Model combinations through revised base-rates

Fotios Petropoulos
School of Management,
University of Bath UK
and
Evangelos Spiliotis
Forecasting and Strategy Unit,
School of Electrical and Computer Engineering,
National Technical University of Athens, Greece
and
Anastasios Panagiotelis
Discipline of Business Analytics,
University of Sydney, Australia

July 26, 2022

Abstract

Standard selection criteria for forecasting models focus on information that is calculated for each series independently, disregarding the general tendencies and performances of the candidate models. In this paper, we propose a new way to perform statistical model selection and model combination that incorporates the base-rates of the candidate forecasting models, which are then revised so that the per-series information is taken into account. We examine two schemes that are based on the precision and sensitivity information from the contingency table of the base-rates. We apply our approach on pools of either exponential smoothing or ARMA models, considering both simulated and real time series, and show that our schemes work better than standard statistical benchmarks. We test the significance and sensitivity of our results, discuss the connection of our approach to other cross-learning approaches, and offer insights regarding implications for theory and practice.

Keywords: forecasting, model selection/combination, information criteria, exponential smoothing, cross-learning.

1 Introduction

Model selection and combination (or averaging) have long been fundamental ideas in forecasting for business and economics (see Inoue and Kilian, 2006; Timmermann, 2006, and references therein for model selection and combination respectively). In both research and practice, selection and/or the combination weight of a forecasting model are usually *case-specific*. By this, we mean that they are based on criteria such as the Akaike's information criterion (Kolassa, 2011), the predictive log score (Geweke and Amisano, 2011; Pettenuzzo and Timmermann, 2017), or standard accuracy measures (Koutsandreas et al., 2021) that summarise forecasting performance through time series validation processes (Bergmeir and Benítez, 2012) computed only on the *series of interest itself*. Since these criteria are typically justified by assumptions (asymptotic or otherwise) that may not hold in practice, a selected model may not perform best out of sample. As a result, forecasts can potentially be improved by exploiting environmental information, in particular the propensity across multiple time series, for the model chosen by a selection criterion to differ from the model with the best out-of-sample forecasting performance.

In this paper we propose easy to implement and general algorithms for model selection and combination in forecasting that exploit revised base-rate information by using a collection of reference series. Examples of such reference series could include the large collections of macroeconomic time series (Stock and Watson, 2012) or the time series from the M forecasting competitions (Makridakis et al., 2020). For each of these examples, all reference series are themselves quantities of interest that may be the target of forecasting and we can evaluate forecasting performance by averaging across all series. This motivates the approach we take in our paper. However, in a detailed empirical study, we also show that forecast performance is robust to including additional reference series (e.g., by including series from a different domain to the forecast of in-

terest), being also robust to the number of series considered for extracting the base-rate information.

Rather than use reference series as predictors, they are instead used to revise the probabilities that a model is the "correct" (or "true") model in the sense of having the best out-of-sample forecasting performance. Apart from the reference series, the only other requirements are the choice of a pool of candidate models, a criterion for selecting between these models, and a criterion for evaluating forecasts. As a result, there is scope to tailor our proposed algorithms to applications with specific loss functions. Furthermore, as long as the selection and evaluation criteria are likelihood-free, the set of candidate models can even include models for which the likelihood is intractable or difficult to compute.

To provide the general idea behind our proposed approach, let there be two models under consideration (model A and model B). Let events S_A and S_B refer to models A or B respectively being selected according to some criterion. Similarly, let events C_A and C_B refer to models A or B being the "correct" model, in the sense of being optimal with respect to some evaluation criterion. Also, assume that we have access to a set of reference series, such that we can empirically estimate joint probabilities of models being selected and "correct", and thus populate the cells of the contingency table (such as Table 1). The use of reference series is inspired by the meta-learning literature in forecasting (see Lemke and Gabrys, 2010; Wang et al., 2009; Talagala et al., 2018; Montero-Manso et al., 2020, and references therein). However, in contrast to these papers, the weights we compute have an interpretation as probabilities rather than being the outputs of a "black-box"[1], machine learning algorithm which type (e.g., neural network or decision tree) and hyper-parameter values have to be carefully selected. Moreover, since the weights in our approach are solely estimated using forecasting performance related informa-

[1]Even where weights are determined by a black-box, they can still be interpreted as contributions to aggregate forecast error variance(Wolpert, 1992; Conflitti et al., 2015; Diebold and Shin, 2019).

tion, they are not subject to time series features and general statistics which number, type, and representativeness may be challenging to determine in practice for constructing a successful meta-learning algorithm. In addition, the information exploited by our approach focuses on models instead of series, being also summarised at a global level (forecasting performance is being tracked across the complete set of reference series) instead of being learnt at a local one (forecasting performance is being tracked at each series and the connections between the inputs and outputs of the algorithm are determined accordingly).

Table 1: The contingency table.

	C_A	C_B	Total
S_A	$p(S_A \cap C_A)$	$p(S_A \cap C_B)$	$p(S_A)$
S_B	$p(S_B \cap C_A)$	$p(S_B \cap C_B)$	$p(S_B)$
Total	$p(C_A)$	$p(C_B)$	1

From Table 1, we can observe the general, environmental tendencies for models A and B in terms of (*i*) being the selected model with probabilities $p(S_A)$ and $p(S_B)$, respectively, and (*ii*) being the "correct" model with probabilities $p(C_A)$ and $p(C_B)$, respectively. In the literature, when selection is based on $p(C_A)$ and $p(C_B)$, which correspond to the base-rate information, then this is typically referred to as "aggregate selection" because a single model is used to forecast all series by considering which model provided the most accurate forecasts for a hold-out sample in most of the cases and ignoring their particular characteristics (Fildes and Petropoulos, 2015).

For any new series, we can first evaluate the case-specific event of selecting either model A or model B, i.e., we either observe the event S_A or S_B. Suppose we observe S_A. Rather than use model A, we propose to incorporate base-rate information by selecting model A when $p(C_A|S_A) > p(C_B|S_A)$ and model B otherwise. These conditional probabilities are computed using the values in the contingency table and summarise "precision", i.e., the proportion of cases for which the selected model is actually the "correct" model.

An alternative approach will be to select model A when $p(S_A|C_A) > p(S_A|C_B)$ and model B otherwise. These conditional probabilities offer the revised probabilities for a model being selected, assuming that some other model (or the same model) is the correct one. In contrast to the previous case, these conditional probabilities summarise "sensitivity". We note that $p(S_A|C_A) = (C_A|S_A)$ and $p(S_A|C_B) = (C_B|S_A)$ only when $p(C_A) = p(C_B)$.

If, instead of model selection, model combination is desired, the conditional probabilities discussed above can be used as weights in a forecast combination. Since $p(C|S)$ represents the probability that a model is the "correct" model conditional on observed information, this bears an interesting resemblance to the Bayesian paradigm for model combination. In the Bayesian setting, the choice of model is treated in the same way as other parameters and model combination based on posterior model probabilities arises in a natural way to integrate out model uncertainty. The posterior probability that a given model is the "correct" model can be computed using Bayes theorem, although in practice Markov chain Monte Carlo algorithms are required for exploring the model space. For an extensive review of Bayesian model combination, including key historical references, see Hoeting et al. (1999).

Although the computation of posterior model probabilities can be challenging, there are a number of useful approximations. A prominent example, discussed by Raftery (1996), is based on the Bayesian Information Criterion (BIC), which we use as one of the selection criteria in Section 5.2. Alternatives to the BIC can also be considered. For example, in the forecasting literature Kolassa (2011) uses Akaike's Information Criterion (AIC) in a similar fashion to find forecast combination weights. Furthermore, recent work by Bissiri et al. (2016), Loaiza-Maya et al. (2020) as well as the literature on PAC-Bayes (see Guedj, 2019, for a review) generalise posterior inference to allow loss functions to replace likelihoods and can also be applied to finding model weights. Our proposed approach also uses loss functions in the form of selection and evaluation cri-

teria. It does however differ from existing approaches substantially, by using the conditional probabilities computed from reference series as "proxies" for the posterior model probabilities. By comparing our own proposed methods to forecast combinations that approximate posterior model probabilities without using the reference time series, we can examine the benefits of the proposed cross-learning framework. Furthermore, the use of general loss functions, rather than likelihoods, allow our method to be extended to machine learning methods, such as random forests, which are becoming increasingly popular in business and macroeconomic forecasting (Medeiros et al., 2021).

The remainder of the paper is structured as follows. Section 2 introduces our proposed approach more rigorously including different approaches for computing the weights from the reference time series. Section 3 presents a simulation study that illustrates the intuition behind how the proposed approaches work and provide evidence of their effectiveness. This is followed by an application on real macroeconomic data in Section 4. Section 5 describes the empirical design used to evaluate our proposed approach in more detail. By considering a data set containing more series, both in number and diversity, we are able to investigate the sensitivity of results to different choices of the reference set as well as an additional selection criteria, namely time series validation. Section 6 and Section 7 provide additional discussion and conclude respectively.

