Another look at forecast selection and combination: evidence from forecast pooling

Nikolaos Kourentzes[a,*], Devon Barrow[b], Fotios Petropoulos[c]

[a]*Department of Management Science, Lancaster University Management School, Lancaster University, UK*
[b]*Faculty of Business, Environment and Society, Coventry University, UK*
[c]*School of Management, University of Bath, UK*

Abstract

Forecast selection and combination are regarded as two competing alternatives. In the literature there is substantial evidence that forecast combination is beneficial, in terms of reducing the forecast errors, as well as mitigating modelling uncertainty as we are not forced to choose a single model. However, whether all forecasts to be combined are appropriate, or not, is typically overlooked and various weighting schemes have been proposed to lessen the impact of inappropriate forecasts. We argue that selecting a reasonable pool of forecasts is fundamental in the modelling process and in this context both forecast selection and combination can be seen as two extreme pools of forecasts. We evaluate forecast pooling approaches and find them beneficial in terms of forecast accuracy. We propose a heuristic to automatically identify forecast pools, irrespective of their source or the performance criteria, and

*Correspondance: N Kourentzes, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK. Tel.: +44-1524-592911
Email address: nikolaos@kourentzes.com (Nikolaos Kourentzes)

demonstrate that in various conditions it performs at least as good as alternative pools that require additional modelling decisions and better than selection or combination.
Keywords: Forecasting, Model Selection, Forecast Combination, Forecast Pooling, Cross-Validation

1. Introduction

There is nearly 40 decades of research and empirical evidence in favour of forecast combination over selection (Elliott and Timmermann, 2016; Barrow and Kourentzes, 2016). While the correct identification of the best forecast for a given time series can lead to significant gains in accuracy and dependent decisions (Fildes, 2001; Strijbosch et al., 2011; Fildes and Petropoulos, 2015), the uncertainty associated with identifying a 'best model' makes this a challenging problem. These include sample, parameter and model uncertainty (Breiman, 1996; Kourentzes et al., 2014a). Different sample size will result in different parameter estimates, which in turn may result in different model forms. Parameter estimation uncertainty may originate from the estimation algorithm and setup; for instance different initial values may result in different estimates. Different model structures may impose specific restrictions in parameters, simplifying, or not, the estimation problem, and so on. Given these uncertainties, a standard approach in forecast building is to use multiple alternative forecasting models or methods and pick the one that is identified as most appropriate, given the data at hand.

Assuming that we consider a family of models to produce the forecasts, we can rely on information criteria, such as the Akaike Information Criterion (AIC, Akaike, 1974). More generally, when we consider multiple model families or forecasts without a formal model, some appropriate fit criterion or cross-validation statistic can be used (Fildes and Petropoulos, 2015; Barrow and Crone, 2016). Naturally, using different criteria may lead to different forecast selections and all these criteria are subject to the aforementioned uncertainties. Therefore, they are not guaranteed to result in the best possible forecasting performance. Several of these criteria, especially those based on likelihood or one-step ahead in-sample fit, suffer from an additional limitation: implicitly they assume that the postulated forecasting model is true. Otherwise, the likelihood function is not appropriate for any multi-step forecast that we require from the model (Chatfield, 2000; Xia and Tong, 2011). Fildes and Petropoulos (2015) provide empirical evidence of the disadvantage of one-step ahead forecast based selection criteria.

Given these challenges, alternatively we can combine multiple forecasts. There is ample literature that discusses why combinations of forecasts are beneficial, or how to best perform these (for examples see, Timmermann, 2006; Kolassa, 2011; Aye et al., 2015; Elliott and Timmermann, 2016). However, the quality of the combined forecasts is always dependent on the individual forecasts that are combined. Although this is an intuitive point, it is often overlooked in the literature, where these forecasts are assumed to be as required, for example, uncorrelated (Clemen, 1989; De Menezes et al.,

2000) or having sufficient diversity (de Menezes and Bunn, 1998; Brown et al., 2005; Lemke and Gabrys, 2010) or not encompassed (Fang, 2003; Harvey and Newbold, 2005). Alternatively the issue of forecast quality is subsumed in the question of how to best weight the different forecasts that are combined (de Menezes and Bunn, 1998; De Menezes et al., 2000; Tian and Anderson, 2014). For example, Granger and Ramanathan (1984) propose using a restricted regression to estimate the combination weights of different forecasts, where a forecast can in principle be attributed zero weight and therefore effectively excluded. Given that the estimation of weights is subject to the various uncertainties previously discussed, several alternative weighting schemes have been proposed (an empirical evaluation is provided by Barrow and Kourentzes, 2016), while a common finding in the literature is that unweighted combinations perform very competitively (Timmermann, 2006), the latter effectively not excluding any forecasts.

It becomes obvious that a caveat in combination approaches is that they assume that the forecasts to be combined are reasonable. As a mental experiment, consider dealing with a series that exhibits no seasonality or trend and combining only seasonal forecasts. Unweighted combinations will fail, as will more complex approaches such as using AIC derived combination weights (Kolassa, 2011). Since only seasonal forecasts will be combined, irrespective of the weights, the resulting final forecast will be inappropriate. To overcome this, an additional step can be considered: forecast pooling, which instructs that from the complete set of forecasts only a subset is deemed relevant to

be combined. For example, Aiolfi and Timmermann (2006) investigate the construction of forecast pools by using forecasts that belong to arbitrarily chosen top performing quantiles, or based on clustering methods. The authors find that pooling can be beneficial, but recognise that the methods proposed depend on multiple subjective choices by the modeller. Geweke and Amisano (2011) conclude that pooling performs well, even when the true model in not part of the considered models, contrasting results of typical selection or weighting schemes. This is very relevant to practice, as in business forecasting this is the norm.

In this paper we investigate pooling for business forecasting. We take the view that the forecast selection criteria and the different approaches that are used to combine forecasts, can be considered as two independent types of operations, that follow pooling. For example, forecast selection can be seen as nothing more than combining 'all' forecasts from a pool of a single top performing forecast. On the other hand, model combination assumes that all forecasts add value and therefore are retained in the pool, yet the combination weights are based on criteria that have equivalences to the criteria used for model selection. For example, AIC-weights are the combination analogue to AIC model selection (Burnham and Anderson, 2002). Forecast selection and forecast combinations can be seen as the two extremes of a spectrum that is defined by forecast pooling, combined with some selection/weighting operator (figure 1).

This paper has three aims: (i) demonstrate empirically the usefulness of

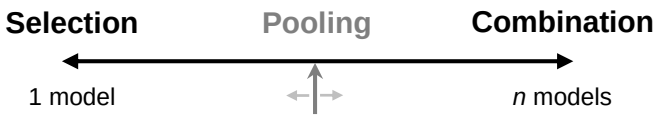


Figure 1: The spectrum of forecast pooling, with selection and combination featuring as extreme cases, irrespective of the scoring criterion.

pooling in a variety of conditions (e.g., forecasting at a disaggregate product level or at a more aggregate one, and at different tiers of a supply chain); (ii) propose a heuristic that can automate the selection of the pool size that is independent of the performance criterion, therefore making it easily implementable in practice and in existing setups; and (iii) focus the attention of the business forecasting literature beyond the dichotomy of model combination or selection, as both are potential outcomes of pooling.

The practical importance of this work is that it provides an approach to increase forecast accuracy further, with minimal assumptions or requirements for the individual forecasts that contribute to the pool. It is easy to accommodate various selection and combination operators, making it applicable to a variety of existing forecasting support systems, and incorporate innovations in either forecasting combination or selection. At the same time, the forecast selection problem is avoided, simplifying the forecasting process for users, with the decision making benefits stemming from forecast combinations. For instance, in terms of inventory management, forecasting combinations have been shown to result in lower requirements for safety stocks (Barrow and Kourentzes, 2016).

The paper is organised as follows; in section 2 an overview of forecast selection and combination is presented, followed by section 3 that discusses pooling and introduces the proposed forecast islands. Section 4 outlines the empirical evaluation that is conducted to benchmark the performance of pooling and presents the results. Section 5 discusses further properties of the proposed pooling. Section 6 concludes with final remarks and future research.

