BANK OF CANADA
BANQUE DU CANADA

Working Paper/Document de travail
2013-11

Forecasting with Many Models: Model Confidence Sets and Forecast Combination

by Jon D. Samuels and Rodrigo M. Sekkel

www.bank-banque-canada.ca

Bank of Canada Working Paper 2013-11

April 2013

Forecasting with Many Models: Model Confidence Sets and Forecast Combination

by

Jon D. Samuels[1] and Rodrigo M. Sekkel[2]

[1]BEA, Department of Commerce, and IQSS, Harvard University
jon.samuels@bea.gov

[2]Canadian Economic Analysis Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
Corresponding author: rsekkel@bankofcanada.ca

ISSN 1701-9397

Acknowledgements

This paper is based on chapters of our dissertations at Johns Hopkins University. We would like to thank our advisors, Jon Faust and Jonathan Wright, for their advice and comments on earlier drafts. We would also like to thank Natsuki Arai, Cheng Hsiao, Maral Kichian, Eva Ortega, Gabriel Perez-Quiros, Tatevik Sekhposyan and seminar participants at Johns Hopkins University, Bank of Spain, BlackRock, Bank of Canada and INSPER for comments and useful conversations, as well as participants in the 2012 Computing in Economics and Finance and Latin American Econometric Society conferences. Jon Samuels thanks the SRC for the Robert M. Burger Fellowship, and Rodrigo Sekkel, Capes/Fulbright and the Campbell Fellowship for financial support.

Abstract

A longstanding finding in the forecasting literature is that averaging forecasts from different 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 one based on Model Confidence Sets that take into account the statistical significance of historical out-of-sample forecasting performance. In an empirical application of forecasting U.S. macroeconomic indicators, we find significant gains in out-of-sample forecast accuracy from our proposed trimming method.

JEL classification: C53
Bank classification: Econometric and statistical methods

Résumé

Les études consacrées au travail de prévision ont fait ressortir depuis longtemps que la moyenne des projections de plusieurs modèles a souvent un degré de précision plus élevé que les projections tirées d'un seul modèle, et qu'à ce titre, la technique qui consiste à établir une moyenne en pondérant les prévisions avec les mêmes coefficients donne de très bons résultats. Les auteurs se demandent ce qu'apporterait l'élagage de modèles avant le calcul des projections moyennes. À cette fin, ils comparent différentes méthodes d'élagage et proposent une nouvelle démarche (*Model Confidence Set* ou approche MCS) fondée sur la sélection de modèles selon un seuil de confiance défini par la valeur statistique de la qualité passée des prévisions hors échantillon. Un exercice empirique – la projection d'indicateurs macroéconomiques pour les États-Unis – leur permet de constater que leur démarche améliore de manière notable la précision des prévisions hors échantillon.

Classification JEL : C53
Classification de la Banque : Méthodes économétriques et statistiques

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 where individual models are subject to structural breaks and misspecified by varying degrees, a strategy that pools information from the many models typically performs better than methods that try to select the best forecasting model.[2] To use this strategy, the forecaster faces two basic choices: which models to include in the pool of models, and how to combine the model predictions. With the ease of access to large macro panel data sets, a vast body of research has investigated optimal model combination, yet have repeatedly found that a simple average of the forecasts produced by individual predictors is a difficult benchmark to beat, and commonly outperforms more sophisticated weighting schemes that rely on the estimation of theoretically optimal weights. This is the forecast combination puzzle.

While there is a large literature examining model combination weights, Capistrán et al. (2010) points out that little research has focused on how to choose the models to combine given a pool of potential models. Theoretically, if a potential model has any information for forecasting, that information should be used. Nevertheless, in small samples, when parameter estimation error is often pervasive, it is possible that discarding predictions, that is assigning them zero weight, leads to better final forecast combinations. Parameter estimation error will be particularly acute when, as argued by Aiolfi and Timmermann (2006) and Hsiao and Wan (2011), the number of

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

[2] See Hendry and Clements (2004) among many others.

models is large relative to the sample size, as it is often the case with large macroeconomic datasets. In such cases, trimming models could lead to better estimates of the remaining models' combination weight. When the relevance of a particular model is particularly small, and the estimation of its coefficients is subject to considerable uncertainty, trimming these models prior to forecast combination should also lead to better final combinations. Hence, the benefits of adding one additional variable to the combination should be weighed against the cost of estimating additional parameters.

In this paper, we use model confidence sets, as proposed by Hansen et al. (2011) in order to form forecast combinations conditional on model's past out-of-sample performance. We compare this method with the simpler and 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 model accuracy. In the model confidence approach, the number of models trimmed is not exogenously fixed by the econometrician, but is determined by a statistical test comparing model accuracy. In our application of forecasting macroeconomic indicators in the US, we use the often-used approach of averaging the forecasts of many bivariate models,[3] and find substantial improvements in forecast combination accuracy after trimming the set of potential models to be combined with both schemes, but larger and more robust gains with the MCS approach.

The idea of trimming the set of potential models before forecast combination is not novel. Makridakis and Winkler (1983) studies the effects of adding forecasts to a simple combination. They find that the marginal benefit of adding forecasts to a

[3]See, for example, Stock and Watson (2004); Faust et al. (2011); Wright (2009).

simple combination decreases very rapidly once a relatively small number of forecasts are included. In the same spirit, Timmermann (2006) argues that the benefit of adding forecasts should be weighed against the cost of introducing increased parameter estimation error. He considers three straightforward trimming rules: combining only the top 75%, top 50% and top 25% models based on the models out-of-sample MSPE.[4] The author finds that aggressive trimming yields better results, in other words, that fewer models included in the combination leads to better forecasts. In a stock return forecasting context, Favero and Aiolfi (2005) also finds that aggressive trimming rules based on model's R^2 improves forecasts. In their application, trimming 80% of the forecasts leads to the best results. When combining forecasts from various models for inflation in Norway, Bjørnland et al. (2011) argues that a strategy that combines only the 5% best models leads to the best forecast combination. We add to this literature by proposing a selection rule based on the Model Confidence Set that takes into account the statistical significance of differences between model performance, and hence is more robust than the simple strategy of ranking the models by their past performance. Whereas for the fixed trimming method, significant gains are restricted to strategies that aggressively trim 80% to 95% of the models, 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 confirms the intuition that forecast accuracy gains from trimming models based on their historical out-of-sample performance arise mainly in environments where some of the models have a very small predictive ability relative to others.

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

The outline of the paper is as follows: Section 2 lays out the trimming schemes, while Section 3 details the data and models. Section 4 discusses the benefits of trimming under different combination methods and cutoffs, and presents the results of our trimming exercise. Section 5 presents the results from a Monte Carlo study. Section 6 compares forecasts based on trimming to commonly used alternative data-rich forecasting methods. Finally, Section 7 concludes.

2 Trimming Rules

Consider a situation where the forecaster has a toolbox of different models to forecast a variable of interest y. Each model i implies a forecast $\hat{y}_i$. These models might comprise naive autoregressions, Bayesian vector autoregressions, factor models, DSGE, etc., among others. The question we address in this paper is how should the forecaster decide which of the forecasts should be included in forming the forecast combination.

We propose a conditional forecast combination strategy based on the model confidence set concept in Hansen et al. (2011). We first provide an introduction to the MCS and then detail how we use it as a trimming device to parse models and form forecast combinations conditional on the recent out-of-sample performance of each model. We then contrast the results obtained with the MCS with a trimming rule that simply ranks the models according to their out-of-sample forecasting performance and trims a fixed share of the worst performing models.[5]

[5]See Bjørnland et al. (2011) for a recent application of this method to the combination of inflation forecasts in Norway, as well as Timmermann (2006) for the US.

2.1 Exogenous Trimming

In the fixed-rule trimming scheme, the number of forecasting models to be discarded is exogenously fixed. To construct a conditional forecast combination, we rank the models according to their past MSPE, discard a fixed proportion of models, and use 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 also combined) is exogenously fixed, there is nothing constraining the procedure to discard the same models at each forecast period. Different models will be trimmed and used according to their respective MSPE rank in the periods preceding the forecasting period.

With this trimming rule, the forecaster has to decide the proportion of models to be trimmed. We do a careful analysis of showing how the MSPE of the final combination would change for the complete range of proportions.

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 performance 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. Hence, we propose a trimming rule that takes into account the statistical significance of the differences in model performance. We use Hansen et al. (2011) model confidence set method to identify the set of best models. We then trim the models that are excluded from the MCS prior to forecast combination. The model confidence set approach is, from a frequentist perspective, a tool to summarize

the relative performance of an entire set of models by determining which models can be considered statistically superior, and at what level of significance. The interpretation of a MCS for a set of models is analogous to a confidence interval for a parameter, in the sense that the models covered by the MCS are those that can not be rejected from the set of best models for a given level of confidence. By attaching p-values to models, it is easy to ascertain at what level of significance individual models would be in the set of superior models, and which would be eliminated as statistically inferior. By eliminating models with the MCS, we examine a trimming method that has a clear grounding in statistical theory. The MCS provides the analyst with a less arbitrary trimming strategy that has a clear frequentist interpretation.