2 Methodology

The proposed forecasting algorithm involves two steps. First, a collection of reference time series is used to compute the probabilities in a contingency table. Second, models are fit to the actual time series of interest and combination weights are derived for each model according to one of three schemes, two of which depend on the contingency table. We now describe each of these steps in turn.

2.1 Populating contingency table

Let $\mathcal{Z} := \left\{z^{(1)}, \ldots, z^{(N)}\right\}$ be a collection of N *reference* time series. Let $\mathcal{M} := \{M_1, \ldots, M_K\}$ be a set of K *models* that can be used for forecasting. Let S be a *selection criterion*, computed using only in-sample information and with value $S_k^{(n)}$ for reference time series $z^{(n)}$ and model M_k. Similarly, let C be an *evaluation criterion*, used to determine the "correct" model, that is computed using only out-of-sample information and has value $C_k^{(n)}$ for reference time series $z^{(n)}$ and model M_k. Without loss of generality, we will assume that lower values of $S_k^{(n)}$ and $C_k^{(n)}$ indicate better performing models. Although, the selection and evaluation criteria will be of a statistical nature in this paper, in certain applications, model combinations could be based on context specific loss functions (for an example from finance see Caldeira et al., 2016, who consider model combinations based on Sharpe ratios).

Let $\boldsymbol{W}$ be a $K \times K$ matrix corresponding to the contingency table, with element $w_{i,j}$ in the i^{th} row and j^{th} column. These entries measure the joint probability that for a randomly selected reference time series, the "correct" model is model j when the selected model is model i. Algorithm 1 provides details on how $w_{i,j}$ are computed. We note that in cases where the computational burden is small, step 7 of Algorithm 1 can involve a rolling window evaluation.

2.2 Forecasting algorithm

Now let $\boldsymbol{y}$ be the time series of interest that needs to be forecast. In this paper, we are interested in forecasting all of the reference series, that is $\boldsymbol{y}$ will be each element of $\mathcal{Z}$. However, in general, $\boldsymbol{y}$ may only be a single element of $\mathcal{Z}$ or may not be in $\mathcal{Z}$ at all. Forecasts from all models in $\mathcal{M}$ will be produced. Three schemes are considered for model selection and combination. *Criterion*-based schemes ignore the information in the contingency table entirely. *Precision*-based schemes derive weights based on the probabil-

Algorithm 1 Algorithm to populate cells of contingency table

procedure ContTab($\mathcal{Z}$,$\mathcal{M}$,S,C)
  Set $w_{i,j} \leftarrow 0$ for all $i = 1,\ldots,K$, $j = 1,\ldots,K$ ▷ Initialise
  for $n = 1,\ldots,N$ **do** ▷ Loop over reference time series
    Split $z^{(n)}$ into a training sample $z^{(n)}_{train}$ and a test sample $z^{(n)}_{eval}$.
    for $k = 1,\ldots,K$ **do** ▷ Loop over models
      Fit model M_k to $z^{(n)}_{train}$ and compute $S^{(n)}_k$.
      Compute $C^{(n)}_k$ using $z^{(n)}_{eval}$
    end for
    Set $i^* = \underset{i}{argmin} S^{(n)}_i$ ▷ if larger values of $S^{(n)}_i$ indicate better models, use *argmax* instead
    Set $j^* = \underset{j}{argmin} C^{(n)}_j$ ▷ if larger values of $C^{(n)}_j$ indicate better models, use *argmax* instead
    Set $w_{i^*,j^*} \leftarrow w_{i^*,j^*} + 1$
  end for
  Set $\boldsymbol{W} \leftarrow \boldsymbol{W} \big/ N$ ▷ Normalise cells of contingency table
end procedure

ities that a model is the "correct" model conditional on it being selected ($p(C|S)$). *Sensitivity*-based schemes derive weights based on probabilities of selecting a model given the "correct" model ($p(S|C)$). These are equivalent to $p(C|S)$ if the assumed distribution $p(C)$ is uniform (a priori the new time series is equally likely to be best forecast by any model). The computation of weights is outlined in detail in Algorithm 2.

The final forecasts are either based on selection or combination and one of the three weighting schemes. In the results of Section 3, 4 and 5, *criterion-select*, *precision-select* and *sensitivity-select* respectively refer to using the model corresponding to the element that is maximal for w^{crit}, w^{prec} and w^{sens}. Alternatively, *criterion-average*, *precision-average* and *sensitivity-average* take a weighted average of forecasts using w^{crit}, w^{prec} and w^{sens} respectively as weights.

Algorithm 2 Algorithm to compute criterion, precision and sensitivity combination weights

procedure CompW($\mathbf{W}$,$\mathcal{M}$,S,$\boldsymbol{y}$)
  for k=1,...K **do** ▷ Loop over models
    Fit model M_k to $\boldsymbol{y}$ and compute $S_k^{\boldsymbol{y}}$ and a forecast $\hat{y}_{T+h,k}$.
  end for
  Set unnormalised criterion weights to $w_k^{\text{crit}} \leftarrow exp(-S_k^{\boldsymbol{y}}/2)$
  Set $i^* = \underset{i}{argmin} S_i^{\boldsymbol{y}}$ ▷ if larger values of $S_i^{\boldsymbol{y}}$ indicate better models then use $argmax$ instead
  Compute unnormalised precision weights $w_k^{\text{prec}} \leftarrow w_{i^*,k}$
  Compute unnormalised sensitivity weights $w_k^{\text{sens}} \leftarrow w_{i^*,k} \Big/ \sum_i w_{i,k}$
  Normalise all weights, $\boldsymbol{w}^{\text{g}} \leftarrow \boldsymbol{w}^{\text{g}} \Big/ \sum_k w_k^{\text{g}}$, for g $\in$ {crit, prec, sens} and where bold $\boldsymbol{w}$ denotes vectors of length K.
end procedure

3 Simulation Study

3.1 Simulation design

We now turn our attention to a simulation study that provides intuition into how our proposed methods work as well as evidence of their effectiveness. We simulate $n = 200$ times series from the ARMA class of models

$$(1 - \phi(L))(1 - \Phi(L^m))y_t = (1 + \theta(L))(1 + \Theta(L^m))\epsilon_t,$$

where y_t is the time series at time t, ϵ_t is a white noise term at time t, L is the lag operator, m is the seasonal frequency (e.g., 4 and 12 for quarterly and monthly data, respectively), $\phi(L)$ is an AR polynomial of order p, $\Phi(L)$ is a seasonal AR polynomial of order P, $\theta(L)$ is an MA polynomial of order q and $\Theta(L)$ is a seasonal MA polynomial of order Q. Of the $n = 200$ series, 80 are simulated from a non-seasonal ARMA(1,1) model (i.e., $p = q = 1$, $P = Q = 0$) and 120 series are simulated from a seasonal AR(1) model (i.e., $P = 1$, $p = q = Q = 0$) with $m = 12$. Different AR and MA parameters are generated

for each series from $U(0.5, 0.75)$ distributions and the variance of the white noise terms of each series are generated from $U(0.8, 1.6)$ distributions.

A data generation process (DGP) similar to the one described above could arise in the context of factor models with variables loading onto seasonal factors, non-seasonal factors, or some combination of both (see Nieto et al., 2016, for an example of such a model). In this setting, although the number of variables is large, each variable is a linear combination of only a small number of factors. Linear combinations of ARMA models are also ARMA models, with the AR and MA orders of the linear combination constrained by the orders of the constituent processes (see Lütkepohl, 1984, for general results). As a result, in a factor model, each series will follow one of only a small number of ARMA processes despite the number of series being large. This enables "borrowing strength" from all time series when trying to select a forecasting model (or combination weights) for each individual series.

3.2 Contingency Table

For the simulation study we consider all seasonal ARMA models with p, P, q, and Q equal to either 0 or 1, leading to a total of 16 candidate models. These are numbered with letters from A to P as outlined in Table 2. Models A-D are all non-seasonal models ($P = Q = 0$) while models E-P all allow for some form of seasonality. Recall that all data are either generated from model D (non-seasonal ARMA(1,1)) or model E (seasonal AR(1)). All models are estimated using the `ARIMA()` function in the *forecast* package in R.

Table 2: Summary of labels used to indicate the 16 ARMA models considered in the simulation study.

Order	Label															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
p	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
P	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
q	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Q	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

To populate the contingency table in 3, for each series we first use a training sample of $T_{train} = 60$ observations to compute the BIC of the 16 candidate models and use it as a criterion for selecting the most appropriate model. The counts of these selections correspond to the rows of the contingency table. The evaluation criterion used to determine the "correct" model (corresponding to the columns of the contingency table) is the Mean Absolute Error (MAE). To compute MAE, we form one step ahead forecasts, and repeat this process over a non-overlapping rolling window for the next $T_{eval} = 90$ observations. Note that the "correct" model according to MAE may differ from the true DGP which is always either model *D* or model *E*.