2. Selection and combination of forecasts

Although the question how to best select or combine forecasts is not resolved, there has been a lot of research in both areas. In the following subsections, we summarise the main findings and highlight the most common approaches to perform these.

2.1. Selection

Identifying and selecting the best forecast is a challenging task. It is so as there are many uncertainties that one needs to consider, but also because what a 'best' forecast is often ill-defined. In the literature there have been several advances in both aspects, yet there is no widely accepted best method to select a forecast.

Traditionally simplistic criteria, such as the coefficient of determination (R^2), and its adjusted version that penalises models with more parameters, have been commonly used to choose between alternatives. These have been

shown to suffer from multiple weaknesses, in particular in a predictive context, where more advanced metrics, such as information criteria, are nowadays the norm (Burnham and Anderson, 2002; Montgomery et al., 2015). Nonetheless, simplistic criteria can often still be useful in the absence of a full forecasting model. Irrespectively, criteria that do not penalise in-sample fit will typically lead to choosing over-fitted methods or models and should be avoided.

Akaike (1974) proposed an information criterion to select the best model amongst various alternatives, which came to be known as the Akaike Information Criterion (AIC). It balances the quality of fit, as measured by the likelihood function for the model $\mathcal{L}$, and its complexity as measured by the number of parameters. Due to Burnham and Anderson (1998), who used an information-theoretic approach to ground AIC, its use became more widespread.

In using AIC we have to keep note of its requirements. AIC requires: model parameters to maximise its likelihood; and the various alternative models considered to be estimated on the same sample (Burnham and Anderson, 2002). The latter translates to same sample size, but also scale, which implies that even within a model family certain transformations do not allow us to compare AIC values of competing models. For example, Autoregressive Integrated Moving Average (ARIMA) models, when formulated using the Box-Jenkins notation (Box et al., 2015), cannot be compared on AIC when the order of differencing varies, as the sample is not identical.

Related metrics have appeared in the literature, for example, for small sample sizes the AICc is preferable, while the Bayesian Information Criterion (BIC Schwarz, 1978) imposes a stronger penalty for model complexity than AIC. An obvious question is which information criteria to use in practice. The literature provides diverging opinions (for a discussion see Burnham and Anderson, 2002; Yang, 2005), however the work by Billah et al. (2005) is interesting in that it demonstrates that although the criteria may result in different selection of models, in terms of predictive accuracy the differences are small.

Information criteria attempt to balance the quality of model fit against its complexity, so as to avoid over-fitting. A more direct approach to this is to use separate sample to fit the various alternative models and another to choose the best one. This falls under the general framework of cross-validation (CV). CV in its basic implementation separates the available sample in s separate subsets. A model is estimated s times, using each time $s-1$ subsets as fitting set and the one remaining sample to evaluate the out-of-sample performance of the model. This is repeated until all s samples have been used as out-of-sample. The hold-out performance is then averaged, providing the cross-validation error. There are several variations of the basic cross-validation idea (for a comprehensive review see Arlot et al., 2010), however many cannot be applied in a time series context. Often time series models try to capture time dynamics in the series, which does not permit splitting the time series in any desirable way. Note that depending on the model used

different types of cross-validation may become feasible (Barrow and Crone, 2016). For example, cross-validating multiple regression models is trivial. When autoregressions are present these have to be accounted for, and the problem becomes even more challenging when moving average components are considered.

For time series models, at minimum, we can split a series into a fitting and a validation set, where the later follows the first. We can then measure the error in the validation set to choose the appropriate model. This 'hold-out' approach has been used successfully in the past (Makridakis et al., 1998). Fildes and Petropoulos (2015) explored this in detail and found that using rolling forecasts in the validation set of suitable steps-ahead forecast resulted in appropriate selection of forecasts, as judged by forecasting accuracy. In this format, cross-validation can be applied to any time series forecasting model, and notably method. In contrast to information criteria no likelihood expressions are required, nor even optimal parameters (as these cannot be always uniquely defined for ad-hoc methods). Furthermore, the evaluation metric can be any that the user deems appropriate.

It has been shown that AIC and BIC are connected to forms of cross-validation asymptotically (Stone, 1977; Shao, 1997; Burnham and Anderson, 2002; Fang, 2011). Although this well known result is often used as an argument to the sufficiency of information criteria over cross-validation, which is more complex to implement, in practice this argument has limited importance, as cross-validation allows us to consider a variety of performance

metrics, relevant forecast horizons and can be applied generally. However, cross-validation comes at a sample cost, since it requires the use of both fitting and validation sets, while information criteria need only fitting sample. At the same time, the quality of the cross-validation result depends on the size of the validation set and its representativeness, a problem that has been discussed to a great extent in the forecast evaluation literature (Tashman, 2000).

As different selection approaches may lead to different forecasts being selected, a relevant question is what constitutes a 'best' forecast. The forecasting competition literature has discussed this issue in length (Makridakis and Hibon, 2000; Ord, 2001; Fildes and Ord, 2002). Yet, it is clear that the relevant application driven forecasting objective should be considered, at each case. This can often mean that a sub-optimal forecast with respect to the alternative approaches discussed above, may still be desirable as it exhibits other useful properties. For example in a production or inventory management setting, a very robust to shocks forecast will lead to easier planning which may outweight potential inaccuracy costs. Another relevant example is when the selection method would advise switching forecasting models too frequently, with adverse effects to planning, but also to the confidence that users will put in statistical forecasting.

2.2. Combination

An alternative to selecting a single forecast is to combine different forecasts in an aggregate one. It is widely accepted that forecast combinations are beneficial (Elliott and Timmermann, 2016), leading to a reduction of forecast error variance, as well as mitigating the forecast selection problem. Futhermore, Chan and Pauwels (2018) demonstrate that simply selecting a single model on cross-validated errors will lead to biased selection, supporting a combination approach. The research has focused mainly on two questions: which are useful combination operators and what are the best combination weights.

Typically, the combined forecast is constructed as a linear combination of the initial forecasts:

$$\tilde{\mathbf{y}} = \hat{\mathbf{Y}}\mathbf{w}, \tag{1}$$

where $\tilde{\mathbf{y}}$ is the resulting vector of combined forecasts for various forecast horizons, $\hat{\mathbf{Y}}$ a matrix containing the separate individual predictions for the various forecast horizons and $\mathbf{w}$ is a vector of combination weights. Forecast combination research has primarily focused on the problem of weight estimation (Newbold and Granger, 1974; Granger and Ramanathan, 1984; Diebold and Pauly, 1990; Kolassa, 2011; Tian and Anderson, 2014; Elliott and Timmermann, 2016). A full overview of various alternatives to derive combination weights is beyond the scope of this paper, however we will draw some analogues to the selection approaches discussed above.

Burnham and Anderson (2002) provide an extensive discussion of combination weighting schemes that are based on information criteria, which Kolassa (2011) demonstrates to result in superior accuracy over selection by information criteria. One such common approach is using AIC-weights. Given a set of k forecasting models, for which comparable $\text{AIC}(k) = \{\text{AIC}_i\}$ and $i = 1, \ldots, k$ are available, the following steps can be used to derive the combination weights $\mathbf{w}$:

$$\Delta\text{AIC}_i = \text{AIC}_i - \min(\text{AIC}(k)),$$
$$w_i = \frac{e^{\left(-0.5\Delta\text{AIC}_i\right)}}{\sum_{i=1}^{k} e^{\left(-0.5\Delta\text{AIC}_i\right)}}.$$

Similar weights can be derived from AICc and BIC. Akaike weights calculated in this manner can be interpreted as being the probability that a given model is the 'best' model, given the model set and data.

