International Journal of Forecasting 33 (2017) 48–60

Contents lists available at ScienceDirect

International Journal of Forecasting

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

# Model Confidence Sets and forecast combination

Jon D. Samuels [a], Rodrigo M. Sekkel [b,*]

[a] BEA, United States

[b] Bank of Canada, Canada

ARTICLE INFO

*Keywords:*
Model combination
Performance-based weighting
Trimming

ABSTRACT

A longstanding finding in the forecasting literature is that averaging the forecasts from a range of models often improves upon forecasts based on a single model, with equal weight averaging working particularly well. This paper analyzes the effects of trimming the set of models prior to averaging. We compare different trimming schemes and propose a new approach based on Model Confidence Sets that takes into account the statistical significance of the out-of-sample forecasting performance. In an empirical application to the forecasting of U.S. macroeconomic indicators, we find significant gains in out-of-sample forecast accuracy from using the proposed trimming method.



## 1. Introduction

Since the original work of Bates and Granger (1969), a myriad of papers have argued that combining predictions from alternative models often improves upon forecasts based on a single best model.[1] In an environment in which individual models are subject to structural breaks and misspecified to varying degrees, a strategy that pools information from many models typically performs better than methods that try to select the best forecasting model. When using this strategy, the forecaster faces two basic choices: which models to include in the model pool, and how to combine the model predictions. With the present easy access to large panel data sets, a vast body of research has investigated optimal model combination, but found repeatedly that a simple average of the forecasts produced by individual models is a difficult benchmark to beat, and commonly outperforms more sophisticated weighting schemes that rely on the estimation of theoretically optimal weights. This is known as the forecast combination puzzle.

While a large body of literature has examined model combination weights, Capistrán et al. (2010) pointed out that there has been little research focusing on how to choose the models to combine, given a pool of potential models. Theoretically, a potential model should be used for forecasting if has any useful information. Nevertheless, in small samples, where parameter estimation error is often pervasive, it may be that discarding predictions, that is, assigning them a zero weight, will lead to better final forecast combinations. As Aiolfi and Timmermann (2006) argued, the problem of parameter estimation error is particularly acute when the number of models is large relative to the sample size, as is often the case with large macroeconomic datasets. In such cases, trimming models could lead to better estimates of each model's weight in the combined forecast. Hence, the benefits of adding one additional forecast to the combination should be weighed against the cost of estimating additional parameters.

This paper uses a novel approach to select the models to be included in the forecast combination. In particular, we use the concept of the model confidence set (Hansen et al., 2011) to determine the statistically superior set of best models, conditional on the model's past out-of-sample performance. We compare this method with the commonly-used approach of fixing the proportion of models to keep and discarding the remaining models, without regard for the statistical significance of differences in

* Corresponding author.
*E-mail address:* rsekkel@bankofcanada.ca (R.M. Sekkel).

1 See Clemen (1989), Clemen and Winkler (1986), Hendry and Clements (2004), Makridakis and Winkler (1983), Stock and Watson (2004), and Timmermann (2006), among many others.

http://dx.doi.org/10.1016/j.ijforecast.2016.07.004

model accuracy. In the model confidence set approach, the number of models trimmed is not fixed exogenously by the econometrician, but is determined by a statistical test comparing model accuracies. In our application to the forecasting of macroeconomic indicators in the US, we employ the often-used approach of averaging the forecasts of many bivariate models,[2] and find substantial improvements in forecast combination accuracy after trimming the set of potential models to be combined using both the fixed and MCS schemes, but the gains from using the MCS approach are larger and more robust.

The idea of trimming the set of potential models prior to forecast combination is not novel. Makridakis and Winkler (1983) studied the effects of adding forecasts to a simple combination, and found the marginal benefit of adding forecasts to a simple combination to decrease very rapidly once a relatively small number of forecasts has been included. In the same spirit, Timmermann (2006) argued that the benefit of adding forecasts should be weighed against the cost of introducing an increased parameter estimation error. He considered three straightforward trimming rules: combining only the top 75%, 50% or 25% of models, based on the models' out-of-sample MSPEs.[3] The author found aggressive trimming to yield better results; in other words, including fewer models in the combination led to better forecasts. In stock return forecasting, Favero and Aiolfi (2005) also found that aggressive trimming rules based on models' $R^2$ values improved forecasts. In their application, trimming 80% of the forecasts led to the best results. When combining forecasts from various models for inflation in Norway, Bjørnland et al. (2011) argued that a strategy that combines only the 5% best models leads to the best forecast combination.