Table 3: Contingency table for the simulation study. Rows correspond to the "selected" model, while columns correspond to the "correct" model. Model labels *A*-*P* are outlined in Table 2. Row I and Cell (F,O) are discussed in the text therefore are respectively highlighted in italics and bold.

		Correct															
	Model	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	Total
Selected	*A*	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	2
	B	0	4	6	17	0	2	3	3	1	1	1	1	1	1	0	4	45
	C	0	2	1	4	0	0	1	0	0	1	0	0	0	2	0	0	11
	D	0	8	5	25	0	2	0	1	1	3	0	4	0	2	1	2	54
	E	0	0	0	0	34	7	11	5	2	0	1	0	2	0	2	0	64
	F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	1
	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	H	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3
	I	*0*	*0*	*0*	*0*	*7*	*2*	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*1*	0	*11*
	J	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4
	K	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
	L	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	3
	M	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Total	0	17	12	51	43	14	15	10	4	5	3	5	5	5	5	6	200

We can draw a number of conclusions from Table 3 that provide insights into how the proposed cross-learning methods work. Focusing on row *I* (italics in Table 3), there are

11 series for which model I (a seasonal MA(1) model) is selected by BIC. For the majority of these series (7 series), the model identified as "correct" by validation is model D (seasonal AR(1)). For all 11 of these series the true DGP is in fact model D. Here *criterion-select*, which selects purely on the basis of BIC, would lead to an incorrect model (model I) being used to forecast these 11 series. In contrast, our *precision-select* method would use the "correct" model (model D) whenever model I is selected by BIC. In this way our cross-learning approach can correct possible tendencies to select incorrect models by revising them through base-rate information.

As a caveat, we should note there are cases where the proposed approach can break down. For example, there is one series with model F selected for which the "correct" model was misidentified as model O (bold in Table 3). The true DGP for this series is model D. In this case, both the *criterion-select* and *precision-select* approaches would fail to choose the "correct" model.

Finally, we make some comments on the expected asymptotic behaviour of the contingency table. As $T_{eval} \to \infty$ the MAE loss function will be minimised for the true model[2] leading all non-zero values to be concentrated in columns D and E. Since BIC is a model consistent criterion, as $T_{train} \to \infty$ eventually all non-zero values will concentrate in rows D and E. However, there are a number of settings where the selection criteria do not lead to concentration on the true model. One is where T_{train} is finite, another is where the selection criterion is not model consistent (e.g., AIC), and yet another is where the true DGP in not included in the candidate models (Diebold, 1991). In all of these cases we can exploit knowledge from the contingency table and improve model selection and combination through revised base-rates.

[2]All models we consider are symmetric in the sense that the predictive mean and median are equal. If the predictive mean is used for forecasting then choosing the model based on MAE may not converge to the true DGP.

3.3 Measuring performance

Following Makridakis et al. (2020), we consider two forecasting performance measures. The first focuses on point forecast accuracy, while the second assesses the performance of prediction intervals. Let y_t be the observation of $\boldsymbol{y}$ at time period t and f_t, u_t, and l_t the point forecast, the upper, and the lower bound of the prediction interval for the same period, respectively. Also, let T be the length of in-sample data for $\boldsymbol{y}$ and m its seasonal frequency. Since we will evaluate forecasts of all simulated series, we require measures that can be averaged across many time series. For point forecasts a widely-used measure of forecasting accuracy with this property is the Mean Absolute Scaled Error (MASE), defined as

$$\text{MASE} = \frac{1}{h} \frac{\sum_{t=T+1}^{T+h} |y_t - f_t|}{\frac{1}{T-m} \sum_{t=m+1}^{T} |y_t - y_{t-m}|}.$$

The Mean Scaled Interval Score (MSIS) is used as a measure of the performance of the prediction intervals. It is the scaled average difference between the upper and lower bound of the prediction interval plus a penalty for the instances where the actual observation lies outside the intervals. This penalty is linked to the desired confidence level, $(1-\alpha)\times 100\%$. In this study, we set $\alpha = 0.05$ (95% confidence level). The MSIS is defined as

$$\text{MSIS} = \frac{1}{h} \frac{\sum_{t=T+1}^{T+h} \left(u_t - l_t + \frac{2}{\alpha}(l_t - y_t)\mathbb{1}\{y_t < l_t\} + \frac{2}{\alpha}(y_t - u_t)\mathbb{1}\{y_t > u_t\} \right)}{\frac{1}{T-m} \sum_{i=m+1}^{T} |y_t - y_{t-m}|},$$

in which $\mathbb{1}\{\cdot\}$ is an indicator function. Note that the scaling of MSIS is the same as in

MASE. For both measures, lower values are better.

3.4 Results

We now compute the performance measures introduced using a double rolling window. For each outer window we first compute the contingency table using an "inner" rolling window of 29 one step ahead forecasts and a training sample size of 60 points. We then retrain all models using all 90 available data points used to compute the contingency table and produce one step ahead forecasts. The size of the outer rolling window is 90. Both here and in the remaining empirical applications, information on all series up to, but not past the forecast origin is used to form forecasts. We also note that in our contingency table there are rows entirely made up of zeroes (models *G*, *N*, *O*, and *P*). If these models are selected during the validation, the forecasts for the *precision-select* and *sensitivity-select* approaches are set to be the same as for the *criterion-select* approach. A similar strategy is used for the methods based on averages.

Along with the results for the various selection and combination schemes, we offer the results for three benchmarks. The first is selecting the best model according to the evaluation criterion for all series, which we refer to as *aggregate-select*. The second is an equally weighted forecast combination, referred to as *EQW-average*. The third is the method of Diebold and Shin (2019) which involves regressing realisations of the target variable on forecasts, but shrinking towards equal weights using an L1 or L2 penalty. We implement the Diebold and Shin (2019) method using an L1 penalty and the R package `glmnet` with default settings for selecting the regularisation parameter by cross-validation. We note that this approach can provide negative weights making it ill-suited to computing prediction intervals. Therefore, results for MSIS are not presented for the Diebold and Shin (2019) method. We also implemented a modification of the Diebold and Shin (2019) method whereby all negative weights are set to zero and

weights are then normalised to sum to one. Here and in the remainder of the paper this is referred to as *normalised DS*.

Table 4: Mean Absolute Scaled Error (MASE) and Mean Scaled Interval Score (MSIS) for the simulated data described in Section 3.1. The top four lines refer to selection methods, while the bottom four lines to combination methods. Precision/Sensitivity are our proposed methods for exploiting cross-learning. The other methods, serving as benchmarks, do not exploit revised base-rates. The best results in each column and each panel are shown in bold.

Method	MASE	MSIS
Aggregate-select	0.956	6.82
Criterion-select	0.928	7.04
Precision-select	0.928	7.23
Sensitivity-select	**0.905**	**6.68**
EQW-average	**0.850**	**5.66**
Diebold & Shin (2019)	0.886	-
Normalised DS	0.853	5.75
Criterion-average	0.903	7.04
Precision-average	0.887	6.62
Sensitivity-average	0.875	6.33

The MASE and MSIS are reported in Table 4. We can make some general conclusions. First, *aggregate-select* uses the forecasts from the most commonly selected model (in most windows, model E) for all series. This performs poorly since 40% of the series are drawn from a DGP without seasonal behaviour. Second, model averages outperform the corresponding selection methods. Third, and most critical, *sensitivity-select* and *sensitivity-average* outperform *criterion-select* and *criterion-average*, respectively. Simi-

larly, *precision-select* is on par with *criterion-select* and *precision-average* is more accurate than *criterion-average*. While these differences are small, they are statistically significant according to a non-parametric Nemenyi test that controls for multiple comparisons (Koning et al., 2005). The best performing method for this particular simulation is equal weights - a robust choice that is often difficult to beat, with *normalised DS* also performing well. Nonetheless, the simulation demonstrates the advantages that can be accrued by exploiting cross-learning relative to the approaches that only use series specific information.

4 Macroeconomic Forecasting

4.1 Data

We now follow a similar exercise to the simulation study, this time using real macroeconomic data. The FRED-MD data[3] contains over 100 macroeconomic time series measured at a monthly frequency that are continuously updated and revised. The same double rolling window setup used in the simulation study is used again here. As such, the starting point we use for the data is February 2003, a date chosen to exclude observations affected by the onset of Covid-19. A small number of series with missing data were excluded from the analysis leading to $n = 115$ series. We note that the FRED-MD data are pre-processed with differencing and log transformations applied to each series to ensure stationarity. This approach that is common in the literature of macroeconomic forecasting with a large number of series (see McCracken and Ng, 2016, and references therein). In light of this, it is suitable to only consider the stationary ARMA models rather than the richer ARIMA class of models.

[3]These are publicly available from the St Louis Federal Reserve.