Although there is accuracy evidence supporting the use of information criteria for combination over selection, one should not overlook the more subtle benefits. The parameters of any model will be estimated given some uncertainty; combining forecasts will mitigate this. Furthermore, selecting a single model assumes that the resulting model is close to the underlying true data generating process, if one exists (Chatfield, 2000). On the other hand, combination avoids this strong assumption and in particular when AIC-weights (or similar) are employed the various models are weighted according to the evidence of how appropriately each model describes the observed data (Burnham

and Anderson, 2002). Similar arguments can be made for other weighting schemes. However, Geweke and Amisano (2011) notes that even with combination, many weighting schemes will tend to show preference to a small number of models, even when none of these is the true underlying model.

Naturally, following the same logic, one can construct ad-hoc weights from any selection metric, expanding the calculation of combination weights across different model families and forecasting methods. Barrow and Crone (2016) show that cross-validation, in its various forms, can be used for forecast combination, improving forecasting accuracy over model selection.

Empirically unweighted combination has been found to perform very well, often at least as good as complex weighting schemes (Genre et al., 2013; Elliott and Timmermann, 2016), even though the later appear to be more theoretically elegant. Smith and Wallis (2009) argue that this is due to estimation uncertainty of the combination weights. Claeskens et al. (2016) distinguish between fixed and random combination weights and demonstrate the importance of the weight estimation uncertainty, explaining further the strong performance of simple , sub-optimal, weighting schemes. Elliott (2011) notes that there is second part to this argument, that is the relative gain from optimally combining forecasts, given any losses due to estimation uncertainty. He provides forecast clustering (most experts erring on the same side of the actual) as an example where combination gains would be relatively small and out-weighted by estimation issues. Petropoulos and Kourentzes (2015) provide a similar argument as to why forecast combinations do not seem to

provide benefits for intermittent demand forecasts.

Irrespective of the combination weighting scheme, the other line of research in forecast combination has looked at the combination operator. Equation (1), when no constant is incorporated, is a weighted average. Agnew (1985) found that the median outperformed the mean, and recommended its use. Barrow and Kourentzes (2016) found the median performing best amongst a large variety of alternative combination schemes, as it was robust against outlying forecasts. Alternatively, one can employ the trimmed mean (Elliott and Timmermann, 2016). On the other hand, Stock and Watson (2004) found support for the mean, while McNees (1992) found no significant differences between the two. Kourentzes et al. (2014a) compared the use of mean, median and mode of forecasts, as estimated using kernel density, and found that the mean required a substantial number of forecasts to converge to a stable good forecasting performance, while the median converged very fast. When an adequate number of forecasts was available for the kernel density estimator (around 30) then the mode performed best. However, weighting schemes have not been explored for such combination operators, although such extensions are simple.

In the literature, combination of forecasts have been gaining popularity, resulting in more exotic approaches. Examples of this are algorithms such as the Random Forests, that combines multiple decision trees for classification and regression problems (Breiman, 2001), bagging of time series to improve the performance of exponential smoothing (Bergmeir et al., 2016), and com-

bining forecasts from different temporally aggregated versions of the data (Kourentzes et al., 2014b; Kourentzes and Petropoulos, 2015; Athanasopoulos et al., 2017), amongst others. The motivation behind all these combination approaches is to reduce the modelling uncertainty and avoid relying on a single model, while potentially gaining accuracy benefits. Nonetheless, forecast combination has parallels to model selection, and although the choice of a single model is avoided, the modeller fundamentally has to decide on the criterion to assess the performance of each forecast to be combined, so as to construct appropriate combination weights, given a combination operator.

3. Pooling methods

With pooling a subset of the available forecasts is used instead of using all available ones. The aim is to attempt to reduce forecast errors further, while improving computational efficiency and lowering the number of forecasting approaches that need to be maintained by the users.

As we reasoned in the introduction, pooling is a separate step from selecting the criteria to rank the forecasts or how the combination is performed. The first is associated with the allocation of appropriate combination weights, or forecast selection if the pool becomes a single model. For instance, these could be based on information criteria (see examples by Burnham and Anderson (2002) and Kolassa (2011) for the respective combination weights or Chatfield (2000) and Hyndman et al. (2002) for forecast selection), or forecast errors, considering error correlation, error variance and error covariance

(Bates and Granger, 1969; Newbold and Harvey, 2002; Timmermann, 2006; Barrow and Kourentzes, 2016). The second is associated with the operator that is used for constructing the combined forecast, which at a fundamental level can be unweighted mean, median or mode, or weighted variants that were discussed in section 2.2.

In the forecasting literature there is limited discussion of pooling. De Menezes et al. (2000), based on a review of prior research, suggest combining forecasts which are uncorrelated, to avoid high weight estimation errors (Miller et al., 1992). Approaches using the error covariance matrix are not without issues. The estimation of the error variances and covariance can be challenging due to limited sample size, changes of the behaviour of forecast errors over time and other unexplained variations that may occur in the data (Newbold and Harvey, 2002; Tian and Anderson, 2014; Elliott and Timmermann, 2016).

Aiolfi and Timmermann (2006) take a different approach. They argue in favour of conditional combination strategies, as they find that there is strong evidence of persistence for top and bottom performing forecasts. They investigate grouping the forecasts either by assigning them into quartiles or k-means clusters based on their historical performance. The authors then consider a variety of ways to combine the forecasts within a pool. Geweke and Amisano (2011) demonstrate the benefits of pooling in predicting S&P 500 returns, noting that models that are clearly inferior by the usual scoring criteria, result in well performing pooled forecasts. Elliott (2011) combines aspects of unweighted averaging and optimal combination weights using the

Best Subset Averaging procedure to construct pools of forecasts. This procedure performs robustly against unweighted average or optimal weights combinations, in a variety of settings. Matsypura et al. (2017) successfully use pooling to combine expert forecasts.

Below we discuss in more detail the use of quantiles to form groups and propose a heuristic that allows us to automatically identify appropriate cut-off points from forecast pools. The key advantage of the latter is that it does not rely on an appropriately chosen quantile by the modeller. Although we discuss pooling without assuming a specific performance criterion, the resulting pools will depend on that. This is in line with our previous argument that pooling provides a data driven continuum between selecting a single forecast or combining.

3.1. Top quantiles

Let $C = \{c_i\}$ be the values of an appropriate criterion to assess the forecasts for $i = 1, \ldots, k$ forecasts. This criterion can be an information criterion like AIC, a CV statistic, or a weaker metric, such as the adjusted R^2. Depending on the criterion this may be based on in-sample data (for instance AIC or adjusted R^2) or some validation sample (for example for cross-validated errors).

Irrespective of the nature of C we rank forecasts from best to worst performing, and use the top quantile to form a forecast pool to be combined. It is up to the modeller to decide what quantilisation to use. The extreme cases

are using k-quantiles, where using the top is equivalent to forecast selection, and using 1-quantile, where all forecasts are included in the single quantile and is equivalent to forecast combination. If we use quartiles, we can pool together the top 25%, 50% or 75% of the forecasts. However, there is no statistic to guide our choice and the cut-off point is selected in an arbitrary manner.

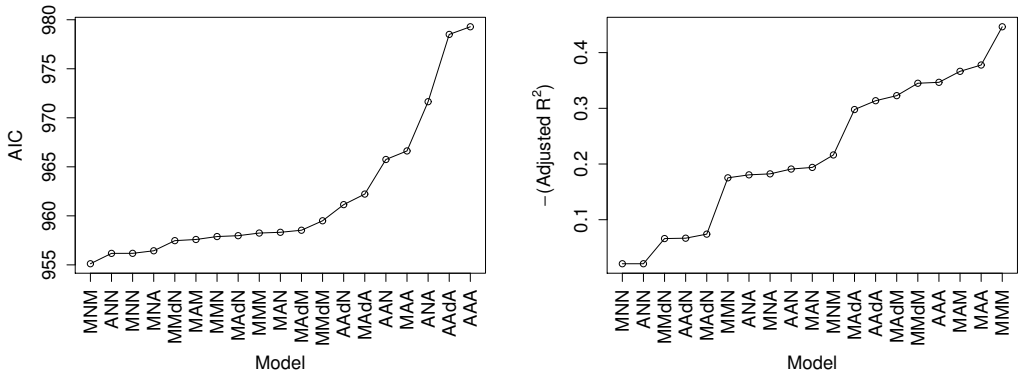
Note that using the top quantiles, as described here, is different from the use of quartiles described by Aiolfi and Timmermann (2006), who combine all forecasts within the different quartiles and then perform a weighted combination of the combined quartile forecasts. Their motivation is to reduce the number of forecasts for which optimal combination weights need to be calculated, rather than reducing the number of base forecasts used that we focus on here.