We keep Hansen et al. (2011) notation, and refer the reader to the original paper for a more detailed exposition. The starting point for our application is a finite set $\mathcal{M}_0$ of forecasting models. The MCS aims at identifying the set $\mathcal{M}^*$, such that:

$$\mathcal{M}^* = \{i \in \mathcal{M}_0 : u_{i,j} \leq 0 \text{ for all } j \in \mathcal{M}_0\}$$

where $u_{i,j} = E(d_{ij,t})$ is the expected loss differential between models and $d_{ij,t} = L_{i,t} - L_{j,t}$ is the model loss differential, where in our application we choose $L_{i,t} - L_{j,t}$ as the squared forecast error. That is, given the set of all forecasting models $\mathcal{M}_0$ in the comparison set, the MCS searches for the set of models that cannot be rejected as statistically inferior at a chosen level of confidence.

The MCS is based on the following algorithm: Starting from the set of all models $\mathcal{M}_0$, repeatedly test the null hypothesis at significance level α of equal predictive accuracy, $H_{0,\mathcal{M}} : u_{i,j} = 0 \ \forall i, j$. If the null is rejected, the procedure eliminates a model

from $\mathcal{M}$, and repeats until the null of no difference between models can not be rejected at the chosen level of significance. The set $\hat{\mathcal{M}}^*_{1-\alpha}$ with the remaining models is denoted as the MCS, $\mathcal{M}^*$.[6]

To test H_0, which is done sequentially through the case when the null is not rejected, we construct t-statistics based on $\bar{d_{ij}} \equiv n^{-1}\sum_{t=1}^{n} d_{ij,t}$, the relative loss of model i relative to model j. The pertinent test statistics are

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{v\hat{a}r(d_{ij})}}$$

$$T_{\mathcal{M}} = \max_{i,j\in\mathcal{M}} |t_{ij}|$$

The $T_{R,\mathcal{M}}$ statistic imposes that whether or not the null of no difference in model performance is rejected depends only on the model that has the greatest relative loss. This statistic is particularly convenient in implementation because the decision rule of which model to eliminate is given by $e_{R,\mathcal{M}} = \arg\max_{i\in\mathcal{M}} sup_{j\in\mathcal{M}} t_{ij}$, that is the model with the largest t-statistic. The t-statistic test is particularly useful since it does not require the estimation of a large variance-covariance matrix of the forecast errors.

The asymptotic distribution of this test statistic is non-standard, as it depends on the cross-section correlation of the $t'_{ij}s$. In order to address these issues, the MCS procedure uses bootstrap methods to estimate the distribution of the test statistic. So that estimates of the distribution reflect the persistence in the $d_{i,j,t}$, the MCS employs a stationary bootstrap, as proposed by Politis and Romano (1994). In our implementation,

[6]If the null is not rejected in the first round, $\mathcal{M}^* = \mathcal{M}_0$.

the expected size of block depends on the forecast horizon.[7]

The p-values associated with each model in the set $\mathcal{M}$ reflect the sequential nature of the testing. Defining $p(i)$ as the p-value of the individual model i that gets eliminated if the null is rejected, $p_{mcs}(i) = \max_{j \leq i} p(i)$. This definition accounts for the cases where the p-values for individual models decreases relative to earlier models that have been eliminated, but eliminated with a higher p-value than the current model to be eliminated. The intuition is that if the model eliminated earlier in the sequential testing has a relatively high p-value, the next model can not be eliminated with a higher level of confidence, given the relatively lower confidence of the earlier elimination. By convention, the p-value of the last surviving model in the set is defined to be 1.0 because the test that the model is as good as itself can never be rejected. P-values constructed in this way are convenient because, like standard p-values, each $p_{mcs}(i)$ allows one to determine at which level of significance the model would be in the set of best models.

The MCS approach allows us to make statements about the significance level at which models belong to the set of best models that are valid in the frequentist sense. Thus, we are able to answer questions like: if one could observe a large number of realizations of this sample, given a confidence level, what percentage of times would you expect a particular model to be in the set of best models?

We use the MCS to trim models that do not belong to the set of best models as selected by the MCS. To determine which models to use at a given point in time, we use the MCS to test the accuracy of the models in $\mathcal{M}^0$ using a rolling training sample, and let the MCS procedure indicate which models survive the elimination algorithm at

[7]For the 1-quarter ahead forecasts, we use a block size of 2 quarters. For the 2 and 4-quarter ahead forecasts, we use blocksizes of 3 and 6 quarters, respectively.

our chosen level of significance.

When constructing the model confidence set $\hat{\mathcal{M}}^*$ we perform a systematic analysis to choose our baseline confidence level of α and then keep models with an associated p-value greater than or equal to α. We based our choice of α on an analysis of results for a range of α going from from 1 percent to 99 percent, as shown below.

3 Models and Combination Methods

In order to examine the benefits of trimming the number of available models before forecast combination, we apply the trimming methods discussed above to the commonly-used setting of averaging forecasts from many bi-variate models estimated from a panel of macroeconomic data.[8]

The forecasts are based on linear AR models and one additional predictor per model. Let t date the predictors, and y_t be the annualized growth rate from *t-1* to t, of the variable to be forecasted and, x_t, the n x 1 vector of predictors. y_{t+h} is the h-quarter ahead value of the cumulated growth rate to be forecasted, $y_{t+h} = \sum_{i=1}^{h} y_{t+i}/h$. We estimate the models for $h = 1$, 2, and 4. For each individual series $\{x_{i,t}\}_{i=1}^{n}$ in our macroeconomic and financial panel, we estimate the following model for our variable of interest:

$$y_{t+h} = \alpha_i + \sum_{j=0}^{P} \beta_i y_{t-j} + \gamma_i x_{i,t} + \varepsilon_{i,t+h} \tag{1}$$

The single predictor based models described above are estimated with rolling sam-

[8]See for example, Stock and Watson (2004); Faust et al. (2011); Wright (2009) among many others.

ples. Hansen et al. (2011) recommends a rolling window to guard against non-stationarity in the model's loss differentials, which is a requirement when comparing loss functions over time using the model confidence set approach. P is based on the Bayesian Information Criteria (BIC) of the univariate AR, of Y_t on its lags and is calculated for each rolling sample, but is held fixed across each predictor based model i.

We use a in-sample rolling sample size of R quarters to estimate the parameters of the models. We then generate a rolling training sample of S quarters that will subsequently be used to weight the different models and to evaluate which models should be trimmed. Our first forecast combination will be for period $R + S + 1$. As baseline values, we use $R = 40$ and $S = 20$. In an supplemental appendix, we analyze the effects of different choices of training sample.

3.1 Data

The macroeconomic data set of potential predictors we use consists of 126 economic and financial variables. This large panel contains data on aggregate and disaggregate macroeconomic data, surveys, and financial indicators. Table 1 details the series contained in the panel. The panel starts in 1959Q3 and ends at 2010Q4. Table 1 also details the transformations applied to each series to eliminate trends. The panel closely resembles that of Stock and Watson (2002).

We use this dataset to predict measures of economic activity and inflation, namely: Gross Domestic Product (GDP), Nonfarm Payroll (EMP), Industrial Production (IP), Housing Starts (HST) and the Gross Domestic Product Deflator (DEF)[9]. As our base-

[9]We also examined the robustness of the results to other measures of inflation, namely CPI and PCE, and found similar results across all measures.

line exercise, we report results for a forecasting exercise with real-time data for these series. The dataset was obtained at the Real-Time Dataset for the Macroeconomist at the Federal Reserve Bank of Philadelphia. The macroeconomic panel was gleaned from a number of data sources, but mostly from the St. Louis Fed FRED data base.

Unfortunately, there is not a real-time data set of macroeconomic variables that covers all of our predictors over our whole sample to perform the exercise fully in real-time. Nevertheless, as shown by Bernanke and Boivin (2003) and Faust and Wright (2009), the use of real-time or revised data does not affect the relative forecast accuracy of similar forecasting models as the ones in this paper.

3.2 Forecast Combination Methods

After estimating and selecting the individual models, we weight the predictions to produce the final combined forecast. We form these combinations with and without trimming in order to analyze the gains from the trimming methods described above. We perform two commonly used forecast combination techniques to combine the forecasts based on the individual models, as well as on the set of best predictors as chosen by our trimming methods. Here we briefly describe these forecast combination methods.