Table 5: Contingency Table for the Fred-MD data. Rows correspond to the "selected" model, while columns correspond to the "correct" model. Model labels *A*-*P* are outlined in Table 2.

		Correct																
		A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	Total
Selected	*A*	8	6	6	1	0	0	1	0	0	0	0	0	4	0	2	0	28
	B	6	2	4	0	1	0	1	1	0	2	1	0	2	0	2	0	22
	C	3	2	5	3	0	0	5	0	0	0	1	9	2	0	4	0	34
	D	1	5	2	1	0	1	0	1	0	2	0	2	0	0	0	0	15
	E	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	F	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	I	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	J	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
	K	0	0	2	1	0	0	0	0	0	0	0	0	1	0	0	0	4
	L	0	0	2	0	0	0	0	0	0	1	0	1	1	0	1	0	6
	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	2
	P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Total	19	16	22	6	2	1	7	2	1	6	2	12	10	0	9	0	115

Table 5 displays the contingency table for one window of the FRED-MD macroeconomic data. We observe a tendency to select non-seasonal models (models *A*-*D*) even though the results of out-of-sample validation suggest that seasonal models, especially

models L and M, may be better suited to forecasting some series. Our cross-learning approach will correct for this tendency.

The results comparing our proposed methods to the same benchmarks used in Section 3.4 is shown in Table 6. Regarding MASE and the selection methods, Precision select and sensitivity select do not outperform criterion select. However the for the averaging methods outperform the precision average is the best method overall outperform all selection methods, criterion average and even equal weights. These results further demonstrate the potential of our cross-learning approach in a real-world setting.

The results in Table 5 summarise results over all time series in the database. While a difference of 0.002 may seem small, it should be noted that most series are transformed using log differences, and to summarise forecast accuracy across series, MASE is used which further standardises the forecast error metric. To give an idea of how precision average can give an economically significant improvement over criterion average, we can consider the series *total business inventories* for which precision-average yields an mean absolute error of 0.094 compared to 0.179. Since this series is transformed to log differences a difference in MAE of represents a difference in forecast error in the growth rate of 8% or, given the scale on which inventories are measured, roughly $US 160bn. As a caveat we would note that methods that exploit cross-learning do not lead to superior forecasts for every series. For example, for the series *number of employees in financial services* criterion average outperforms precision average, with a difference in MAE of 0.0175, which when interpreted as a difference in forecasting accuracy for a growth rate, is economically significant.

Table 6: Mean Absolute Scaled Error (MASE) and Mean Scaled Interval Score (MSIS) for the FRED-MD data described in Section 4.1. The top four lines refer to selection methods, while the bottom four lines to combination methods. Precision/Sensitivity are our proposed methods for exploiting cross-learning. The best results in each column and each panel are shown in bold.

Method	MASE	MSIS
Aggregate-select	0.553	**5.04**
Criterion-select	**0.530**	5.06
Precision-select	0.537	5.10
Sensitivity-select	0.544	**5.04**
EQW-average	0.532	**4.84**
Diebold & Shin (2019)	0.553	-
Normalised DS	0.533	**4.83**
Criterion-average	0.527	4.99
Precision-average	**0.525**	4.85
Sensitivity-average	0.527	4.84

5 Large-scale empirical evaluation

5.1 Models

In this section, we focus on the exponential smoothing (ETS) family of models. In exponential smoothing models, up to three components are estimated (level, trend, and seasonality) and the forecasts are based on the estimates of these components. The exponential smoothing family consists of 30 models in total, which are all possible combinations of different types of error (additive or multiplicative), trend (none, additive,

or multiplicative; damped or not), and seasonality (none, additive, or multiplicative). An exponential smoothing model form is usually summarised by three or four letters that represent the types of the components in the model. For instance, an exponential smoothing model with additive error, additive trend, and multiplicative seasonality is acronymised as ETS(AAM), or simply AAM. Similarly, ETS(MAdN) is a model with multiplicative error, additive damped trend, and no seasonal component.

In practice not all models are used, either because some combinations result in estimation difficulties (such as additive error term with multiplicative seasonality) or in unrealistic and explosive forecasts (such as multiplicative trends). In this study, we use the `ets()` which is part of the very popular *forecast* package for the R statistical software (Hyndman et al., 2020), which is used for this study, considers by default 15 out of the 30 theoretically possible models. For non-seasonal data (such as time series with a yearly frequency), the number of available exponential smoothing models drops from 15 to 6.

5.2 Selection and evaluation criteria

We consider two selection criteria, which are described below. The evaluation criterion that we use in this study is the mean absolute error (MAE). The application of the proposed algorithm for populating the contingency table (subsection 2.1) requires the splitting of each reference time series into a training set ($z^{(n)}_{train}$ – on which the selection criteria values are calculated) and a test set ($z^{(n)}_{eval}$ – on which the evaluation criterion values are calculated). Let $T^{(n)}$ be the length of the reference time series $z^{(n)}$ and h the required forecast horizon for $\boldsymbol{y}$. The first $T^{(n)}-h$ observations of $z^{(n)}$ serve as the training data, with the last h being the test data.

The first selection criterion is an information criterion. Information criteria are based on the in-sample performance of a model, penalised for the size of the model

(number of parameters that need to be estimated). Information criteria values can be calculated using the training data of each reference time series, $z^{(n)}_{train}$. In this empirical evaluation, we present results for the BIC, similarly to the simulation and macroeconomic data of Section 3 and Section 4, respectively. However, the insights are consistent for other information criteria such as the AIC or its corrected version for small sample sizes (AICc).

The second selection criterion is time series validation. Replacing information criteria with time series validation allows us to directly match the cost functions of the selection and the evaluation criteria in constructing the base-rate matrix. However, this requires further splitting the series such that selection via validation is enabled. We first consider the first $T^{(n)} - 2h$ observations of $z^{(n)}$ and prepare forecasts for the periods $T^{(n)} - 2h + 1$ to $T^{(n)} - h$. The model that performs best (based on MAE) on the periods $T^{(n)} - 2h + 1$ to $T^{(n)} - h$ is the selection via time series validation. Next, we take the first $T^{(n)} - h$ observations of $z^{(n)}$, corresponding to $z^{(n)}_{train}$, and prepare forecasts for $z^{(n)}_{eval}$ to calculate the evaluation criterion, similarly to information criteria. We note that the benchmark due to Diebold and Shin (2019) also requires a splitting of the sample, therefore can only be compared to the validation results below.

5.3 Data

We use the yearly, quarterly, and monthly data from the M, M3, and M4 forecasting competitions (Makridakis et al., 1982; Makridakis and Hibon, 2000; Makridakis et al., 2020). The forecasting horizon considered in this study is different per data frequency, which matches the original design of the aforementioned competitions: $h = 6$, 8, and 12 for the yearly, quarterly, and monthly data respectively. Following the process described in 2.1, we populate a separate contingency table per frequency and selection criterion (BIC or time series validation).

The various exponential smoothing models available have different numbers of parameters to be estimated and, as such, require a different minimum number of available observations. Because of that, we need to ensure that models are selected for their merits and not their data requirements. This would be important for the shorter of the series available, as their inclusion would introduce a bias when populating the contingency tables. As such, for each selection criterion, only the series that could be fitted over all available exponential smoothing models were considered. Table 7 provides the respective counts.

Finally, the out-of-sample evaluation takes place on the series that both selection criteria can be applied, which matches the counts of series for constructing the BIC's contingency tables. That means that for the case of selection with validation, there is not an absolute match between the series used for constructing the base-rate matrices and the series that are finally evaluated. This mismatch would be a normal situation for cases where only a small number of series needs to be forecast, with the corresponding contingency table being populated using a wider, representative set of time series.

Table 7: Number of series considered for constructing the base-rate matrices for each selection criterion, i.e., BIC and time series validation.

Frequency	BIC base-rate	Validation base-rate
Yearly	20,616	15,315
Quarterly	24,820	24,327
Monthly	49,998	49,477
Total	95,434	89,119

5.4 Out-of-sample evaluation

Tables 8 presents the average values of MASE and MSIS on the out-of-sample performance for the various selection and combination schemes considered (see section 2.2), for each selection criterion (see section 5.2), and for each and data frequency. As de-

scribed in section 5.3, the out-of-sample evaluation takes place over 95,434 yearly, quarterly, and monthly series for which both selection criteria (BIC and Validation) can be applied. The best performances (lower MASE or MSIS value) for each scheme (selection or combination) are highlighted in boldface. We excluded, both from the selection and combination schemes, the exponential smoothing models where the information criterion values could not be estimated or the lower or upper prediction interval of the furthest horizon was an outlying value, as determined by the interquartile range of the forecasts produced by the examined models.

Table 8: The out-of-sample performance of the various selection and combination schemes.