3.2. Forecast Islands

We propose a heuristic to form forecast pools, irrespective of whether the forecasts are originating from models for which a likelihood can be calculated, or forecasting methods that lack such derivations.

Given some appropriate criterion of performance C, first, we transform it so as to ensure that a smaller value is better. No change is required for information criteria such as AIC, AICc and BIC, or CV statistic which are already sorted in this manner. For metrics, such as adjusted R^2 where a higher value is better, we multiply them by -1. Next we order the forecasts

from best to worst. Figure 2 provides an example for AIC and Adjusted R^2 considering 19 alternative exponential smoothing (ETS) models (for the example we use the first monthly series of the M3 competition dataset, N1402; see section 4.1). The models are named following the convention introduced by Hyndman et al. (2008b); ETS(Error, Trend, Season), where each component can be: 'N' for none, 'A' for additive, 'M' for multiplicative. For the trend component an additional letter 'd' indicates damped trend. Observe that AIC provides a gradual increase, while for Adjusted R^2 several steep increases are observed.

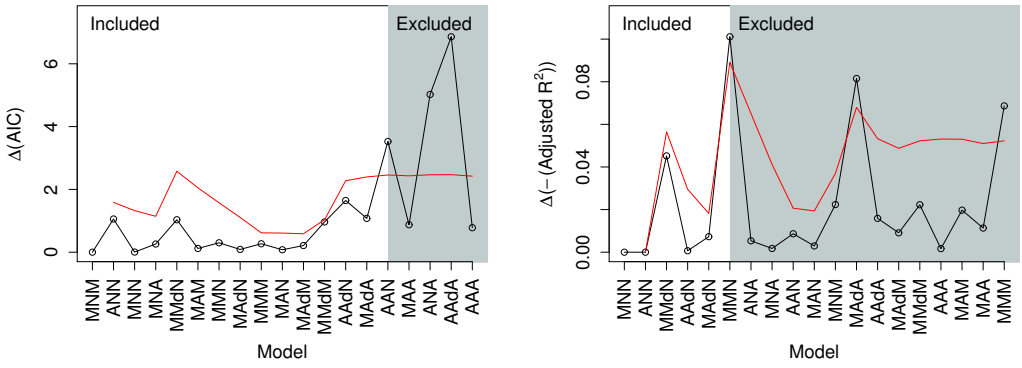


(a) AIC values (b) $-$(Adj. R^2) values

Figure 2: Sorted metric values for the 19 alternative ETS models considered.

Next, from the sorted metric we construct $C' = \{0, \Delta C\}$, where Δ is the differencing operator, and a 0 is included for the first forecast, which is assigned no value by ΔC. C' captures the rate of increase of the metric assigned to each forecast. Based on this, we include in the pool all forecasts until the first steep increase.

To detect the first steep increase we resort to using the same approach used for detecting outliers in boxplot, i.e. $T = \text{Q3} + 1.5\text{IQR}$, where Q3 is the 3rd quartile and IQR is the inter-quartile range. T is calculated gradually as each additional forecast is considered, as illustrated in figure 3. We include all forecasts in the pool up until $C' \geq T$. Observe how the different metrics in our example provide different pools of forecasts. Note that the calculation of T is appropriate as C' does not exhibit any trend, due the differencing in its construction.



(a) C' for AIC (b) C' for $-$(Adj. R^2)

Figure 3: C' together with threshold T, as it is updated for each additional forecast considered. Once $C' \geq T$ we stop adding forecasts to the pool.

Once the pool has been identified then the forecasts included can be combined, using any desired combination approach. Note that the process described here is identical to considering k-quantiles with the arbitrarily selection of the cut-off point of how many quantiles to use, replaced with the proposed approach.

We name this approach forecast 'islands', due to the several small groups

of forecasts that can be seen in figure 2. Although the proposed approach is ad-hoc, it can be applied in a wide range of cases, as it does not require the forecasts to be outputs of models. Any method or even judgmental forecasts can be used with an appropriate performance metric. Even if model based forecasts are available, there is no need for these to belong to the same family, again assuming the use of an appropriate metric, such as CV statistic.

If forecasts are linearly combined, using (1), the combination weights for model independent criteria, such as cross-validated errors, with x_i being the value of the criterion for $i = 1, \ldots, k$ different forecasts, are calculated as:

$$w_i = \frac{x_i^{-1}}{\sum_{i=1}^{k} x_i^{-1}}. \tag{2}$$

This expression ensures that $\sum_{i=1}^{k} w_i = 1$ and that $0 \leq w_i \leq 1$.

3.3. When should we expect pooling to be beneficial?

The existing theoretical framework of the combination literature focuses on identifying when the variance of the combined forecast improves. Side-stepping the weight estimation issue (for a summary the reader is referred to Elliott and Timmermann, 2016), we compare the variance of combination of all forecasts (y_c) and a pooled forecast (y_p) with given (fixed) weights. Let $\mathbf{y} = (y_1, \ldots, y_k)'$ vector of unbiased forecasts, $\mathbf{w} = (w_1, \ldots, w_k)'$ the respective combination weights with $\sum_i w_i = 1$, and $\mathbf{\Sigma}_{yy}$ the finite variance of $\mathbf{y}$. The combined forecast is $y_c = \mathbf{w}'\mathbf{y}$ and its variance $var(y_c) = \mathbf{w}'\mathbf{\Sigma}_{yy}\mathbf{w}$ (Claeskens et al., 2016).

Pooling can be seen as an operation to eliminate forecasts (y_i) from the combination, which is reflected in assigning zero weights to these forecasts. Suppose the weights are calculated as $\mathbf{w} = \mathbf{x}^{-1}\mathbf{1}/\mathbf{1}'\mathbf{x}^{-1}\mathbf{1}$, where $\mathbf{x}$ is a diagonal matrix containing the values of the criterion used for each forecast and $\mathbf{1}$ is a vector of k ones. Note this is the same as (2). We can devise a $k \times k$ matrix $\mathbf{p}$, where the diagonal for any included forecast is equal to 1 and all other elements are zero and calculate the pooled weights as:

$$\mathbf{w_p} = \frac{(\mathbf{pxp})^{-1}\mathbf{1}}{\mathbf{1}'(\mathbf{pxp})^{-1}\mathbf{1}}. \tag{3}$$

Note that by construction $\mathbf{pxp}$ is singular, as its determinant will always be zero due to $\mathbf{p}$. However, given the structure of $\mathbf{p}$ and $\mathbf{x}$, its pseudoinverse is simply $\mathbf{px}^{-1}\mathbf{p}$. From (3) we can see that two things will happen, all excluded forecasts will be given zero weight, and the remaining weights will be re-weighted (increased) to ensure that their sum is equal to 1.

In general, for pooling to be beneficial over combining all forecasts $var(y_p) \leq var(y_c)$, that is $\mathbf{w_p}'\mathbf{\Sigma}_{yy}\mathbf{w_p} \leq \mathbf{w}'\mathbf{\Sigma}_{yy}\mathbf{w}$. Bringing all terms to the left side and expanding it to its bilinear form we get:

$$\sum_{p_i \neq 0}\sum_{p_j \neq 0} \sigma_{i,j}(w_{pi}w_{pj} - w_i w_j) - \sum_{p_i = 0}\sum_{p_j = 0} \sigma_{i,j} w_i w_j \leq 0, \tag{4}$$

where $\sigma_{i,j}$ and $p_{i,j}$ are elements of $\mathbf{\Sigma}_{yy}$ and $\mathbf{p}$ with $i, j = (1, \ldots, k)$. Remember that when $p_{i,j} \neq 0$, then $w_{pi,j} > w_{i,j}$. Therefore, (4) tells us that

any contributions to the variance of the combined forecast by the increased pooled weights, must be smaller to the contributions by the excluded forecasts, for pooling to be beneficial. Naturally, this is easy to observe when the pool of forecasts contains very outlying and poor fitting forecasts, rather than when all forecasts are well designed. The first case can happen easily with many existing forecasting support systems in companies, such as SAP APO that is widely used in practice, which offers a fixed repertoire of forecasting methods that often do not match at all the data on hand. Furthermore, when the number of forecasts considered is small, it is relatively easy to seek the optimal set of forecasts to contain in the pool. However, when k increases, the computational cost can increase substantially as well (as $\mathbf{w_p}$ will change with the pool, thus complicating the search), and therefore heuristics such as the one proposed above can be helpful.