The methods we use to combine forecasts from individual models are mostly weighted averages of each of the individual forecasts. Let $\hat{y}_{i,t+h}$ denote the i-th individual pseudo out-of-sample forecast, estimated at time of predictor i's availability in time t. The combined forecast is constructed as:

$$\hat{y}_{t+h} = \sum_{i=1}^{m} w_{i,t} \hat{y}_{i,t+h} \tag{2}$$

where $\hat{y}_{t+h}$ is the final combination forecast, $w_{i,t}$ is the weight assigned to each i individual forecast, $\hat{y}_{i,t+h}$ at period t.

3.2.1 Equal Weights

The simplest and often most effective forecast combination method is the simple mean of the panel of forecasts. With this approach,

$$w_{i,t} = 1/M, \tag{3}$$

where M is the total number of models. Hence, all forecasts contribute with an equal constant weight. Stock and Watson (2004) finds that the equal weights combination for output forecasts produce forecasts that beat a series of more elaborate weighting schemes when forecasting output growth in the G7.

3.2.2 Inverse MSE weights

We combine forecasting models by weighting by the inverse of each model's MSPE. By this method, models that have a lower mean squared prediction error get a higher weight in producing the combined forecast. Because we want to consider the out of sample performance of the models, we use a rolling training sample of $S = 20$ quarters to calculate the out of sample MSPEs for the individual predictor based models.[10] The sample gets rolled forward as each additional out of sample forecast is produced. Calling v the number of periods in the rolling training sample, the weight for model i used in forecasting period t is

[10] In a supplemental appendix we analyze the effect of different choices of training samples.

$$w_{i,t} = \frac{MSE^{-1}_{i,(t-1-v,t-1)}}{\sum_{i=1}^{M} MSE^{-1}_{i,(t-1-v,t-1)}}. \quad (4)$$

The weights will thus be bounded between 0 and 1.

4 Gains from Trimming

Figure 1 and 2 show the ratio of the trimmed forecast combination to the non-trimmed forecast combination's MSPE for the 1-year ahead forecasts for the fixed and MCS trimming methods. The models are combined with equal weights. The figures for the inverse MSPE weights are very similar to the equal weights ones. In a supplemental appendix, we provide the same figures for the 1 and 2-quarters ahead forecasts. A ratio smaller than 1 means that the trimmed forecast combination has a smaller MSPE than the non-trimmed one. As each of the different trimming schemes here examined depend on different choices for their implementation, we conduct a careful analysis in order to map how the results vary with different cutoff options.

As shown by Stock and Watson (2007) and Tulip (2009), the predictable component of these macroeconomic series was significantly reduced during the GM, especially in the case of inflation. Hence, the majority of the fluctuations in these series are mainly driven by idiosyncratic shocks, which cannot be predicted. For this reason,, we also split the results into two sub-samples: a pre-Great Moderation (from 1975Q4 to 1984Q4) and the Great Moderation period (1985Q1 to 2007Q2). We have chosen to end the Great Moderation period slightly before the beginning of the financial crisis, and the Great Recession that followed it. Not surprisingly, the forecasting errors for this period

are dramatically higher than the ones from the GM. Hence, for obvious reasons, we have decided not to merge this period within the Great Moderation sub period. We discuss the evidence for each of the trimming methods below in more detail.

4.1 Gains from Fixed trimming

Figure 1 shows the MSPE of the trimmed forecast combinations relative to the non-trimmed forecast combinations using the equal weights combination scheme, so a ratio below 1 indicates the trimmed forecast outperforms its non trimmed counterpart. We start by trimming all but the 2% best performing models. All of the ratios converge to one when all models are kept. The figures for each of the indicators clearly show that very aggressive trimming is required to improve forecasts over the simple average combination scheme. In order to achieve sizable gains from trimming, when choosing an exogenous fixed proportion of models to be trimmed, one needs to trim a large share of the models. For all the variables forecasted in this paper, a fixed rule that trims around 90% of the models, and hence combines fewer than 10% of the models, provides the most accurate forecast combination. With this level of trimming, one can achieve sizable reductions of around 25% in MSPE over combining all models predictions. As we discussed above, other papers have also found that aggressive trimming rules tend to be superior. Favero and Aiolfi (2005) advocates trimming 80% of the models when forecasting stock returns. When forecasting inflation in Norway, Bjørnland et al. (2011) finds that trimming 95% of the worst models generates better combinations than trimming a more modest, but still sizable, 50% of the models. Our results corroborate these previous findings. Based on this analysis, as our benchmark we choose a fixed trimming

rule that keeps only the top 10% of models when we compare this approach to other data-rich forecasting alternatives.

The most aggressive rule trims 98% of the models, combining only the remaining 2%, or the top 3 models only, given the size of our dataset. This most aggressive level trimming (98%) does not appear to be optimal for most of the variables we forecast, indicating that using very few models to forecast usually produces subpar results.

The figures show that for all variables the gains from trimming differ markedly across the forecasting periods. For all forecasting horizons, gains from trimming prior to the GM are substantially higher than during the GM. Again, this is consistent with the forecasting literature that shows that the predictable component of macroeconomic variables virtually disappeared during the Great Moderation, so most of the gains from trimming occurred before that period.

4.2 Gains from MCS trimming

In this section, we analyze the effectiveness of Hansen et al. (2011) Model Confidence Set approach as a trimming device. When using the MCS to trim the set of worst forecasts, one must choose with what level of confidence (α) one wishes to select the set of best models, and implicitly trim those not in this set. Choosing a low α will result in fewer models being trimmed, whereas a high α induces more models to be trimmed. The intuition for this is that at low α only models that have very low p-values, i.e. those for which there is strong evidence against the null, are rejected from the set of best models. At higher α, more models can potentially be rejected from the set of best models.

Compared to the fixed trimming method, a key advantage of the MCS approach to trimming is that one chooses a confidence level, a concept with clear interpretation in statistical theory. Given the confidence level, if the data is not informative, many models will be included in the set of best models, as the MCS will have difficulty distinguishing between the models and few of them will be trimmed. On the other hand, if the data is informative, the set of best models will be comprised of fewer models. Given the chosen level of confidence, one does not fix the proportion of models to be discarded without any regard to the statistical significance of the forecasts, in contrast to the fixed trimming above.

We provide a complete picture of the gains of trimming with the MCS using p-values varying from 1% to 99%. Figure 2 shows the ratio of MCS-trimmed to non-trimmed forecasts where forecasts are combined by equal weighting. Again, a ratio smaller than one means that trimming leads to better final combinations than averaging the whole set of forecasts. For all variables, with the exception of GDP deflator, there are U shape gains in forecasting accuracy from trimming. With a p-value of 1%, only the very strongly statistically inferior forecasts that get a p-value between zero and 1% are discarded. Hence, differences between the MCS-trimmed and non-trimmed combinations are small, as evidenced by the fact that most of the ratios start approximately at one. As we increase the level of significance required to select a forecast to the set of best forecasts, more forecasts are trimmed and the gains from MCS trimming increase. The highest gains from trimming are achieved with p-values between 30% and 60%. Above this 60% level, increasing the p-value cutoff leads to worse forecasts for all variables we analyze. Based on this we choose a baseline cutoff p-value of 50% for our

comparison between the trimmed forecast combinations and other data-rich forecasting methods.

A difference between the MCS and fixed-trimming is the robustness of the results. Whereas in the latter approach, only when a very high trimming cutoff is selected, thus discarding a large share of the forecasts, are there sizable accuracy gains; for the MCS trimming results are relatively unchanged for a wide range of p-values cutoffs. By taking into account the significance of the statistical differences between the forecasts, one is able to select more carefully which models should be trimmed. Note that it could be the case that a cutoff p-value of 50% might actually trim more or less than 90% of the models, depending on the informativeness of the data.

As with fixed trimming, the forecasting gains from MCS-trimming arise more strongly prior to the Great Moderation period. For p-values in the interval of 20% to 70%, the forecasting gains for the one-year horizon are of the order of 30% during the first subsample for all variables but GDP, where the gains are in excess of 50%. Results for the whole period are more modest, generally around 10% less than the gains from trimming prior to the GM.

4.3 Which models are chosen?

In the previous section, we investigated the gains in forecasting accuracy of restricting the set of models to be combined by either choosing a fixed proportion of models to be discarded prior to combination, or by using the Model Confidence Set to choose the set of best forecasts, hence taking into account the statistical significance of differences between the forecasts. In this section, we highlight the frequency with which the models

are selected. Are some of the models persistently chosen to the set of best forecasts? If so, which are those models? Are the differences in selection rate between the best and worst models large?

Figures 3 and 4 show the proportion of times in our out-of-sample forecasting exercise each of the 126 models is selected to the set of best forecasts with the MCS and the fixed trimming approach, respectively, for the 1-year ahead forecasts. In the MCS approach shown in Figure 4, the models are selected with a p-value of 50%, whereas in the fixed approach we keep the 10% best models. The models (x-axis) are sorted from lowest to highest proportion rates (y-axis). Under both schemes, a subset of models are never selected to the set of best forecasts, followed by a larger group with increasing selection rates. Finally, on the other point of the spectrum, there is a small group of models with significantly higher selection rates. Therefore, there appears to be a significant persistence in the out-of-sample forecasting performance of good and bad models.