Selection Criterion	Scheme	Yearly		Quarterly		Monthly	
		MASE	MSIS	MASE	MSIS	MASE	MSIS
BIC	Aggregate-select	3.512	45.524	1.192	9.953	0.988	8.893
	Criterion-select	3.412	33.175	**1.166**	**9.503**	0.949	**8.175**
	Precision-select	3.490	44.494	1.184	9.888	0.985	8.649
	Sensitivity-select	**3.309**	**32.329**	1.174	10.368	**0.948**	8.327
	EQW-average	3.231	29.225	1.174	9.099	0.948	8.213
	Criterion-average	3.351	31.652	1.152	9.332	0.942	8.098
	Precision-average	3.247	29.935	**1.147**	**9.023**	**0.916**	**7.933**
	Sensitivity-average	**3.212**	**29.180**	1.155	9.032	0.922	7.961
Validation	Aggregate-select	3.512	45.524	1.192	**9.953**	0.988	8.893
	Criterion-select	**3.358**	**38.937**	1.179	10.182	**0.942**	**8.640**
	Precision-select	3.511	45.265	1.188	10.083	0.972	8.756
	Sensitivity-select	3.374	39.119	**1.178**	10.144	0.951	8.789
	EQW-average	3.231	29.225	1.174	9.099	0.948	8.213
	Diebold & Shin (2019)	5.084	-	1.583	-	1.346	
	Normalised DS	3.249	30.387	1.173	9.104	0.946	8.246
	Criterion-average	3.348	37.631	1.176	9.934	0.936	8.478
	Precision-average	3.251	30.004	**1.159**	**9.071**	**0.925**	**8.053**
	Sensitivity-average	**3.214**	**29.149**	1.170	9.083	0.935	8.115

We observe that, when a single model is selected, the *criterion-select* scheme (regardless of the selection criterion) provides a reasonably good performance. *Sensitivity-select* is better than *criterion-select* in the yearly frequency and when BIC is used. Also,

precision-select offers better performance with respect to prediction intervals for the quarterly data and the Validation selection criterion. However, it would be fair to say that there is not much to be gained from the contingency tables and the environmental information when a single model is selected. Overall, and as expected, *aggregate-select* results in worse performance compared to other selection schemes, with the only exception being the MSIS for the quarterly frequency.

The situation is different when models are combined. In this case, the *criterion-average* scheme focuses on weights that have been estimated based on the information criteria values (for the BIC) or the validation performance of the models for a single time series, disregarding the general tendencies and performances of these models. On the other hand, *precision-select* and *sensitivity-select* take into account the performance of each model when applied to a large set of series. *Sensitivity-average* is the best approach for the yearly frequency, outperforming all other selection and combination approaches, including *EQW-average*. Similarly, *precision-average* is the best approach for the seasonal (quarterly and monthly) frequencies. The gains from the application of *precision-average* and *sensitivity-average* are especially evident in the case of the performance of the prediction intervals. For example, the MSIS value for the *sensitivity-average* at the yearly frequency is 7.8% and 22.5% lower than the respective values of *criterion-average* for the BIC and Validation criteria, which, by turn, are lower than the respective *criterion-select* values.

Comparing the results of Table 8 for the two selection criteria considered, BIC and Validation, we can generally notice small differences. However, BIC is overall better than Validation for the various selection and combination schemes. Regardless, the proposed selection/average schemes are by definition applicable for any selection criterion, as long as the respective contingency tables can be populated.

Next, we perform non-parametric multiple comparisons using the Friedman and the

post-hoc Nemenyi tests. The results from the application of these tests allow us to check whether or not the differences between the performance of the various selection and combination schemes are statistically significant. It is worth noting that these tests do not rely on distributional assumptions, while they focus on the ranked rather than the absolute performance of each scheme. We use the `nemenyi()` function of the *tsutils* package for R (Kourentzes, 2020). The significance results at a 5% level are presented in figures 1 and 2 for the MASE and the MSIS respectively. The considered schemes are presented from best (top row) to worst (bottom row) based on their average ranks. The columns' order follows the presentation of the schemes in the Table 8. For each row, the black cell represents the scheme being tested; blue cells suggest that the scheme depicted in the row has an average rank that is not statistically different than the scheme in the respective column; and white cells suggest statistically significant differences. As an example, focusing on the first panel of Figure 1 (yearly data and BIC selection criterion), the equal-weighted average ("EQW-Ave"), which has an average rank of 4.51, is not statistically different, at a 5% level, to *sensitivity-select* ("Sens-Sel"), but it is statistically different to all other selection and combination schemes.

Three major observations arise from Figures 1 and 2. First, the *precision-average* scheme is ranked always first in terms of MASE, regardless of the frequency of the data or the selection criterion. Moreover, it is statistically better than all other schemes, with the only exception being the quarterly data and the BIC criterion where there is no evidence of statistical different average ranks between *precision-average* and *criterion-average*. Second, the good performance of the *precision-average* scheme is also evident in the yearly and monthly frequencies for the MSIS measure. However, the other combination scheme that utilises revised base-rate information, the *sensitivity-average*, is significantly better than all others in the quarterly data. Third, there is no evidence that one of the selection schemes, *aggregate-select*, *criterion-select*, *precision-select*, and

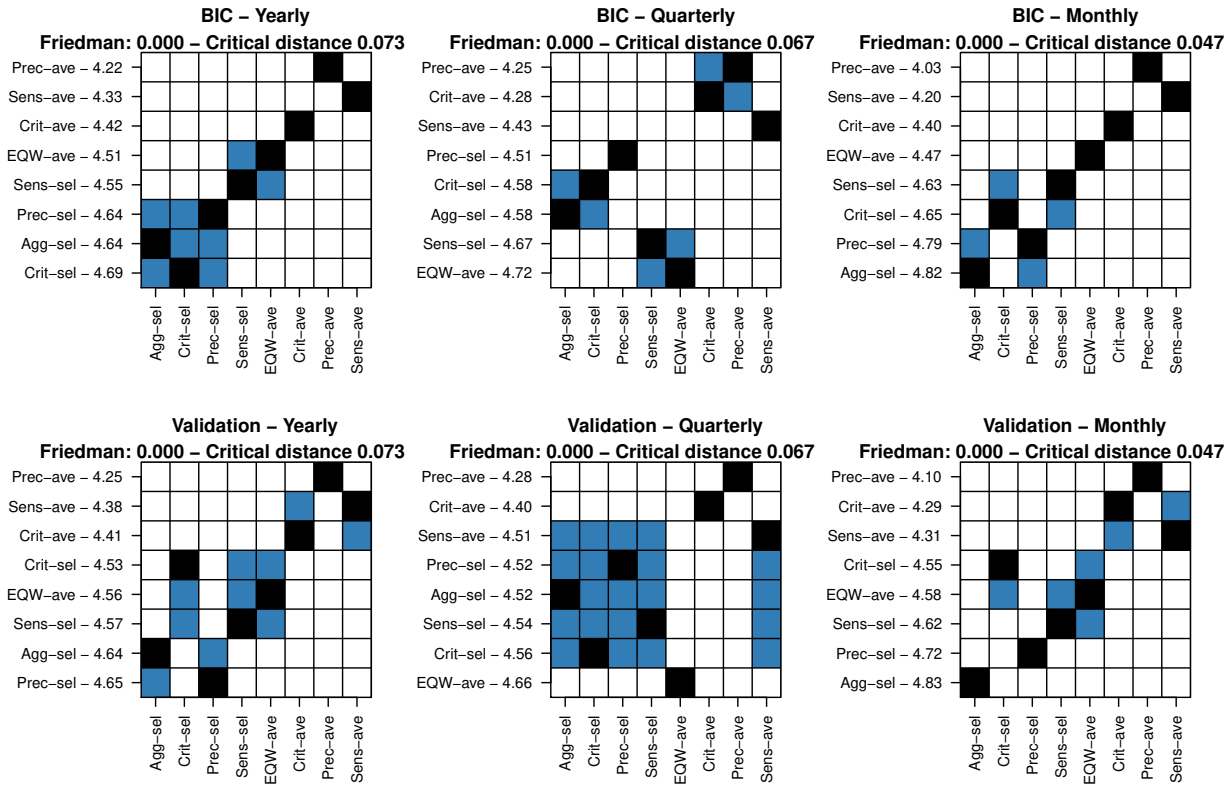

Figure 1: Nemenyi test results at a 5% significance level for the MASE.

sensitivity-select, performs consistently better than the others.

Koning et al. (2005) showed that the above significance test may tend to over-reject the null when analysing a large number of series. We repeated the above analysis focusing only on the series labelled as macroeconomic, which are roughly 20% of the M competition series. The *precision-average* scheme is still the best option, frequently being statistically better than the other approaches.