4. Empirical evaluation

In this section we outline the dataset and the experimental setup that we use to empirically evaluate pooling against conventional forecast selection and combination, as well as present the results.

4.1. Data

We use four datasets to evaluate the benefits of pooling. The first set of time series is comprised by monthly M3 competition time series (Makridakis and Hibon, 2000) that have at least 120 observations. This dataset includes

1,020 time series with sample size ranging from 120 to 144 periods. The time series capture various items that are broadly classified in microeconomic, macroeconomic, industry, finance, demographic and other series. The second set of series contains monthly time series from the Federal Reserve Economic Data (FRED, Federal Reserve Bank of St. Louis, 2016), filtered with the tags: *inventories*; *nsa* (non-seasonally adjusted); and *monthly* that have only positive values and sample size of 120 observations or more. The dataset has 323 time series, ranging from 120 to 588 observations. The series in this dataset describe size of inventories in various sectors and goods.

The third dataset originates from a UK fast moving consumer good manufacturer, and contains 229 weekly times of 173 observations each. The series record sales at stock keeping unit level of household and personal care products. The fourth dataset originates from a US supermarket chain and describes sales of various food related products. Sales for 854 items are recorded for 91 days.

The behaviour of the series in all datasets is quite diverse, offering a wide variety of time series patterns, including disaggregate produce level sales, or aggregate figures, at different sampling frequencies. Time series have adequate sample to construct both validation and test sets, when needed. The validation set is necessary for the calculation of the cross-validated error, otherwise it is not used. We also restrict time series to be positive so that we can use models with both additive and multiplicative components, giving a wider pool of potential forecasts.

4.2. Experimental setup

Here we provide details of the evaluation scheme and metrics, the forecasting models, and the selection and combination operators used in the empirical evaluation.

4.2.1. Evaluation scheme

For the empirical evaluation we use a rolling origin evaluation scheme (Ord et al., 2017). From each forecast origin we produce the required forecasts, and expand the in-sample set by one observation and repeat the process. This provides us with multiple forecast error measurements, reinforcing the validity of our results, as the effect of potential outliers is mitigated. At each origin the forecasting models are re-optimised, following the recommendations by Fildes and Ord (2002). Given a validation or test set of size m and forecast horizon h, for each set $q = m - h + 1$ forecasts are constructed. For this experiment both validation and test sets are 36 periods and we forecast up to 18 periods ahead, providing 19 forecast traces in each set. The choice of forecast horizon is based on the M3 competition that most time series originate from (Makridakis and Hibon, 2000) and is retained for all datasets for convenience in presenting the results.

To measure the performance of the competing approaches we use the Average Relative Mean Absolute Error (AvgRelMAE) by Davydenko and

Fildes (2013):

$$\text{AvgRelMAE}_i = \sqrt[n]{\prod_{r=1}^{n}\left(\frac{\text{MAE}_{i,r}}{\text{MAE}_{b,r}}\right)},$$
$$\text{MAE} = (qh)^{-1}\sum_{j=1}^{q}\sum_{t=1}^{h}|y_{t+j-1} - \hat{y}_{t+j-1}|,$$

where n is the number of time series that we summarise accuracy over, MAE_i is the Mean Absolute Error of forecast i, over m origins, and y_t and $\hat{y}_t$ are the actuals and forecasts respectively. MAE_b is the error of the benchmark forecast.

As AICc and cross-validated MSE use quadratic loss, we also construct the Average Relative Root Mean Squared Error (AvgRelRMSE), following the same logic as AvgRelMAE:

$$\text{AvgRelRMSE}_i = \sqrt[n]{\prod_{r=1}^{n}\left(\frac{\text{RMSE}_{i,r}}{\text{RMSE}_{b,r}}\right)},$$
$$\text{RMSE} = \sqrt{(qh)^{-1}\sum_{j=1}^{q}\sum_{t=1}^{h}(y_{t+j-1} - \hat{y}_{t+j-1})^2}.$$

Both metrics are very simple to interpret. When their value is below 1 then the forecast is better than the benchmark and vice versa. Davydenko and Fildes (2013) discuss the advantages of AvgRelMAE over several over common accuracy metrics, such as the Mean Absolute Percentage Error or the Mean Absolute Scaled Error that are biased. Therefore, our choices

are due to their desirable statistical properties and ease of interpretation. However, note that as the ratios are formed on summary error statistics, reported differences by AvgRelMAE and AvgRelRMSE tend to be appear smaller than with other metrics. To test whether the reported differences in accuracy are statistically significant we rely on the use of the nonparametric Friedman and post-hoc Nemenyi tests, following the suggestions by Koning et al. (2005) and Demšar (2006), to avoid performing multiple comparisons.

4.2.2. Forecasting models

To produce forecasts we use the complete family of exponential smoothing models, as formulated in the state space framework (Hyndman et al., 2008a). Exponential smoothing is one of the most widely used forecasting models in business, with multiple papers attesting to its good performance and robustness (Holt, 2004). Furthermore, exponential smoothing is implemented in most commercial forecasting systems, making it very relevant for practice (Gardner, 2006).

Different model forms allow capturing varying trend (no trend, additive or multiplicative, which may be damped or not) and seasonality (none, additive or multiplicative). Additionally the error term may be additive or multiplicative, giving in total 30 alternative models. We follow the notation introduced by Hyndman et al. (2008a), as introduced in section 3.2. For example, the well known single exponential smoothing is denoted as ETS(A,N,N), the damped trend model as ETS(A,Ad,N) and so on.

The smoothing parameters, as well as any initial values, are optimised by maximum likelihood. For each model we can calculate various information criteria, which can be used for model selection and combination. We restrict the generation of our forecasts within a single model family so as to be able to use both information criteria and simpler performance metrics in our analysis.

Given how established the model is in both research and practice, for brevity we will not provide any further details here, but instead refer the reader to Hyndman et al. (2008a) or Ord et al. (2017). All forecasts are produced using the *smooth* package for the R language (Svetunkov, 2018).

4.2.3. Selection and combination operators

We use three alternatives for selecting a forecast, or analogously calculating combination weights: (i) uninformative (EW); (ii) $AICc$; and (iii) cross-validated mean squared error (CV).

The *uninformative* relates to the equal weight combination, where we do not have information that any forecast is preferable to others. As a selection operator this translates to choosing a forecast at random. Although there is evidence that equal weights combination is effective, naturally there is no expectation that this is successful as a forecast selection scheme. Given k forecasts, each forecast is given $1/k$ combination weight, or probability to be chosen as best.

Using $AICc$ for model selection is well established and has been shown to

be effective for the exponential smoothing family of models (Hyndman et al., 2002), but also for creating combined forecasts using AICc derived weights (Kolassa, 2011). Nonetheless, it is important to stress that AICc, as other information criteria, have strict requirements as discussed in section 2.1. This restricts its use to forecasting model families, or even narrower, and therefore cannot be used to select or combine forecasts from disparate model families or ad-hoc methods. As the calculation of AvgRelMAE requires a benchmark, we use as such the performance of AICc selected forecast.

A more general approach that avoids such restrictions is based on cross-validated mean squared errors. Although there are many alternatives for calculating cross validation statistics (Barrow and Crone, 2016), not all of them are generally applicable to time series modelling, as they break the continuity of the series. Here we use a validation set, over which rolling origin forecasts are produced and assessed. The assessment metric matches the forecast horizon, as recommended by Fildes and Petropoulos (2015). We use mean square errors, but this is not necessary and different metrics can be used. Note that there is no need to use scale-independent metrics that typically introduce various biases and calculation problems. To select a forecast we pick the one that has the minimum cross-validated error.