Next, we investigate which are the models most frequently selected. Table 6 shows the five most frequently chosen models for the 1-year ahead forecasts with both trimming methods, again using our baseline cutoffs. A general pattern emerges: Models that include housing and interest rate spreads are the most commonly selected for all economic activity variables forecasted. For inflation, the top models are the ones that include measures of employment, housing and economic activity in general.

5 Monte Carlo Evidence

In the previous sections, we showed that there are substantial gains in forecasting accuracy from carefully trimming the set of models based on their past out-of-sample forecasting performance before forecast combination. In order to shed more light into the benefits of our trimming approach, we conducted a monte carlo study along the lines of Inoue and Kilian (2008). We posit that there are N=50 predictors for y_{t+1}. The simulations are based on 500 replications, with $T = 150$.

The data generating process (DGP) for y_{t+1} is given by:

$$y_{t+1} = \beta' x_t + \varepsilon_{t+1} \tag{5}$$

where $\varepsilon_{t+1} \sim$ NID(0,1). We generate 50 random predictors with two methods. The first assumes that the predictors are independent:

$$x_{it} \sim NID(0_{50\times 1}, I_{50})$$

while the second takes into account the factor structure in the data. We assume that the DGP for the X's is driven by three common factors and idiosyncratic shocks:

$$x_{it} = F_{it}\Lambda_i + \epsilon_{it}$$

where Λ is constructed based on a subset of 50 series from the initial panel of predictors, covering real activity, prices, housing and financial variables. F_{it} is generated from a standard normal distribution.

Following Inoue and Kilian (2008), we propose five different scenarios for the slope parameter vector:

Design B1. $\beta = c_1[1, 1, 1, ..., 1]'$

Design B2. $\beta = c_2[50, 49, 48, ..., 1]'$

Design B3. $\beta = c_3[1, 1/2, 1/3, ..., 1/50]'$

Design B4. $\beta = c_5[1_{1\times 10}, 0_{1\times 40}]'$

Design B5. $\beta = c_6[e^{-1}, e^{-2}, e^{-3}, ..., e^{-50}]'$

In design B1, all variables are equally important predictors of y_{t+1}. In such an environment, one would not expect to find any gains from trimming the set of potential predictors based on their past forecasting performance, as all predictors should on average have equal predictive power. In all other designs, the predictive power of the 50 generated predictors are significantly different. In design B4, a small group of variables (ten) have equal importance, whereas the majority of predictors have no importance at all (zero loadings). In designs B2, B3 and B5, we incorporate smooth decays in the relative importance of each x_i. Whereas in design B2 this decay is slow, in design B5 (exponential), it is very fast, meaning that few variables will have relatively high predictive power for y_t, whereas a significant number of predictors will have approximately zero forecasting power. One would expect that the gains from trimming the set of predictors should be particularly large in situations like the ones proxied by design B5. As in Inoue and Kilian (2008), the scaling constant c_1 are chosen such that the the population R^2 of the forecasting models are the same across

all designs. We show the results for a R^2 of 25% and 50%.

After generating the series with each design, we compare the performance of untrimmed forecast combination and the trimmed forecast combinations, both with equal weights. For the fixed trimming, we examine performance with three different cutoff choices: 95, 75 and 50% of the worst performing models are trimmed. Given the higher computational cost of trimming with the model confidence set approach, we restrict ourselves to our baseline cutoff choice of a p-value of 50%.

Tables 2 to 4 show the results for fixed trimming for a cutoff of 90, 75 and 50% of the worst performing models, respectively. The results for the MCS trimming are shown in Table 5. A ratio smaller than one means that the trimmed forecast combination's MSPE is smaller than the untrimmed one. It is clear from the tables that the gains from trimming are larger for designs B3 and B5 for the independent predictors and B3, B4, and B5 for the predictors that are based on the factor structure. Both these designs have a fast monotonic decrease in gains of accuracy as the importance of the predictors relative to the errors (R^2) decreases.

The pattern of results for fixed trimming resembles the one we find in our empirical exercises. The more aggressive strategy, the one that only keeps the 10% best performing models performs better than softer trimming rules. Discarding 50% of the worst performing models leads to negligible gains in forecasting performance with independent predictor and minor gains with predictors based on the factor structure. Finally, it is also noteworthy that trimming with the model confidence set leads to bigger improvements in forecasting performance in the monte carlo setting than with fixed trimming in the case of independent predictors. In the case of predictors based

on the factor structure, keeping the top ten percent of models performs slightly better than the model confidence set approach, though both lead to significant forecasting gains compared using all of the predictions.

The monte carlo indicates that greater gains from trimming are expected when many of the predictors have weak forecasting power, a common situation in economic forecasting. Hence, by assigning weak forecasts a weight of zero, one benefits from a better bias-variance trade-off. The bias from under fitting is more than compensated by the fall in estimation uncertainty.

6 Forecast Combination and Other Data-Rich Forecasts

We have shown in the previous section that, in our application, trimming the set of models can lead to significant improvements in forecast combinations. In this section we address how these trimmed forecast combinations compared to other competing data-rich forecasting methods. We do this comparison by forming model confidence sets over the pool of forecasts based on each trimming and combination schemes and other commonly used data-rich forecasting methods.

We consider two additional data-rich forecasting methods. The first one is the commonly used factor model, as in Stock and Watson (2002). By summarizing the vast amount of information in our panel of series into few common factors, it allows the breadth of information to be incorporated in a linear forecasting model in a parsimonious way. Finally, we also compare the combinations with a Bayesian Model Averaging

approach that has previously been shown to provide competitive forecasts. We briefly provide more information about these methods below.

6.1 Factor Forecasts

Following Stock and Watson (2002) and Forni et al. (2000), among many others, we estimate principal component based forecasts from our panel of macroeconomic series and use the estimated factors in Factor Augmented Autoregressions of the form

$$y_{t+h} = \alpha + \sum_{i=0}^{p} \rho_i y_{t-i} + \sum_{j=1}^{m} \gamma_j f_{jt} + \varepsilon_{t+h} \tag{6}$$

where $\{f_{jt}\}_{j=1}^{m}$ are the first m principal components of our panel of macroeconomic series $\{x_{it}\}_{i=1}^{n}$. Like for the individual predictor models, p is chosen by the BIC. Since there are periods when very few models are chosen to the model confidence set, we restricted ourselves to forecasting with only the first principal component, and fix $m = 1$. We standardize all variables prior to extracting the principal components so that each variable has a zero mean and unit variance. Common factors give us another option of using the vast amount of information in the panel, without expanding the number of parameters in the model beyond feasibility.

6.2 Bayesian Model Averaging

A growing number of papers have shown the usefulness of Bayesian Model Averaging (BMA) at producing macroeconomic forecasts. Koop and Potter (2004) shows that BMA provide appreciable accuracy gains over forecasting with a single model, more so

than factor forecasts. Wright (2008, 2009) show that BMA is a useful alternative for forecasting the exchange rate and inflation in the US, respectively. Finally, Faust et al. (2011) argues that credit spreads are strong predictors of economic activity in the US using a BMA approach.

Consider individual models given by the representation:

$$y_{t+h} = \alpha + \sum_{j=0}^{p} \beta y_{t-j} + \gamma_i x_{i,t} + \varepsilon_{t+h} \tag{7}$$

Every model uses only one variable x_i at a time, has the same number of lags p chosen by the BIC criteria, and $\varepsilon_t \sim N(0, \sigma^2)$. As is common in the literature, we assign a prior that each individual model is equally likely to be the true model. Hence, $P(M_i) = 1/n$, where n is the number of models in the pool. For the model parameters, the assigned priors follow Fernandez et al. (2001) and were used in a forecasting environment by Faust and Wright (2009) and Wright (2008), among others. The priors for β and σ are uninformative and proportional to $1/\sigma$. The prior for γ_i is given by Zellner (1986)'s g-prior, $N(0, \phi\sigma^2(X'X)^{-1})$. The hyperparameter ϕ governs the strength of the prior. Higher values of ϕ are associated with a less dogmatic prior. We use a $\phi = 4$ as our baseline.

The BMA forecast for each individual model i at time t will be given by

$$\tilde{y}^i_{t+h} = \hat{\alpha} + \sum_{j=0}^{P} \hat{\beta} y_{t-j} + \tilde{\gamma}_i x_{i,t} \tag{8}$$

where $\tilde{\gamma}_i = (\frac{\phi}{1+\phi})\hat{\gamma}_i$ represents the posterior mean of γ_i and $\hat{\beta}$ is the OLS estimator of β.