Additionally, we also apply the model confidence set (Hansen et al., 2011) that selects a set of models, jointly non-significantly different from one another, and containing the best model with a certain level of confidence. We use the `MCSprocedure()` function of the *MCS* package for R (Bernardi, 2017), with the default 0.15 alpha level

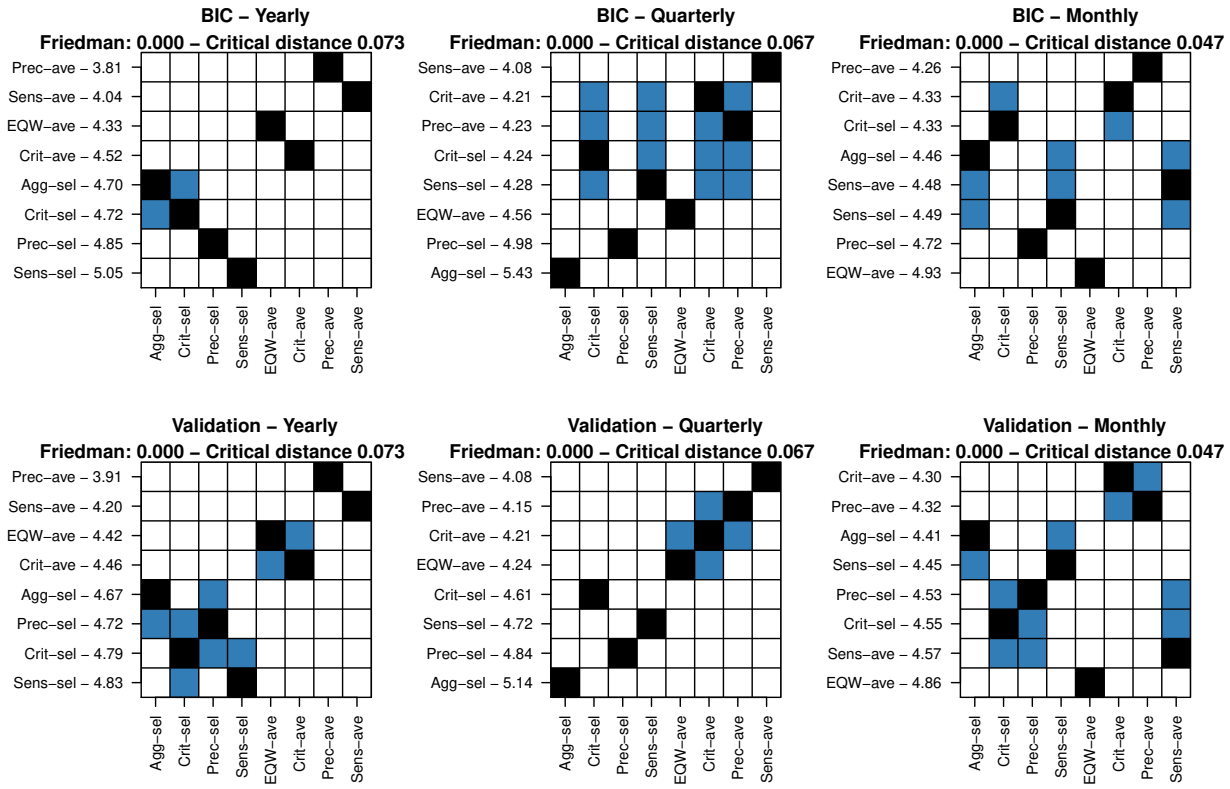

Figure 2: Nemenyi test results at a 5% significance level for the MSIS.

(85% confidence level) and 5000 bootstrap replications. Focusing on forecast accuracy (MASE), we observe that *sensitivity-average* is identified as superior to all other approaches for the yearly data (for both selection criteria, BIC, and Validation), while *precision-average* is the superior approach for the quarterly and monthly data. Focusing on MSIS, *sensitivity-average* and/or *precision-average* are identified as the superior approaches. The only instance that one of the other six approaches is identified as superior is the case of MSIS for the yearly data when BIC is used as the selection criterion, where *EQW-average* is tied with *sensitivity-average* as the superior approaches. Table 9 summarises the results for the model confidence test.

Next, we focus on the frequencies with which the four selection schemes opt for a

Table 9: Superior approaches based on the model confidence test.

Measure	Frequency	Selection Criterion BIC	Validation
MASE	Yearly	Sensitivity-average	Sensitivity-average
	Quarterly	Precision-average	Precision-average
	Monthly	Precision-average	Precision-average
MSIS	Yearly	EQW-average; Sensitivity-average	Sensitivity-average
	Quarterly	Sensitivity-average; Precision-average	Precision-average
	Monthly	Sensitivity-average; Precision-average	Precision-average

model that is within the top, middle, or bottom third of the respective pool of available models. For example, given that the model pool consists of 15 exponential smoothing models (6 for the yearly data), then a scheme points to a model in the top 1/3 of the models when that model is ranked, based on its point forecast accuracy, in $[1, 5]$ (or $[1, 2]$ for the yearly data). Figure 3 presents the respective selection frequencies for each selection criterion and data frequency. We observe that *precision-select* and *aggregate-select* point to a model in the top-third more often than the other two selection schemes (*criterion-select* and *sensitivity-select*) for the yearly and quarterly data, irrespective of the selection criterion. However, *precision-select* and *aggregate-select* also opt more often than the other selection schemes a model that is ranked in the bottom-third of the models. Despite the similarities in the model ranks selected by *aggregate-select* and *precision-select*, the latter offers better overall performance, as observed in Table 8.

Another interesting observation arises from the ranked performance of the models selected by *sensitivity-select*. In three of the six panels (BIC - Yearly, BIC - Quarterly, and Validation - Yearly), we see that *sensitivity-select* opts significantly more frequently than the other two schemes for a model ranked in the middle-third and less frequently for models in either the top or bottom thirds. This suggests that *sensitivity-select* selects less frequently the best models but also avoids more frequently the worst models. In that sense, *sensitivity-select* works similarly to how humans select models (Petropoulos et al.,

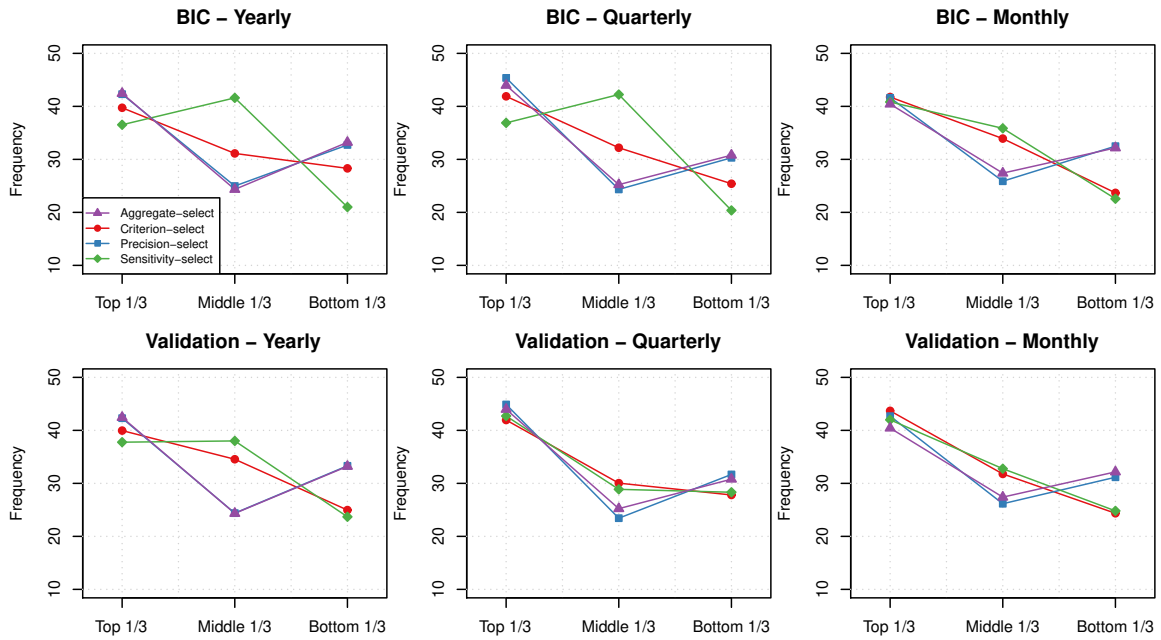

Figure 3: Selection frequencies of the top, middle, and bottom-ranked models for each selection scheme, analysed per data frequency and selection criterion.

2018). *Sensitivity-select* results in models with similar ranks to *criterion-select* for the monthly data frequency but also for the quarterly data and the validation criterion.

5.5 Sensitivity analysis

So far, our main empirical results focused on the use of very large sets of reference series, without paying attention on the domain/context of each series or even whether the number of series in the reference set would alter the results. In this section, we will empirically show that the performance of the sensitivity and precision approaches (both selection and average) are insensitive to the domain of series. We will also show that these approaches work well even in the cases of limited number of reference series.