For this analysis we combine forecasts linearly, as prescribed by equation (1). Six alternatives are considered: (i) top performing forecast, which is equivalent to model selection; (ii) use all forecasts, which is the conventional forecast combination; (iii)–(iv) form forecast pools using 25% and 50%

quartiles of top performing forecasts respectively; and (v)–(vi) form forecast pools using forecast islands based on AICc and cross-validated mean squared errors. These are combined with the three selection operators above to give in total 18 alternatives. Note that we do not report the results for random model selection, corresponding to uninformative selection due to its poor performance. Furthermore, in this case, any forecast pools also include randomly selected forecasts and so are not reported.

Finally, we use one additional benchmark, the Best Subset Averaging Procedure, which was introduced by Elliott et al. (2013) and discussed for the univariate forecast case by Elliott (2011), hereafter named *Subset*. This approach aims to merge the advantages of the empirically successful unweighted averaging and the theoretically elegant approach of optimal combination weights. First, given k forecasts, we construct all possible subsets of 2 up to k forecasts. Then, we calculate the unweighted average of the forecasts in each subset and finally select the combined forecast that exhibits the lowest error. For example, if $k = 3$, then we construct one subset containing all 3 forecasts and three subsets containing all possible pairs of the forecasts. We construct from the subsets the four average forecasts and pick the best. This process is easy to implement for any forecast, irrespective of its source and for a small number of forecasts it is very fast. However, when the number of forecasts increases, then the number of combinations can become unwieldy. Elliott et al. (2013) consider this problem and find that randomly sampling subsets is a fast solution that does not compromise the performance of the

method. In our case, where $k = 30$, when the number of combinations exceeds 5×10^4 we randomly sample that many subsets, otherwise we consider them all. Obviously, this benchmark is an alternative pooling approach.

Ultimately, all forecasts are benchmarked against model selection and model combination (according to three different criteria) and Subset forecasts, that is a well performing existing forecasting pooling method (for an evaluation the reader is referred to Elliott, 2011).

4.3. Results

Table 1 provides the AvgRelMAE and AvgRelRMSE for the various datasets, while Table 2 presents the overall results across all datasets. The most accurate result in each column is highlighted in boldface (excluding EW that contains only benchmarks). Results that are more accurate than all benchmarks (forecast selection, *Select*, combination of all forecasts, *All*, and *Subset*) in a column, are highlighted in italicised letters.

The two error metrics provide similar insights. Considering the equal weights (EW) results, only benchmark results are provided as any pooling method would pick forecasts at random. The Subset typically improves on the unweighted combination of all forecasts, in agreement with the literature (Elliott, 2011), with the only exception being the Manufacturer dataset. In general the equal weights combinations do not perform very well and this is to be expected, given the diversity of forecasts produced by the 30 forms of ETS.

Table 1: AvgRelMAE and AvgRelRMSE results for the four datasets

Pool	AvgRelMAE			AvgRelRMSE		
	EW	AICc	CV	EW	AICc	CV
M3 monthly (1020 series)						
Select	-	1.000	0.991	-	1.000	0.992
All	1.033	**0.981**	**0.955**	1.036	**0.982**	**0.958**
Subset	0.991	0.991	0.991	0.993	0.993	0.993
Quartile 25%	-	**0.981**	0.966	-	0.983	0.969
Quartile 50%	-	**0.981**	0.958	-	**0.982**	0.961
Islands (AICc)	-	**0.981**	0.967	-	0.983	0.969
Islands (CV)	-	0.982	0.967	-	0.983	0.968
Inventory (323 series)						
Select	-	1.000	1.016	-	1.000	1.017
All	1.077	**0.992**	1.020	1.078	**0.993**	1.022
Subset	1.033	1.033	1.033	1.036	1.036	1.036
Quartile 25%	-	**0.992**	*0.991*	-	**0.993**	*0.993*
Quartile 50%	-	**0.992**	***0.988***	-	**0.993**	***0.990***
Islands (AICc)	-	**0.992**	*0.995*	-	**0.993**	*0.996*
Islands (CV)	-	0.993	*0.994*	-	**0.993**	*0.995*
Manufacturer (229 series)						
Select	-	1.000	1.009	-	1.000	1.008
All	1.019	**0.991**	1.002	1.021	**0.993**	1.004
Subset	1.041	1.041	1.041	1.041	1.041	1.041
Quartile 25%	-	0.992	1.002	-	**0.993**	1.004
Quartile 50%	-	**0.991**	*1.000*	-	**0.993**	*1.002*
Islands (AICc)	-	0.993	*0.999*	-	0.994	***1.000***
Islands (CV)	-	0.992	***0.998***	-	0.994	***1.000***
Supermarket (854 series)						
Select	-	1.000	1.019	-	1.000	1.012
All	1.041	0.991	1.021	1.018	0.990	1.004
Subset	1.037	1.037	1.037	1.032	1.032	1.032
Quartile 25%	-	***0.989***	***0.994***	-	***0.988***	***0.985***
Quartile 50%	-	***0.989***	*1.007*	-	***0.988***	***0.995***
Islands (AICc)	-	*0.990*	*1.008*	-	***0.989***	*0.996*
Islands (CV)	-	***0.989***	*1.005*	-	***0.988***	*0.996*

Table 2: Overall AvgRelMAE and AvgRelRMSE across all time series

Pool	AvgRelMAE			AvgRelRMSE		
	EW	AICc	CV	EW	AICc	CV
Select	-	1.000	1.006	-	1.000	1.004
All	1.040	0.987	0.991	1.034	**0.987**	0.987
Subset	1.017	1.017	1.017	1.017	1.017	1.017
Quartile 25%	-	0.987	***0.983***	-	**0.987**	***0.981***
Quartile 50%	-	***0.986***	***0.983***	-	**0.987**	***0.981***
Islands (AICc)	-	0.987	*0.988*	-	**0.987**	*0.985*
Islands (CV)	-	0.987	*0.987*	-	**0.987**	*0.985*

Looking at the AICc column we can observe that any combination is better than selecting a single forecast, but the differences between alternative pools are negligible. Again, given the calculation of AIC weights this is not unexpected, as they effectively reduce the contribution of models that do not fit the series well to the series. Forecasts that are regarded as improbable to match the data generating process are given almost zero weights, which is similar to the effect of the various pooling approaches.

On the other hand, there are more promising gains when CV is used. We can observe that the CV column typically outperforms all other selection/weighting metrics (exception is the Manufacturing dataset). We can also observe that all pooling approaches outperform selecting a single forecast, combining all of them, or using Subsets for most datasets. The combination of all forecasts performs particularly well on the M3 monthly dataset, yet the pools perform relatively close to it, and substantially better than choosing a single model. The differences between the various pooling approaches are

again relatively small.

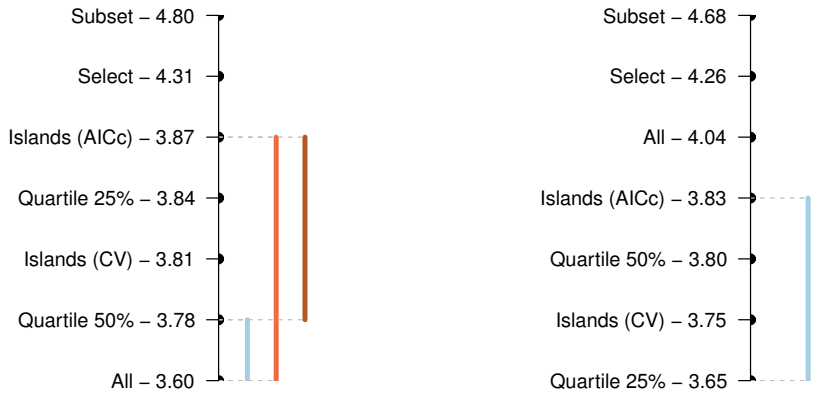
Focusing on the comparison between the two alternative pooling approaches, quartiles and forecast islands, only small differences are observed, in favour of the quartiles. However, the quartiles are based on an arbitrary cut-off, while the forecast islands are not. In this case, we investigated the performance of the two top quartiles, but one could very well attempt to evaluate any number of quantiles. Note that the benchmarks model selection and combination are options of this continuum. Forecast islands avoid this search, by identifying a reasonably performing cut-off point and not requiring the modeller to identify one. We explore this further in section 5.