The marginal likelihood of each individual model i is given by

$$P(D|M_i) \propto \left[\frac{1}{1+\phi}\right]^{\frac{-1}{2}} * \left[\frac{1}{1+\phi}SSR_i + \frac{\phi}{1+\phi}SSE_i\right]^{-\frac{(T-P)}{2}} \tag{9}$$

where SSR is the sum of squared residuals from the restricted regression of y on its own lags and SSE is the squared residuals from the unrestricted regression of y on its own lags and x.

The final BMA forecast will be given by

$$\tilde{y}_{t+h} = \sum_{i=1}^{M} P(M_i|D)\tilde{y}^i_{t+h} \tag{10}$$

where $P(M_i|D)$ is the posterior probability of the i-th model. Hence the BMA forecast is the weighted average of individual forecasts, where the weights are given by each model's posterior probability:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j=1}^{n} P(D|M_j)P(M_j)}. \tag{11}$$

Even though the theoretical justification for BMA relies on strictly exogenous regressors and i.i.d errors, conditions not met in our application, the papers cited above show that BMA provides good forecasts in similar setups.

6.3 Inference

In order to compare the performance of the forecast combinations, before and after trimming, to the above mentioned alternative data-rich forecasts, we make use of Hansen

et al. (2011) Model Confidence Set a second time. $\mathcal{M}^0$, the initial set of all models, consists here of 12 different models: (i) four data-rich forecasts without trimming (EW combination, inverse MSE weights combination, factor and BMA forecasts), (ii) the same initial set of forecasts after fixed trimming with a baseline cutoff of 10%[11], and finally, (iii) the same initial set of forecasts after MCS trimming with a p-value cutoff of 50%.[12]

The MCS selects the set of best models by attaching p-values to each these different forecasts. As in the previous section, we use the stationary bootstrap of Politis and Romano (1994) with 10,000 replications when constructing the p-values. The results below give the MSPE and p-value for each of the models included in $\mathcal{M}^0$.

6.4 Results

In the tables below, we show how the forecast combinations compare with the factor and BMA forecasts before and after trimming. We concentrate our analysis in the 1-year ahead forecasts, and provide additional evidence for the 1 and 2-quarter ahead horizons in a supplemental appendix. Table 7 shows the 1-year ahead MSPE as well as MCS p-value for all the 12 models included in $\mathcal{M}^0$. The first table shows results for the full sample, whereas the next two tables split the results between the forecasts prior to the Great Moderation (1974Q4 to 1984Q4) and the Great Moderation (1985Q1 to 2007Q2).[13]

[11] For this comparison, we only keep the 10% best models.

[12] Only models that receive a p-value of 50% or higher associated with the null hypothesis that the model belongs to the set of best models are kept.

[13] While we include the recent great recession period in the full sample results, we exclude the recent recession from the Great Moderation period.

Several results emerge from this exercise. Among the four data-rich forecasts without trimming, we see that BMA has the lowest MSPE for all predicted variables. Wright (2009) also finds that BMA forecasts are generally more accurate for US inflation than simple averaging. This result is quite consistent across subsamples, and especially strong before the GM and for forecasts made during recessions. Nevertheless, the results show that after trimming, there are minimum to no gains in using BMA to forecast the variables here examined. As a matter of fact, in sharp contrast to the other combination methods (EW and inverse MSE) trimming leads to very small, or even losses in forecast accuracy in BMA forecasts. This result is also consistent across subsamples. This can be explained by the fact that the models being trimmed are the ones that receive a very small posterior weight in the final BMA forecast. Hence, trimming those models leads to little or no difference in the final BMA forecast.

There are also substantial accuracy gains from trimming the panel used in the factor forecasts. As previous research by Bai and Ng (2008) has shown, pre-screening the predictors prior to factor estimation and forecast leads to more accurate final factor forecasts. As we were constrained to consider factor forecasts that include only the first common factor, the unconstrained factor forecast is not competitive with the remaining forecast combination methods. Nevertheless, after trimming with either method, the factor forecast estimated with the subset of data included in the set of best models is competitive, and provides the most accurate forecast for industrial production and inflation.

The MCS-trimmed forecast combinations perform very well when compared to this set of forecasts. For most of the variables, the trimmed forecasts combined with either

EW or MSPE weights are the most accurate forecasts given our baseline cutoffs. MCS trimmed and combined with equal weights outperforms the simple average of models for all of the indicators. Also noteworthy is that once the pool of models is trimmed with the MCS, applying weights other than equal weights to the remaining models has very little benefits, if any, for the resulting combined forecast for each of the macro series.

As previous research has shown, and our results in the section above corroborated, the predictable component of macroeconomic variables was significantly diminished during the Great Moderation. As recently argued by Edge and Gürkaynak (2011), this unpredictability is expected when monetary policy is characterized by a strong stabilizing rule, as was the case during that period. Hence, tables 8 and 9 present the model comparisons for these two periods separately. One manifestation of the fall in predictability of the macroeconomic series here examined is the reduction in MSPE dispersion between the 12 different models. Prior to the GM, there was sufficient information in order to distinguish the forecast performance of the various models. Hence, fewer models are included in the set of best forecast at a 10 and 25% confidence level. During the Great Moderation, the differences in MSPE between the best and worst forecasting model was significantly decreased.

6.4.1 Evidence from Recessions and the Great Recession

As we start our empirical exercise in 1974Q4, our forecasts cover multiple recessions, including the recent Great Recession starting in the third quarter for 2007. A large literature has argued for important nonlinearities in the forecasting performance of

models. Periods of recession juxtaposed to stable environments are a primary example of when one would expect nonlinearities in model performance. Chauvet and Potter (2012) find that the performance of a series of GDP forecasting models deteriorates significantly during recessions.

In this section we single out the forecasts made during periods identified as recessions by the National Bureau of Economic Research (NBER), and compare how both trimming methods perform during these periods. Table 10 shows the MSPE of each of the forecasting schemes, as well as the associated p-value of being in the set of best forecasts. As one would expect, the MSPE of these forecasts are considerably higher than the ones from the whole sample. Nevertheless, we again find a substantial improvement in the MSPE of the forecast combinations after trimming with both methods.

7 Conclusion

In this paper, we proposed the use of model confidence sets to form conditional forecast combination strategies. In an environment where the econometrician has access to a large number of models, we have compared the performance of this proposed method to the more common approach of ranking the models and deciding on a fixed fraction to be discarded, without any regard for the statistical significance of differences between the models.

We show that substantial gains in forecast accuracy can be achieved by discarding the worst performing models before combining the forecasts. We argue that the model confidence set approach offers a more robust procedure for selecting the forecasting

models based on their out-of-sample performance. Applying the MCS to a set of forecasts derived from a large dataset of the US economy, we find that models including housing and interest rate spreads are among the most frequently selected predictors for changes in real activity, whereas the most commonly selected models for inflation forecasting also include measures of employment.

We find that the forecasting gains were concentrated mainly during the less stable environment before the Great Moderation. Given the increase in economic volatility since the beginning of the Great Recession, the substantial gains from trimming could reemerge.

References

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

Bai, J. and S. Ng (2008): "Forecasting economic time series using targeted predictors," *Journal of Econometrics*, 146, 304–317.

Bates, J. and C. Granger (1969): "The combination of forecasts," *OR*, 20, 451–468.

Bernanke, B. and J. Boivin (2003): "Monetary policy in a data-rich environment* 1," *Journal of Monetary Economics*, 50, 525–546.

Bjørnland, H., K. Gerdrup, A. Jore, C. Smith, and L. Thorsrud (2011): "Does Forecast Combination Improve Norges Bank Inflation Forecasts?" *Oxford Bulletin of Economics and Statistics*, 74, 163–179.

Capistrán, C., A. Timmermann, and M. Aiolfi (2010): "Forecast Combinations," *Working Papers*.

Chauvet, M. and S. Potter (2012): "Forescasting Output," *Forthcoming, Handbook of Economic Forecasting Vol.2*.

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

Clemen, R. and R. Winkler (1986): "Combining economic forecasts," *Journal of Business & Economic Statistics*, 39–46.

Edge, R. and R. Gürkaynak (2011): "How useful are estimated DSGE model forecasts for central bankers?" *Brookings Papers on Economic Activity*, 2010, 209–244.

Faust, J., S. Gilchrist, J. Wright, and E. Zakrajsek (2011): "Credit Spreads as Predictors of Real-Time Economic Activity: A Bayesian Model-Averaging Approach," Tech. rep., National Bureau of Economic Research.

Faust, J. and J. Wright (2009): "Comparing Greenbook and reduced form forecasts using a large realtime dataset," *Journal of Business and Economic Statistics*, 27, 468–479.

Favero, C. and M. Aiolfi (2005): "Model uncertainty, thick modelling and the predictability of stock returns," *Journal of Forecasting*, 24, 233–254.