We find that significant gains for the fixed trimming method are restricted to strategies that aggressively trim 80%–95% of the models. On the other hand, the MCS trimming rule results in significant accuracy improvements for a wide range of parameters that govern the confidence level with which the set of best models is identified. Monte Carlo evidence informs the intuition that forecast accuracy gains from trimming models based on their historical out-of-sample performances arise mainly in environments in which some of the models have very little predictive ability relative to others.

The outline of the paper is as follows: Section 2 lays out the trimming schemes, while Section 3 details the results of the Monte Carlo exercise. Section 4 describes our empirical application to the forecasting of US macroeconomic variables. Finally, Section 5 concludes.

## 2. Trimming rules

Our starting point is a situation in which the forecaster has a toolbox of different models with which to predict a variable of interest $y$. Each model $i$ implies a forecast $\hat{y}_i$. These models might include naive autoregressions, Bayesian vector autoregressions, factor models, and DSGE, among others.

We first provide an introduction to the MCS, then detail how we use it as a trimming device to parse models and form conditional forecast combinations. We then contrast the results obtained using the MCS with those obtained using a rule that simply ranks the models according to their past out-of-sample forecasting performances and trims a fixed share of the worst performing models.

### 2.1. Exogenous fixed trimming

In the fixed-rule trimming scheme, the number of forecasting models to be discarded is fixed exogenously. The analysis below refers to this approach as fixed trimming. We construct the conditional forecast combination by ranking the models according to their past MSPEs, discarding a fixed proportion of models, and using the remaining ones to form the set of best forecasts. It is important to note that while the number of models to be discarded (and hence the number to be combined) is fixed exogenously, there is nothing constraining the procedure to discard the same models in each forecast period. Different models will be trimmed and used according to their respective MSPE ranks in the periods preceding the forecasting period. More formally, let $\mathcal{F}_\tau$ be a set of $i = 1, \ldots, n$ candidate models for forecasting in period $\tau$. We estimate each model $i$ using $R$ periods of data. Fixed trimming requires a training sample of $S$ periods of forecasts from each of candidate models. Thus, the first period for which we can apply fixed trimming is $R+S+1$. Individual models are estimated using data from periods $t = \tau - R, \ldots, \tau - 1$, and a rolling sample of $S$ previous forecasts is used to compare model performances. The particular rule that we employ discards a fixed proportion of the models in $\mathcal{F}_\tau$, such that

$$\mathcal{F}_\tau^* = \{i \in \mathcal{F}_\tau : MSE_{i,\tau} \leq P_\tau(x)\}, \tag{1}$$

where $P_\tau(x)$ is the $x$th percentile of $MSE_{i,\tau}$. With this trimming rule, the forecaster has to decide on the proportion of models to be trimmed. We perform a systematic analysis to show how the MSPE of the final combination would change for a wide range of different percentiles.

### 2.2. The model confidence set approach to trimming

An important drawback of the simple trimming rule discussed above is that it does not take into account the statistical significance of differences in the historical performances of the forecasting models. In principle, one might easily conjecture a situation where the best and worst forecasts have mean squared prediction errors that are not statistically different from each other. We use the model confidence set method of Hansen et al. (2011) to identify the set of best models, then trim the models that are excluded from the MCS prior to forecast combination. From a frequentist perspective, the model confidence set approach is a tool for summarizing the relative performances of an entire set of models by determining which models can be considered to be statistically superior, and at what level of significance.

[2] See for example Faust et al. (2013), Stock and Watson (2004), and Wright (2009).

[3] Timmermann (2006) used a recursive weighting scheme based on the MSE. We use a rolling window.