Comparing the two alternative forecast islands specifications, on AICc or CV, we observe small differences in favour of the latter. This reflects the better performance of CV overall. However, it is interesting to observe that the forecast islands in Tables 1 and 2 mix various selection criteria. For example, using both Islands (CV) and CV based weighting is reasonable, however the use of Islands (CV) with AICc derived combination weights is more questionable, as the various forecasts are ranked differently on different criteria. Although we do not advocate mixing the criteria, it is important to note that this case is quite common in practice. For example, consider selecting the best ARIMA model from a pool of ARIMA candidate models. It is not guaranteed that this pool will contain the 'best' ARIMA, and typically the pool will be formed based on some arbitrary modelling decisions. Essentially, the construction of the forecast pool and the selection or combination

of forecasts is typically done on different criteria, and specifically we often do not state explicitly the criterion used for the formation of the forecast pool.

We use the Friedman test to initially test whether there are significant differences in the performance of the forecasts by AICc and by CV. In both cases there is evidence of this (p-value is 0.000 in both cases) and proceed to apply the post-hoc Nemenyi test. Figure 4 presents the results at 5% significance level. For each forecast the mean rank is provided, according to MAE, with the lowest indicating the most accurate one. When there is not adequate evidence to suggest statistically significant differences between forecasts (i.e., the differences of the mean ranks is lower than the critical distance of 0.184), then these are connected by a vertical line. In agreement with the results in Tables 1 and 2, there are only minimal differences between the Island and Quartile pools. For the case of AICc, the weighted combination performs very well, together with the pooling methods. For the case of CV, the pooling methods significantly outperform all three benchmarks.

Overall, we find very strong evidence in favour of forecast combinations, particularly given a reliable performance metric, such as CV. Furthermore, pooling via quartiles or forecast islands offers additional accuracy gains. The latter avoids the arbitrary modelling decision of selecting the cut-off point, or considering the dichotomy between model selection and combination (that are the extreme quantile options) and therefore using forecast islands is recommended.



(a) AICc (b) CV

Figure 4: Visualisation of the Nemenyi test results at 5% significance level. There is no evidence of statistically significant differences between forecasts connected by vertical lines. The multiple lines provided for AICc indicated different groups, depending on the starting forecast.

5. Discussion

Building on the results presented above, we discuss the ability of the proposed heuristic to identify a well-performing cut-off point. In section 3.2 we argued their connection with the top-quantile pools, which require the modeller to choose a cut-off point for the number of forecasts to consider. The performance of islands already suggests their ability to identify useful cut-off points, as seen in section 4.3. Here we explore this connection further. To do this, we use as an example, a time series of monthly wine sales that is available in the *forecast* package for the R language (Hyndman, 2016). We retain the last 3 years as a test set and use the preceding equal sample as validation set, when needed. We set the forecast horizon to a full year. We follow the experimental setup described in section 4.2, with the following

further changes: instead of using only two top-quartiles pools, we construct 30 top-quantiles pools that start from a single up to all thirty forecasts. We also construct forecast island pools and measure the AvgRelMAE, using the performance of AICc forecast selection as benchmark.

Figure 5 plots the AvgRelMAE for pools constructed using AICc and cross-validated errors. We highlight the best performing top-quantile with a vertical line and the forecast island identified cut-off point with a dashed vertical line. We can observe that in both cases the island based cut-off is close to the best possible top-quantile. Note that the best quantile pool is identified on the test set, after the experiment is conducted and would not be known in advance, while the island cut-off is identified using only past data and therefore can be used to produce forecasts.

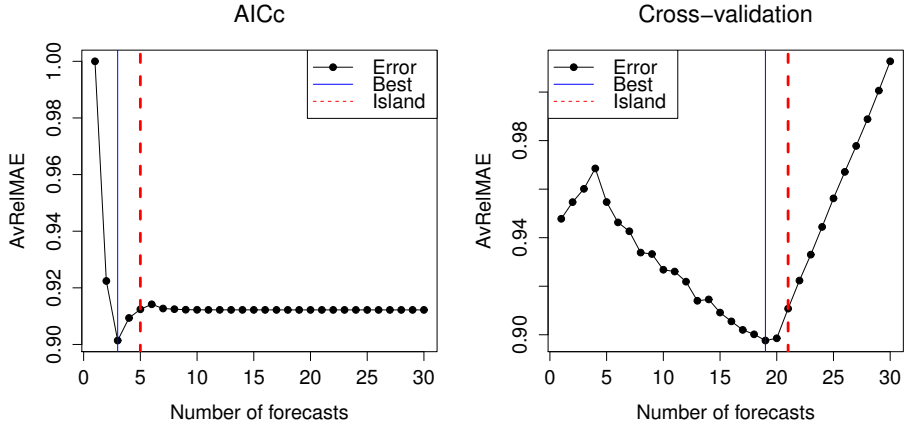


Figure 5: AvgRelMAE performance for top-quantile and forecast island pools constructed using AICc and cross-validated errors. The top performing quantile is marked with a vertical line and the forecast island with a dashed line.

The forecast island pools are close to the best cut-off point that is possible

using top-quantile derived pools and is done without requiring to manually set how many forecasts to include in the pool.

6. Conclusions

Forecast selection and combination have been regarded as two competing alternatives. In the context of forecast pooling these are merely two extreme pools. The way that the individual forecasts are ranked, or weighted, results in the well established alternatives in the literature and practice. Typically we construct arbitrary pools, on which we select or combine forecasts. In this paper we proposed a heuristic to formulate appropriate pools, without having the modeller decide on an arbitrary cut-off point: which forecasts should be included in the combination or not.

Our empirical evaluation over four diverse datasets shows that forecast pooling has overall better forecast accuracy than either selection or combination of all forecasts. This is achieved by eliminating particularly poor forecasts from the combination pool, as well as capitalising on the well established advantages of forecast combination. Moreover, we find that the proposed forecast islands approximate the unknown best-quantile of top performing forecasts that a modeller could have selected only ex-post, for a variety of performance criteria, thus removing that modelling decision and enabling forecast automation further. We argue that this is particularly relevant for practical demand planning situations, as well as wider business forecasting cases, where the number of time series to be forecasting is high,

often with limited expertise and/or supporting tools from the available systems, making reliable automatic forecasting desirable.

We argue that model pooling is part of the forecast building process and should be considered explicitly as such, rather than assuming that the available forecasts are adequate or sensible, which typically are arbitrarily generated and may or may not contain the 'best' forecast. It is important to note that forecast pooling, as discussed in this paper, does not eliminate this aspect fully, but rather allows the modeller to consider a larger number of forecasts that will be streamlined through pooling, before the rest of standard forecasting process takes place.

Forecast pooling and the proposed heuristic are shown to be effective in our empirical evaluation, however, as implemented here, there are several ad-hoc selections and lack a concrete statistical rationale. Forecast islands seem to be able to identify reasonable sets of forecasts, facilitating automation. Although we provide some insight as the to nature of included and excluded forecasts for pooling to be beneficial, this is far from a complete statistical grounding. We argue that this research helps motivate the use of pooling in a supply chain forecasting context, and provides further evidence of good forecast accuracy. Given this promising performance, future research should investigate an appropriate statistical grounding.

References

Agnew, C. E., 1985. Bayesian consensus forecasts of macroeconomic variables. Journal of Forecasting 4 (4), 363–376.

Aiolfi, M., Timmermann, A., 2006. Persistence in forecasting performance and conditional combination strategies. Journal of Econometrics 135 (1), 31–53.

Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19 (6), 716–723.

Arlot, S., Celisse, A., et al., 2010. A survey of cross-validation procedures for model selection. Statistics surveys 4, 40–79.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., Petropoulos, F., 2017. Forecasting with temporal hierarchies. European Journal of Operational Research 262 (1), 60–74.

Aye, G. C., Balcilar, M., Gupta, R., Majumdar, A., 2015. Forecasting aggregate retail sales: The case of South Africa. International Journal of Production Economics 160, 66–79.

Barrow, D. K., Crone, S. F., 2016. Cross-validation aggregation for combining autoregressive neural network forecasts. International Journal of Forecasting 32 (4), 1120–1137.

Barrow, D. K., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: implications for inventory management. International Journal of Production Economics 177, 24–33.

Bates, J. M., Granger, C. W. J., 1969. The combination of forecasts. Operational Research Society 20 (4), 451–468.

Bergmeir, C., Hyndman, R. J., Benítez, J. M., 2016. Bagging exponential smoothing methods using STL decomposition and box–cox transformation. International Journal of Forecasting 32 (2), 303–312.

Billah, B., Hyndman, R. J., Koehler, A. B., 2005. Empirical information criteria for time series forecasting model selection. Journal of Statistical Computation and Simulation 75 (10), 831–840.

Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M., 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Brown, G., Wyatt, J., Harris, R., Yao, X., 2005. Diversity creation methods: a survey and categorisation. Information Fusion 6 (1), 5–20.

Burnham, K. P., Anderson, D., 2002. Model selection and multi-model inference: A practical information-theoric approach. Springer.

Burnham, K. P., Anderson, D. R., 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York.

Chan, F., Pauwels, L. L., 2018. Some theoretical results on forecast combinations. International Journal of Forecasting 34 (1), 64–74.

Chatfield, C., 2000. Time-series forecasting. CRC Press.

Claeskens, G., Magnus, J. R., Vasnev, A. L., Wang, W., 2016. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 32 (3), 754–762.

Clemen, R. T., 1989. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Davydenko, A., Fildes, R., 2013. Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. International Journal of Forecasting 29 (3), 510–522.

de Menezes, L. M., Bunn, D. W., 1998. The persistence of specification problems in the distribution of combined forecast errors. International Journal of Forecasting 14 (3), 415–426.

De Menezes, L. M., Bunn, D. W., Taylor, J. W., 2000. Review of guidelines for the use of combined forecasts. European Journal of Operational Research 120 (1), 190–204.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research 7 (Jan), 1–30.

Diebold, F. X., Pauly, P., 1990. The use of prior information in forecast combination. International Journal of Forecasting 6 (4), 503–508.

Elliott, G., 2011. Averaging and the optimal combination of forecasts. University of California, San Diego.

Elliott, G., Gargano, A., Timmermann, A., 2013. Complete subset regressions. Journal of Econometrics 177 (2), 357–373.

Elliott, G., Timmermann, A., 2016. Economic Forecasting, 1st Edition. Princeton University Press.

Fang, Y., 2003. Forecasting combination and encompassing tests. International Journal of Forecasting 19 (1), 87–94.

Fang, Y., 2011. Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. Journal of data science 9 (1), 15–21.

Fildes, R., 2001. Beyond forecasting competitions. International Journal of Forecasting 17, 556–560.

Fildes, R., Ord, K., 2002. Forecasting competitions–their role in improving forecasting practice and research. A companion to economic forecasting, 322–253.

Fildes, R., Petropoulos, F., 2015. Simple versus complex selection rules for forecasting many time series. Journal of Business Research 68 (8), 1692–1701.

FRED, Federal Reserve Bank of St. Louis, 2016. Inventories monthly non-seasonaly adjusted time series.
URL `https://fred.stlouisfed.org`

Gardner, E. S., 2006. Exponential smoothing: The state of the art - part II. International Journal of Forecasting 22 (4), 637–666.

Genre, V., Kenny, G., Meyler, A., Timmermann, A., 2013. Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting 29 (1), 108–121.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164 (1), 130–141.

Granger, C. W., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 3 (2), 197–204.

Harvey, D. I., Newbold, P., 2005. Forecast encompassing and parameter estimation. Oxford Bulletin of Economics and Statistics 67 (s1), 815–835.

Holt, C. C., 2004. Author's retrospective on "forecasting seasonals and trends by exponentially weighted moving averages". International Journal of Forecasting 20 (1), 11–13.

Hyndman, R., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008a. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media.

Hyndman, R. J., 2016. forecast: Forecasting functions for time series and linear models. R package version 7.1.
URL `http://github.com/robjhyndman/forecast`

Hyndman, R. J., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008b. Forecasting with Exponential Smoothing: The State Space Approach. Springer Verlag, Berlin.

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.

Kolassa, S., 2011. Combining exponential smoothing forecasts using Akaike weights. International Journal of Forecasting 27 (2), 238–251.

Koning, A. J., Franses, P. H., Hibon, M., Stekler, H. O., 2005. The m3 competition: Statistical tests of the results. International Journal of Forecasting 21 (3), 397–409.

Kourentzes, N., Barrow, D. K., Crone, S. F., 2014a. Neural network ensemble operators for time series forecasting. Expert Systems with Applications 41 (9), 4235–4244.

Kourentzes, N., Petropoulos, F., 2015. Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics.

Kourentzes, N., Petropoulos, F., Trapero, J. R., 2014b. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30 (2), 291–302.

Lemke, C., Gabrys, B., 2010. Meta-learning for time series forecasting and forecast combination. Neurocomputing 73 (10), 2006–2016.

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

Makridakis, S., Wheelwright, S. C., Hyndman, R. J., 1998. Forecasting: Methods and Applications, 3rd Edition. John Wiley & Sons, New York.

Matsypura, D., Thompson, R., Vasnev, A. L., 2017. Optimal selection of expert forecasts with integer programming. Omega.

McNees, S. K., 1992. The uses and abuses of 'consensus' forecasts. Journal of Forecasting 11 (8), 703–710.

Miller, C. M., Clemen, R. T., Winkler, R. L., 1992. The effect of nonstationarity on combined forecasts. International Journal of Forecasting 7 (4), 515–529.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2015. Introduction to linear regression analysis. John Wiley & Sons.

Newbold, P., Granger, C. W., 1974. Experience with forecasting univariate time series and the combination of forecasts. Journal of the Royal Statistical Society. Series A (General), 131–165.

Newbold, P., Harvey, D. I., 2002. Forecast combination and encompassing. A companion to economic forecasting, 268–283.

Ord, J. K., Fildes, R., Kourentzes, N., 2017. Principles of Business Forecasting, 2nd Edition. Wessex Press Publishing Co.

Ord, K., 2001. Commentaries on the M3-competition. an introduction, some comments and a scorecard. International Journal of Forecasting 17, 537–584.

Petropoulos, F., Kourentzes, N., 2015. Forecast combinations for intermittent demand. Journal of the Operational Research Society 66 (6), 914–924.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica, 221–242.

Smith, J., Wallis, K. F., 2009. A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics 71 (3), 331–355.

Stock, J. H., Watson, M. W., 2004. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23 (6), 405–430.

Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society. Series B (Methodological), 44–47.

Strijbosch, L. W., Syntetos, A. A., Boylan, J. E., Janssen, E., 2011. On the interaction between forecasting and stock control: The case of non-stationary demand. International Journal of Production Economics 133, 470–480.

Svetunkov, I., 2018. smooth: Forecasting Using Smoothing Functions. R package version 2.4.0.
URL `https://CRAN.R-project.org/package=smooth`

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 16 (4), 437–450.

Tian, J., Anderson, H. M., 2014. Forecast combinations under structural break uncertainty. International Journal of Forecasting 30 (1), 161–175.

Timmermann, A., 2006. Forecast combinations. In: G. Elliott, C. G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Vol. 1. Elsevier, pp. 135–196.

Xia, Y., Tong, H., 2011. Feature matching in time series modeling. Statistical Science, 21–46.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. Biometrika 92 (4), 937–950.