Fernandez, C., E. Ley, and M. Steel (2001): "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, 16, 563–576.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The generalized dynamic-factor model: Identification and estimation," *Review of Economics and Statistics*, 82, 540–554.

Hansen, P., A. Lunde, and J. Nason (2011): "The model confidence set," *Econometrica*, 79, 453–497.

Hendry, D. and M. Clements (2004): "Pooling of forecasts," *The Econometrics Journal*, 7, 1–31.

Hsiao, C. and S. Wan (2011): "Is there an optimal forecast combination?" *Working Paper*.

Inoue, A. and L. Kilian (2008): "How useful is bagging in forecasting economic time series? A case study of US consumer price inflation," *Journal of the American Statistical Association*, 103, 511–522.

Koop, G. and S. Potter (2004): "Forecasting in dynamic factor models using Bayesian model averaging," *Econometrics Journal*, 7, 550–565.

Makridakis, S. and R. Winkler (1983): "Averages of forecasts: Some empirical results," *Management Science*, 987–996.

Politis, D. and J. Romano (1994): "The Stationary Bootstrap." *Journal of the American Statistical Association*, 89.

Stock, J. and M. Watson (2002): "Macroeconomic forecasting using diffusion indexes," *Journal of Business and Economic Statistics*, 20, 147–162.

——— (2004): "Combination forecasts of output growth in a seven-country data set," *Journal of Forecasting*, 23, 405–430.

——— (2007): "Why has US inflation become harder to forecast?" *Journal of Money, Credit and Banking*, 39, 3–33.

Timmermann, A. (2006): "Forecast combinations," *Handbook of economic forecasting*, 1, 135–196.

Tulip, P. (2009): "Has the Economy Become More Predictable? Changes in Greenbook Forecast Accuracy," *Journal of Money, Credit and Banking*, 41, 1217–1231.

Wright, J. (2008): "Bayesian model averaging and exchange rate forecasts," *Journal of Econometrics*, 146, 329–341.

——— (2009): "Forecasting US inflation by Bayesian model averaging," *Journal of Forecasting*, 28, 131–144.

Zellner, A. (1986): "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6, 233–243.

Table 1: Variables and Transformations in our Large Dataset

Variable	Transf	Variable	Transf
Moody's AAA Bond Yield	2	Civilian Unemployed: 27 Weeks and over	5
Moody's AAA Bond Spread	2	Civilian Labor force	5
Avg.Hourly earnings: Construction	5	Avg. Duration of Unemployment	5
Avg.Hourly earnings: Manufacturing	5	Exchange Rate: Switzerland	5
Avg.Weekly Hours: Manufacturing	1	Exchange Rate: Japan	5
Avg.Weekly Overtime Hours: Manufacturing	2	Exchange Rate: UK	5
Moody's BAA Bond Yield	2	Exchange Rate: Canada	5
Moody's BAA Bond Spread	2	S&P Price Dividend Ratio	5
ISM Manufacturing PIM Composite Index	1	S&P Earning Price Ratio	5
ISM Manufacturing Employment Index	1	Real Compensation per Hour	5
ISM Manufacturing Inventory Index	1	Corporate Profits after tax	5
ISM Manufacturing New Orders Index	1	Real Disposable personal income	5
ISM Manufacturing Production Index	1	Real Exports	5
ISM Manufacturing Prices Index	1	Real Final Sales Domestic Products	5
Avg.Weekly Hours: Nondurable Goods	1	Real Gross Domestic Product	5
Avg.Weekly Hours: Durable Goods	1	Real Government Expenditures	5
S&P Returns	1	Real Government Expenditures: Federal Government	5
Fama-French Factor: RmRf	1	Real Government Expenditures: State and Local	5
Fama-French Factor: SMB	1	Real Imports	5
Fama-French Factor: HML	1	Real Compensation per hour: Business Sector	5
Fed Funds Rate	2	Unit Labor Cost: Business Sector	5
1-Year Yield	2	Unit Labor Cost: Nonfarm Business	5
5-Year Yield	2	Real Personal Consumption Expenditures: Services	5
10-Year Yield	2	Real Personal Consumption Expenditures: Durables	5
3-Month Treasury Bill	2	Real Personal Consumption Expenditures: Nondurables	5
6-Month Treasury Bill	2	Real Investment: Structures	5
6-Month minus 3-Month Spread	1	Real Investment: Equipment and Softwares	5
1-Year minus 3-Month Spread	1	Real Investment: Nonresidential Structures	5
10-Year minus 3-Month Spread	1	Real Investment: Residential Structures	5
Personal Saving Rate	2	Nonfarm Business: hours all persons	5
Unemployment Rate	2	Nonfarm Buiness: output per hour all persons	5
U Michigan Consumer Expectations	2	Commercial and Industrial Loans	6
Nonborrowed Reserves: Depository Institutions	3	M1 - Money Stock	6
Nonborrowed Reserves: DI + Term Auction Credit	3	M2 - Money Stock	6
Housing Starts	4	St.Louis Adjusted Monetary Base	6
Housing Starts: Midwest	4	Board of Governors Adjusted Monetary Base	6
Housing Starts: Northeast	4	Board of Governors Adjusted Total Reserves	6
Housing Starts: South	4	PPI: All items	6
Housing Starts: West	4	PPI: Crude Materials	6
Nonfarm Payroll	5	PPI: Finished Foods	6
All Employees: Durable Goods	5	PPI: Industrial Commodities	6
All Employess: Manufacturing	5	PPI: Intermediate Materials	6
All Employees: Nondurable Goods	5	CPI: All items	6
All Employees: Services	5	CPI: Core	6
All Employees: Construction	5	PCE: All items	6
All Employees: Government	5	PCE: Excluding Food and Energy	6
All Employees: Mining	5	PCE: Motorvehicles	6
All Employees: Retail Trade	5	PCE: Food	6
All Employees: Wholesale Trade	5	PCE: Furniture	6
All Employees: Finance	5	PCE: Clothing	6
All Employees: Trade, Transp and Utilities	5	PCE: Gas	6
Excess Reserves of Depository Institutions	5	PCE: Nondurables	6
Industrial Production Index	5	PCE: Housing	6
Industrial Production: Business Equipment	5	PCE: Health care	6
Industrial Production: Consumer Goods	5	PCE: Transportation	6
Industrial Production: Durable Consumer Goods	5	PCE: Recreation	6
Industrial Production: Final Goods	5	PCE: Other	6
Industrial Production: Materials	5	Price of Investment: Structures	6
Industrial Production: Nondurable Goods	5	Price of Investment: Equipament	6
Industrial Production: Durable Materials	5	Price of Investment: Residential Structures	6
Industrial Production: Nondurable Materials	5	Price of Exports	6
Civilian Unemployed: 15 Weeks and over	5	Price of Imports	6
Civilian Unemployed: 15 to 26 Weeks	5	Price of Federal Government Expenditures	6
Civilian Unemployed: 5 to 14 Weeks	5	Price of State and Local Government Expenditures	6
Civilian Unemployed: Less than 5 Weeks	5	Gross Domestic Product Deflator	6

Note: This table shows our dataset, as well as the transformation applied to each one of the series: 1-No change, 2-log, 3-1st difference, 4-2nd difference, 5-1st difference of log, 6-2nd difference of logs.

Table 2: Monte Carlo: Fixed Trimming - 10%

	Independent Predictors		Factor-based Predictors	
	R^2=25	R^2=50	R^2=25	R^2=50
B1	0.99	0.98	0.97	0.87
B2	0.99	0.97	0.93	0.87
B3	0.97	0.90	0.88	0.65
B4	0.99	0.96	0.86	0.62
B5	0.96	0.87	0.87	0.60

Note: This table shows the ratio of MSPE of trimmed-combination over untrimmed-combinations with Equal Weights.

Table 3: Monte Carlo: Fixed Trimming - 25%

	Independent Predictors		Factor-based Predictors	
	R^2=25	R^2=50	R^2=25	R^2=50
B1	0.99	0.98	0.98	0.93
B2	0.99	0.98	0.96	0.92
B3	0.98	0.96	0.92	0.79
B4	0.99	0.97	0.91	0.78
B5	0.98	0.95	0.91	0.77

Note: This table shows the ratio of MSPE of trimmed-combination over untrimmed-combinations with Equal Weights.

Table 4: Monte Carlo: Fixed Trimming - 50%

	Independent Predictors		Factor-based Predictors	
	R^2=25	R^2=50	R^2=25	R^2=50
B1	0.99	0.99	0.99	0.98
B2	0.99	0.99	0.98	0.97
B3	0.99	0.98	0.97	0.92
B4	0.99	0.99	0.97	0.92
B5	0.99	0.98	0.97	0.92

Note: This table shows the ratio of MSPE of trimmed-combination over untrimmed-combinations with Equal Weights.