We first present the accuracy (MASE) and uncertainty (MSIS) results for the monthly data in the M competitions labelled as "macroeconomic". The performance of the different approaches is measured when (*i*) all monthly M competition series or (*ii*) only

the series of the particular category (macroeconomic) are considered to populate the contingency table. We focus on the monthly macroeconomic data as this complements the results of our case study in the previous section. However, the insights presented below are similar for other domains (industry, demographic, micro, and finance) and other data frequencies (yearly and quarterly). The results are presented in Table 10 and refer to the case when the Validation criterion is used. We observe that the average performance of the sensitivity and precision approaches is not affected by the choice of the reference data. More importantly, the relative rankings of all the approaches remains largely the same.

Table 10: The out-of-sample performance of the various selection and combination schemes for the monthly macroeconomic series of the M competitions, when all data or just the macroeconomic data are used as the reference data.

	All data...		Macro data...	
	...as reference data			
Scheme	MASE	MSIS	MASE	MSIS
Aggregate-select	0.991	8.734	0.991	8.735
Criterion-select	**0.960**	9.033	**0.960**	9.033
Precision-select	0.978	**8.684**	0.983	**8.712**
Sensitivity-select	0.973	9.383	0.968	9.285
EQW-average	0.966	8.404	0.966	8.404
Criterion-average	0.956	8.828	0.956	8.828
Precision-average	**0.936**	**8.127**	**0.933**	**8.084**
Sensitivity-average	0.950	8.281	0.953	8.316

Next, we examine the effect of the size of the reference set of series. We do so both for all data but also for the macroeconomic data only. In both cases, we match the reference set with the evaluation set. This means that when we evaluate all data, we also use all data as the reference set, using appropriate subsets of each series to render the out-of-sample forecast evaluation fair. We randomly consider random samples of each reference set such that the percentage of series in the sample is in the range 5% to 100%. Given that the monthly time series in the validation base-rate (see Table 7)

is 49,477, the size of the reference sample for populating the contingency table for "all data" varies between 4,948 and 49,477. Similarly, the size of the reference sample for populating the contingency table for the "macroeconomic data" varies between 1,040 and 10,401. For each possible size, we take five such random samples. The results of this analysis when the Validation criterion is used are presented in figures 4 and 5 for MASE and MSIS respectively.

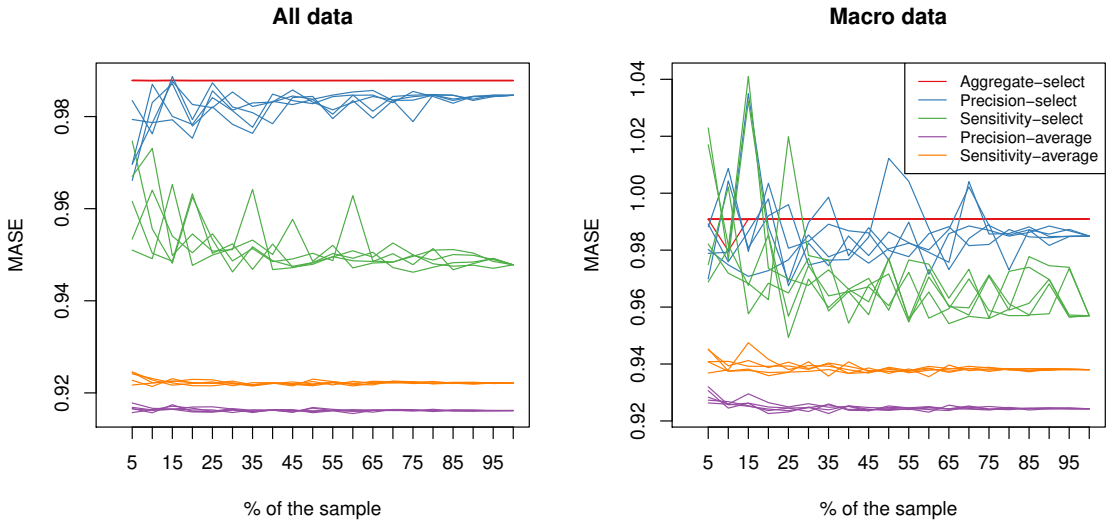

Figure 4: The sensitivity of the MASE results on the size of the reference set. Left panel: all data; Right panel: macroeconomic data. Reference series at 100% of the sample match the evaluation series.

We observe that the *precision-select* and *sensitivity-select* approaches are sensitive to the number of series in the sample size but also to the composition of the sample itself. As such, their performance fluctuates considerably, and in some cases they are worse than *aggregate-select*. Contrary to this, the *precision-average* and *sensitivity-average* approaches are very robust against the sample size, with their performance only slightly affected by the series (and count of series) contributing in the sample. As such, we can conclude that one does not need access to tens of thousands of series to achieve performance improvements using approaches such as *precision-average* and *sensitivity-average*.

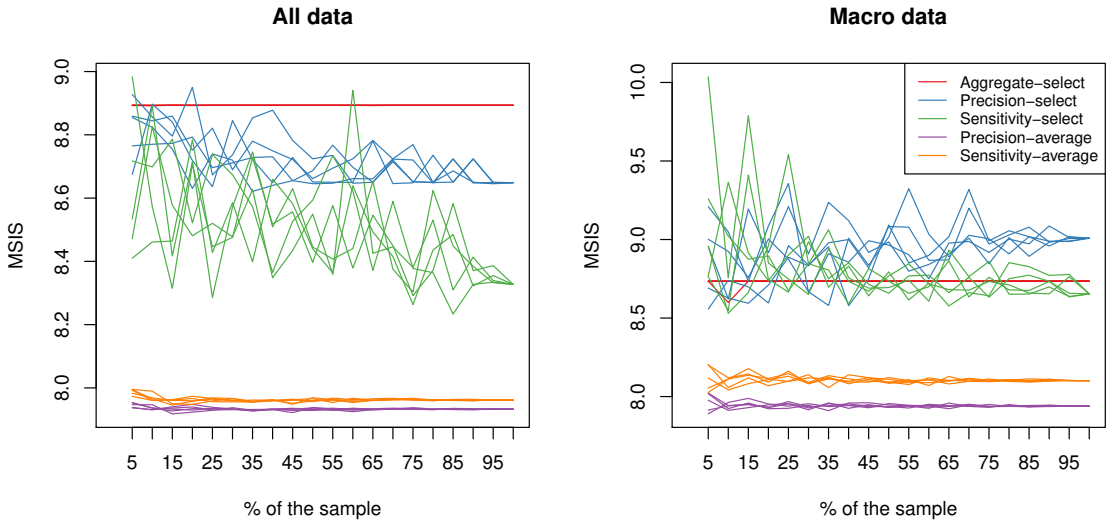

Figure 5: The sensitivity of the MSIS results on the size of the reference set. Left panel: all data; Right panel: macroeconomic data.

As a practical recommendation, we can consider an analysis similar to the above to choose the number of reference series to use in the analysis. In particular we can randomly select different sub-samples of all available reference series, each of size k, and then evaluate the forecasts. We can then choose the number of reference series such that the variance of forecast performance across the different sub-samples of size m is below some threshold.

6 Discussion

Our empirical results suggest that combining models using precision and sensitivity information can significantly improve the performance of a system rather than focusing solely on the per-series information. Effectively, our results show that the base-rate information has a useful role to play in model selection and model combination. Suitably revising the base-rates offers a forecasting performance can be superior to exclusively focusing on the case-specific information. Solely focusing on the base-rates, as showcased

by the *aggregate-select* scheme, is not appropriate and a balance is needed between the environmental (aggregate) and individual (per series) information regarding the performance of a forecasting model.

Our approach is a very simple case of cross-learning, as the base-rate information is built by the application of the models within the pool on a large number of time series. Compared to other approaches that utilise cross-learning, our approach offers more transparency while being more intuitive. Normally, cross-learning (and meta-learning) approaches require extensive feature engineering (Montero-Manso et al., 2020), preprocessing or scaling of the data (Kang et al., 2020), and hyper-parameter optimisation. In contrast, our approach involves a small number of modelling decisions, mostly related to the choice of the pool of forecasting models, the selection of selection and evaluation criteria, and the choice of the reference series. Especially with regards to the latter, our sensitivity analysis shows that considering pools of series that are more generic than the target series does not harm forecast accuracy, while the *precision-average* and *sensitivity-average* approaches are robust against the size of the reference set of series. Our proposition is widely automated and requires limited judgemental input from the modeller. As such, our approach on forecast combinations based on revised base-rates could be offered as part of automatic and batch forecasting solutions.

Similar to other cross-learning approaches, our approach consists of an offline and an online part. The offline part corresponds to the population of the contingency tables, while the online part corresponds to the use of these tables to estimate the revised base-rates. It is worth-mentioning that the additional calculations for the online part, once the values of the selection criteria have been estimated, are trivial and result in negligible additional computational cost. However, populating the contingency tables can be costly if the reference set of series is large, even more for the validation selection criterion compared to information criteria such as the BIC. In any case, it

would be reasonable to assume that, in a relatively constant environment, the offline part would not be updated in every review period (i.e., every time one needs to produce forecasts). Therefore, in forecasting settings where the contingency tables do not have to be re-constructed regularly, the proposed base-rate approaches become "economically" meaningful, even when the accuracy improvements they achieve over simpler model selection and combination methods are relatively small. On the contrary, in forecasting applications that may require frequent contingency tables updates, the trade-off between accuracy improvements and additional computational costs should be carefully evaluated.

