absolute Percentage Errors and Squared Errors in the simple selection approaches in order to select the most promising single forecasting approach from a specific pool of methods. The average *Past Forecast Performance* (*PFP*) of series i for a method m for origins t_1 through t_2 may be calculated as the performance over a fixed lead time (h) or multiple lead times (h_1 to h_2) measured by an *EM* as follows:

Single lead time: $${}_{EM}PFP^{m}_{t_1,t_2}(i|h) = \frac{1}{t_2-t_1}\sum_{t=t_1}^{t_2-h} EM^{m}_{t}(i|h)$$

Multiple lead times: $${}_{EM}PFP^{m}_{t_1,t_2}(i|h_1,h_2) = \frac{1}{h_2-h_1+1}\sum_{h=h_1}^{h_2} {}_{EM}PFP^{m}_{t_1,t_2}(i|h)$$

Note that in the special case where we are only interested in one horizon, h_1=h_2, the two equations are equivalent.

A2. Model Selection Rules

Assuming that we want to perform model selection at the T_2+k origin, the *PFP* is measured between origins T_1 to T_2+k as is the fitted performance The method to be selected is the one with the most promising performance. To this direction, we select the method with the minimum error (the smallest *PFP*), for the different lead times:

Single lead time: $$Best\ Method = argmin\left[{}_{EM}PFP^{m}_{T_1,T_2+k}(i|h), m\right]$$

Multiple lead times: $$Best\ Method = argmin\left[{}_{EM}PFP^{m}_{T_1,T_2+k}(i|h_1,h_2), m\right]$$

In the following, the four simple rules implemented and examined in this research are defined. These rules are applied, as previously mentioned, in a rolling origin matter. As such, the most appropriate method identified and applied for the calculation of the forecasts for the next origin may change over time. Nevertheless, in each origin h point

forecasts are calculated. Note that in all cases *m* is the index referring to each one of the methods examined by a specific rule.

Rule 1. Use the method with best fit as measured by the minimum one-step ahead in sample Mean Squared Error:

$$Best\ Method_{T_{2+k}}\ (i|for\ all\ lead\ times) = argmin\left[{}_{SE}PFP^{m}_{T_1,T_2+k-1}(i|1), m\right]$$

Rule 2. Use the method with the best 1-step-ahead forecast error, in terms of Mean Absolute Percentage Error, and apply that method to forecast for all lead times:

$$Best\ Method_{T_{2+k}}(i|for\ all\ lead\ times) = argmin\left[{}_{APE}PFP^{m}_{T_1,T_2+k}(i|1), m\right]$$

Rule 3. Use the method with best *h*-step-ahead forecast, in terms of Mean Absolute Percentage Error, and apply it to forecast for just the same lead time:

$$Best\ Method_{T_{2+k}}(i|for\ lead\ time\ h) = argmin\left[{}_{APE}PFP^{m}_{T_1,T_2+k}(i|h), m\right]$$

Rule 4. Use the best 1-18 steps-ahead, in terms of Mean Percentage Absolute Error, method to forecast for all lead times:

$$Best\ Method_{T_{2+k}}(i|for\ all\ lead\ times) = argmin\left[{}_{APE}PFP^{m}_{T_1,T_2+k}(i|1{,}18), m\right]$$