Table 5: Monte Carlo: MCS Trimming - Pvalue=50%

	Independent Predictors		Factor-based Predictors	
	R^2=25	R^2=50	R^2=25	R^2=50
B1	1.00	1.00	0.98	0.89
B2	1.01	0.99	0.97	0.91
B3	0.99	0.91	0.94	0.74
B4	1.00	0.99	0.92	0.72
B5	0.96	0.83	0.93	0.69

Note: This table shows the ratio of MSPE of trimmed-combination over untrimmed-combinations with Equal Weights.

Table 6: Top 5 Models with Highest Selection Frequency: 1-Year Horizon

			10% Best Forecasts Set		
	GDP	IP	EMP	HST	DEF
1	Invest. Res. Structures	S&P returns	All Employees: Services	ISM Price Index	Housing Starts Northeast
2	All Employees: Mining	S&P div price ratio	Invest. Res. Structures	10yr TB3M spread	All Employees: Durables
3	All Employees: Finance	Fama French RmRf	10yr TB3M spread	ISM Inventories	Housing Starts West
4	Hours N.F.B.	Invest. Res. Structures	Real Final Sales	Avg. Weekly Hours	ISM: Employment
5	Housing Starts Midwest	ISM Price Index	NAPM New Orders	1 Year Yield	ISM Production
			Model Confidence Set		
	GDP	IP	EMP	HST	DEF
1	Invest. Res. Structures	Invest. Res. Structures	All Employees: Services	ISM Price Index	Housing Starts West
2	Housing Starts Midwest	AAA spread	10yr TB3M spread	10yr TB3M spread	Housing Starts Northeast
3	AAA spread	ISM Price Index	S&P returns	Housing Starts West	Real Final Sales
4	10yr TB3M spread	10yr TB3M spread	NAPM New Orders	ISM Inventories	NAPM
5	ISM Price Index	Unit Labor Costs N.F.B.	Invest. Res. Structures	All Employees: Govt	ISM Employment

Note: This table shows the models that are selected most frequently for the Model Confidence Set of best forecasts (p-value of 50%) and the fixed trimming set of best forecasts (models in the 10th percentile and above) for the 1-year ahead forecast with a training sample of 20 quarters.

Table 7: Final Model Comparisons: Full Sample 1-Year Horizon

	GDP		IP		EMP		HSTARTS		GDP Deflator	
	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue
Equal Weights	4.50	0.26 **	26.08	0.51 **	2.57	0.19 *	468	0.10 *	1.65	0.50 **
MSE Weights	4.40	0.26 **	25.68	0.51 **	2.52	0.21 *	459	0.10 *	1.63	0.50 **
Factor	5.31	0.17 *	29.92	0.32 **	2.96	0.13 *	546	0.10 *	1.70	0.31 **
BMA	3.39	0.57 **	21.82	1.00 **	2.34	0.54 **	412	0.10 *	1.43	0.50 **
Fixed Trimmed - EW	3.80	0.26 **	22.12	0.75 **	2.02	0.54 **	361	0.22 *	1.43	0.50 **
Fixed Trimmed - MSE	3.58	0.57 **	21.84	0.99 **	2.01	0.54 **	355	1.00 **	1.42	0.50 **
Fixed Trimmed - Factor	4.10	0.26 **	26.35	0.51 **	1.68	1.00 **	366	0.22 *	1.43	0.56 **
Fixed Trimmed - BMA	3.65	0.57 **	21.96	0.98 **	2.35	0.54 **	380	0.22 *	1.41	0.56 **
MCS Trimmed - EW	3.01	0.62 **	22.32	0.75 **	2.19	0.54 **	373	0.22 *	1.33	0.90 **
MCS Trimmed - MSE	3.00	1.00 **	22.29	0.75 **	2.19	0.54 **	370	0.22 *	1.33	0.90 **
MCS Trimmed - Factor	3.04	0.62 **	23.73	0.75 **	2.17	0.54 **	393	0.22 *	1.29	1.00 **
MCS Trimmed - BMA	3.55	0.53 **	22.26	0.75 **	2.40	0.54 **	392	0.22 *	1.34	0.90 **

Note: This table gives MSPEs and p-values for each forecasting scheme under the null that each scheme has the same relative loss. PCA uses the first principal component. * and ** indicate the model is in the set of best models at the 10 and 25-percent level, respectively. Fixed trims the worst 90 percent of models based on their MSPEs, MCS trims models with a p-value of less than 50 percent.

Table 8: Final Model Comparisons: Pre-Great Moderation 1-Year Horizon

	GDP		IP		EMP		HSTARTS		GDP Deflator	
	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue
Equal Weights	9.39	0.06	56.61	0.02	5.89	0.05	969	0.11 *	4.74	0.17 *
MSE Weights	9.09	0.06	55.34	0.02	5.75	0.15	943	0.19 *	4.66	0.17 *
Factor	11.88	0.01	63.59	0.01	6.76	0.03	1314	0.11 *	4.65	0.17 *
BMA	4.76	0.47 **	41.14	1.00 **	4.51	0.54	776	0.23 *	3.85	0.17 *
Fixed Trimmed - EW	7.26	0.06	46.09	0.02	4.22	0.54	666	0.41 **	3.91	0.17 *
Fixed Trimmed - MSE	6.50	0.30 **	44.68	0.61 **	4.15	0.54	652	1.00 **	3.88	0.17 *
Fixed Trimmed - Factor	7.06	0.30 **	51.44	0.02	2.79	1.00	737	0.41 **	3.70	0.17 *
Fixed Trimmed - BMA	5.82	0.38 **	42.87	0.61 **	4.88	0.54	693	0.41 **	3.87	0.17 *
MCS Trimmed - EW	3.88	0.47 **	45.97	0.02	4.34	0.54	688	0.41 **	3.47	0.44 **
MCS Trimmed - MSE	3.85	0.47 **	45.84	0.61 **	4.28	0.54	679	0.41 **	3.46	0.44 **
MCS Trimmed - Factor	3.05	1.00 **	46.92	0.02	3.38	0.54	769	0.41 **	3.15	1.00 **
MCS Trimmed - BMA	5.56	0.30 **	45.10	0.30 **	4.99	0.54	707	0.41 **	3.55	0.43 **

Note: Same as Table 7.

Table 9: Final Model Comparisons: Great Moderation 1-Year Horizon

	GDP		IP		EMP		HSTARTS		GDP Deflator	
	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue
Equal Weights	2.25	0.42 **	10.62	0.56 **	1.06	0.35 **	160	0.05	0.59	0.99 **
MSE Weights	2.24	0.42 **	10.58	0.56 **	1.05	0.35 **	159	0.05	0.59	0.99 **
Factor	2.24	0.42 **	10.45	0.56 **	1.13	0.35 **	178	0.05	0.64	0.80 **
BMA	2.26	0.42 **	9.26	1.00 **	1.10	0.78 **	163	0.05	0.60	0.92 **
Fixed Trimmed - EW	2.02	0.42 **	9.32	0.95 **	0.93	0.79 **	143	0.65 **	0.59	0.99 **
Fixed Trimmed - MSE	1.99	0.91 **	9.31	0.95 **	0.92	0.99 **	140	1.00 **	0.59	0.99 **
Fixed Trimmed - Factor	2.23	0.42 **	11.89	0.35 **	0.91	0.99 **	144	0.65 **	0.64	0.54 **
Fixed Trimmed - BMA	2.31	0.42 **	9.96	0.56 **	1.01	0.78 **	154	0.65 **	0.57	1.00 **
MCS Trimmed - EW	1.94	1.00 **	9.90	0.80 **	0.94	0.79 **	164	0.05	0.58	0.99 **
MCS Trimmed - MSE	1.95	0.91 **	9.89	0.56 **	0.94	0.79 **	164	0.65 **	0.58	0.99 **
MCS Trimmed - Factor	2.13	0.42 **	11.45	0.19 *	0.91	1.00 **	170	0.05	0.60	0.96 **
MCS Trimmed - BMA	2.25	0.42 **	9.71	0.83 **	1.02	0.35 **	175	0.05	0.58	0.99 **