Usually, weighted-based combination approaches estimate a unique set of weights for each target series. An example is the *criterion-average* approach, where the weights are estimated based on the values of the selection criterion for each model when applied to a particular (the target) series. However, this is not true for the *precision-average* and *sensitivity-average* approaches. We calculate only K sets of weights for each of these approaches, with each set of weights being applied based on the model that is selected. In this sense, our combination weights are "static" given a reference set of series. It is not the first time that static combination weights are proposed in the literature. For instance, Collopy and Armstrong (1992) proposed the use of static weights when combining between four models towards estimating levels and trends. However, contrary to them, our static weights are not arbitrarily selected but are directly linked with the environmental performance (base-rates) of the models. As such, the combination weights for *precision-average* and *sensitivity-average* will change if the set of reference series used for calculating the base-rate also changes.

One advantage of model combinations through revised base-rates is that they do not rely on specific selection or evaluation criteria, or even a standard pool of models. In this study, we focused on a single evaluation criterion (MAE) solely for purposes

of brevity; the choice of MAE was made so that it links to the performance indicator used for measuring point-forecast accuracy (MASE). However, a different evaluation criterion could be more appropriate in other settings. Moreover, in our work we showed results for two selection criteria, BIC and Validation. The results are similar for other information criteria (such as AIC or AICc), while our approach could work with any other selection criterion, such as cross-validation. Finally, we limited our pool of models to either the exponential smoothing family of models (for the case of the large-scale empirical evaluation) or the ARMA family of models (for the case of the simulation study and the macroeconomic data application). Recall, however, that as long as the selection criteria values are comparable, then one could consider a pool that includes forecasting methods or models from several different families.

In our empirical design, we use suitable sub-series of the target series to form the reference set of series and populate the contingency tables necessary for the precision-based and sensitivity-based schemes. We withheld an appropriate number of observations such that the evaluation criterion is calculated over a period that matches the required forecast horizon of the target series. The use of sub-series of the target series inherently offers contingency tables that are representative to the target series. We also used large, generic reference sets of series to produce forecasts for specific categories of data (see subsection 5.5). Such generic sets of series offer accuracy levels that are very similar (if not better) than using a reference set that closely represents the target data. While one could focus on reference series that are more representative to the target series (see Kang et al., 2017; Spiliotis et al., 2020, for details on measuring series representativeness) to reduce the computational cost, modelling decisions can be simplified by selecting large, diverse sets of reference series.

In this study, we used a large set of real data pooled from three major forecasting competitions (M, M3, and M4 including industry, demographic, micro, and finance se-

ries among other) and FRED (macroeconomic series). We should highlight that our results are based on relatively low frequency data (monthly to yearly) but we have no reason to believe that the insights gained cannot be generalised to other, higher frequency data. Similarly, although our study examined particular data domains, the sensitivity analysis performed suggests that our insights should be applicable to other domains as well. Also, we would like to mention that we do not intend to directly compare the achieved performances presented in this paper with any of the original submissions in the aforementioned forecasting competitions. While we did not use the test data explicitly, having access to the hold-out data renders any comparison with the competitions' participants unfair.

7 Concluding remarks

In this study, we argued that the selection of models for time series forecasting should not exclusively focus on the values of selection criteria applied on each series individually, but the base-rate information should also be taken into account, i.e., how often a particular model performs best on the out-of-sample. We argued that such "environmental" information of the performance of the various models should be revised with the case-specific information towards obtaining probabilities that each of the candidate models is indeed the correct one. Such probabilities can then be used for model (forecast) selection or forecast averaging. Our approach is a very simple case of cross-learning, while also being in-line with the agenda of Bayesian inference through loss functions.

Our empirical analysis was based on the point-forecast accuracy and performance of the prediction intervals using both simulated and real life time series. Our results showed that combination approaches based on precision and sensitivity information

can outperform both individual and aggregate selection or combination, while conceptually being in-between the two. In some cases, the differences in performance were statistically significant. The insights gained were similar for the two selection criteria (BIC and validation), the various sampling frequencies, and the two measures (MASE and MSIS) considered.

Future research could focus on context-specific data and how contingency tables can be populated to better suit the needs of organisations with a small and relatively uniform sets of data. Moreover, in this paper we limited our attention to either exponential smoothing models or ARMA models. As our approach is not structurally limited to these models, it would be interesting to see how it performs when selecting and combining over a more diverse pool of models. A final promising avenue of future research will be to see whether the algorithms proposed can be extended beyond forecasting, to model combination for estimating common parameters across models, in the spirit of (Lavancier and Rochet, 2016).

References

Bergmeir, C., Benítez, J. M., 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences 191, 192–213.

Bernardi, L. C. . M., 2017. MCS: Model Confidence Set Procedure. R package version 0.1.3.

Bissiri, P. G., Holmes, C. C., Walker, S. G., 2016. A general framework for updating belief distributions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78 (5), 1103–1130.

Caldeira, J. F., Moura, G. V., Santos, A. A., 2016. Predicting the yield curve using forecast combinations. Computational Statistics & Data Analysis 100, 79–98.

Collopy, F., Armstrong, J. S., 1992. Rule-Based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. Management Science 38 (10), 1394–1414.

Conflitti, C., De Mol, C., Giannone, D., 2015. Optimal combination of survey forecasts. International Journal of Forecasting 31 (4), 1096–1103.

Diebold, F. X., 1991. A note on bayesian forecast combination procedures. In: Economic Structural Change. Springer, pp. 225–232.

Diebold, F. X., Shin, M., 2019. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. International Journal of Forecasting 35 (4), 1679–1691.

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

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

Guedj, B., 2019. A Primer on PAC-Bayesian Learning.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. Econometrica 79 (2), 453–497.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T., 1999. Bayesian model averaging: A tutorial. Statistical Science 14 (4), 382–401.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2020. forecast: Forecasting functions for time series and linear models.

Inoue, A., Kilian, L., 2006. On the selection of forecasting models. Journal of Econometrics 130 (2), 273–306.

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

Kang, Y., Spiliotis, E., Petropoulos, F., Athiniotis, N., Li, F., Assimakopoulos, V., 2020. Déjà vu: A data-centric forecasting approach through time series cross-similarity. Journal of Business Research.

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., 2005. The M3 competition: Statistical tests of the results. International Journal of Forecasting 21 (3), 397–409.

Kourentzes, N., 2020. tsutils: Time Series Exploration, Modelling and Forecasting. R package version 0.9.2.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., Assimakopoulos, V., 2021. On the selection of forecasting accuracy measures. Journal of the Operational Research Society 0 (0), 1–18.

Lavancier, F., Rochet, P., 2016. A general procedure to combine estimators. Computational Statistics & Data Analysis 94, 175–192.

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

Loaiza-Maya, R., Martin, G. M., Frazier, D. T., 2020. Focused Bayesian Prediction.

Lütkepohl, H., 1984. Linear transformations of vector ARMA processes. Journal of Econometrics 26 (3), 283–293.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R., 1982. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting 1 (2), 111–153.

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

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

McCracken, M. W., Ng, S., 2016. Fred-md: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34 (4), 574–589.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., Zilberman, E., 2021. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. Journal of Business & Economic Statistics 39 (1), 98–119.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., Talagala, T. S., 2020. FFORMA: Feature-based forecast model averaging. International Journal of Forecasting 36 (1), 86–92.

Nieto, F. H., Pena, D., Saboyá, D., 2016. Common seasonality in multivariate time series. Statistica Sinica, 1389–1410.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., Siemsen, E., 2018. Judgmental selection of forecasting models. Journal of Operations Management 60, 34–46.

Pettenuzzo, D., Timmermann, A., 2017. Forecasting macroeconomic variables under model instability. Journal of Business & Economic Statistics 35 (2), 183–201.

Raftery, A. E., 06 1996. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 83 (2), 251–266.

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

Stock, J. H., Watson, M. W., 2012. Generalized shrinkage methods for forecasting using many predictors. Journal of Business & Economic Statistics 30 (4), 481–493.

Talagala, T., Hyndman, R., Athanasopoulos, G., 2018. Meta-learning how to forecast time series. Monash Econometrics and Business Statistics Working paper series 06/18.

Timmermann, A., 2006. Forecast combinations. Handbook of economic forecasting 1, 135–196.

Wang, X., Smith-Miles, K., Hyndman, R., 2009. Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. Neurocomputing 72 (10), 2581–2594.

Wolpert, D. H., 1992. Stacked generalization. Neural networks 5 (2), 241–259.