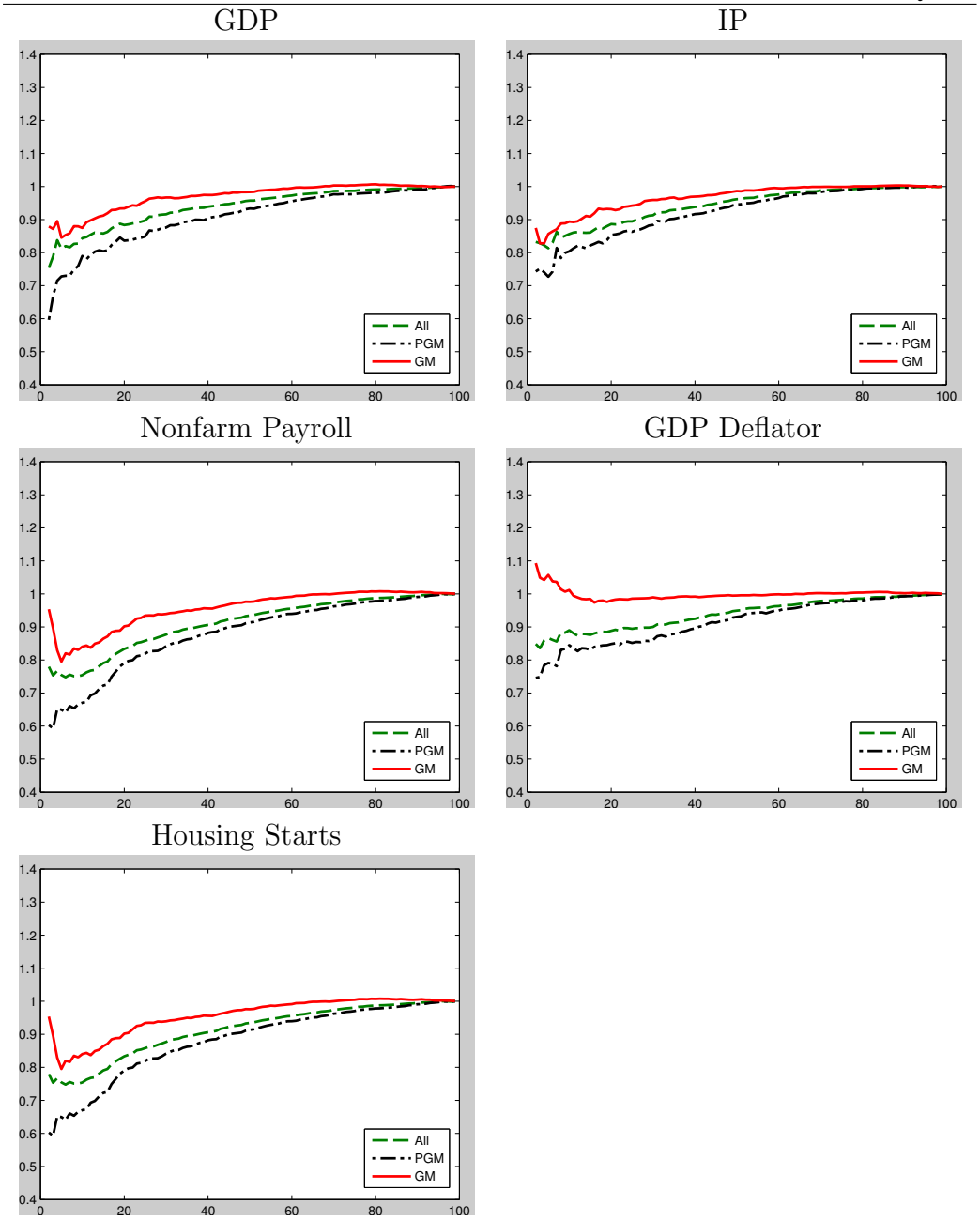
Note: Same as Table 7.

Table 10: Final Model Comparisons: Recessions 1-Year Horizon

	GDP		IP		EMP		HSTARTS		GDP Deflator	
	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue	MSPE	Pvalue
Equal Weights	11.11	0.10 *	71.52	0.13 *	6.97	0.15 *	1361	0.11 *	4.39	0.44 **
MSE Weights	10.87	0.10 *	70.68	0.13 *	6.83	0.15 *	1330	0.11 *	4.26	0.44 **
Factor	13.26	0.10 *	86.81	0.09	8.19	0.06	1480	0.37 **	3.37	0.44 **
BMA	8.22	0.39 **	63.00	0.13 *	7.11	0.15 *	936	0.37 **	2.92	0.44 **
Fixed Trimmed - EW	9.26	0.10 *	64.66	0.13 *	5.48	0.20 *	920	0.37 **	2.98	0.44 **
Fixed Trimmed - MSE	8.78	0.10 *	64.40	0.13 *	5.47	0.19 *	907	0.89 **	2.93	0.44 **
Fixed Trimmed - Factor	8.28	0.10 *	80.04	0.13 *	4.33	1.00 **	881	1.00 **	2.62	0.60 **
Fixed Trimmed - BMA	8.23	0.39 **	60.94	0.92 **	7.09	0.15 *	883	0.97 **	2.84	0.44 **
MCS Trimmed - EW	6.81	0.39 **	59.27	1.00 **	6.22	0.15 *	958	0.37 **	2.46	0.73 **
MCS Trimmed - MSE	6.70	1.00 **	59.32	0.92 **	6.16	0.15 *	950	0.37 **	2.45	0.73 **
MCS Trimmed - Factor	6.95	0.39 **	64.30	0.13 *	5.16	0.29 **	1096	0.37 **	2.21	1.00 **
MCS Trimmed - BMA	7.74	0.39 **	61.92	0.71 **	7.18	0.15 *	958	0.37 **	2.54	0.73 **

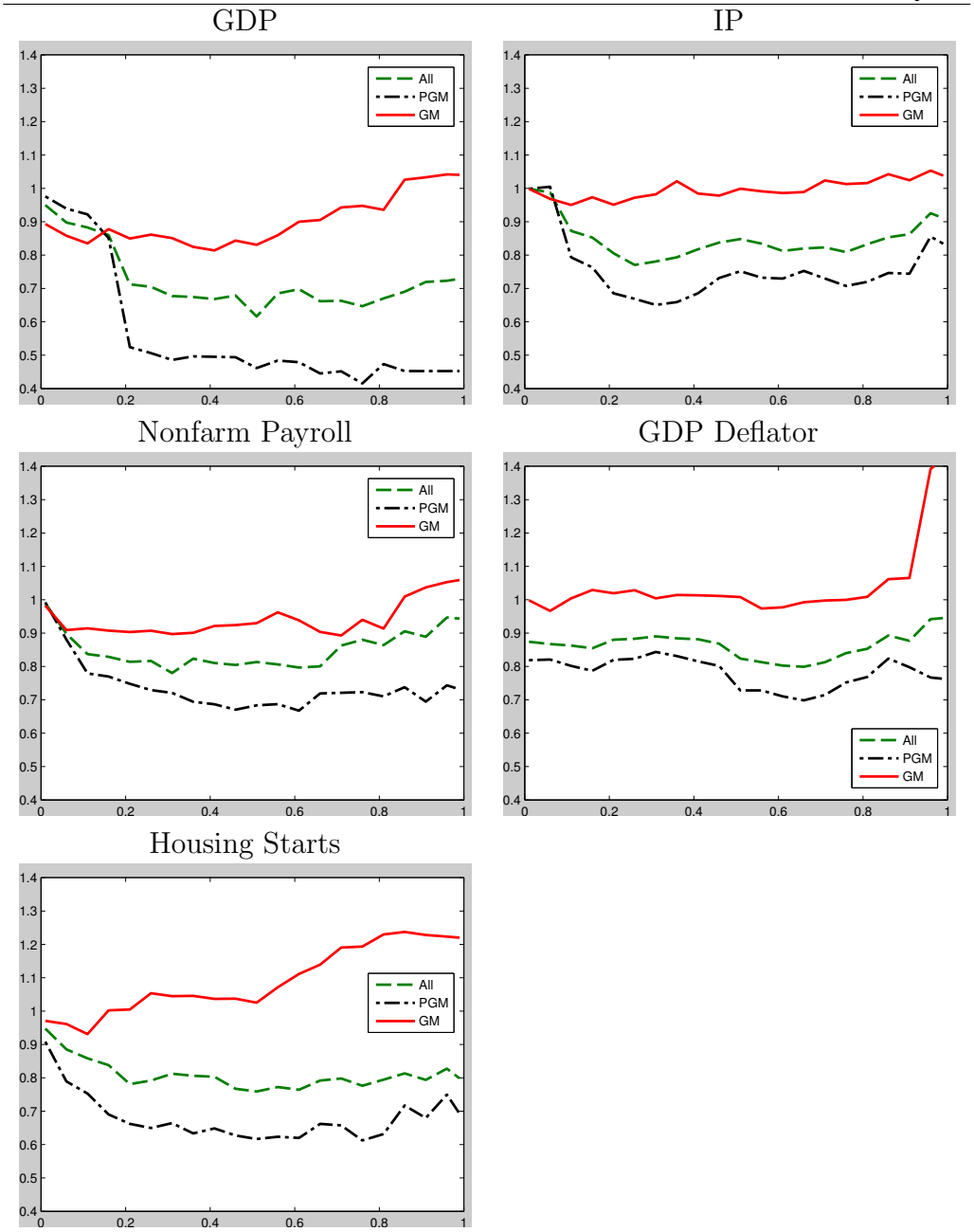
Note: Same as Table 7.

Figure 1: Ratio of Fixed Trimmed to Untrimmed Forecast Combination - 4Q Forecasts



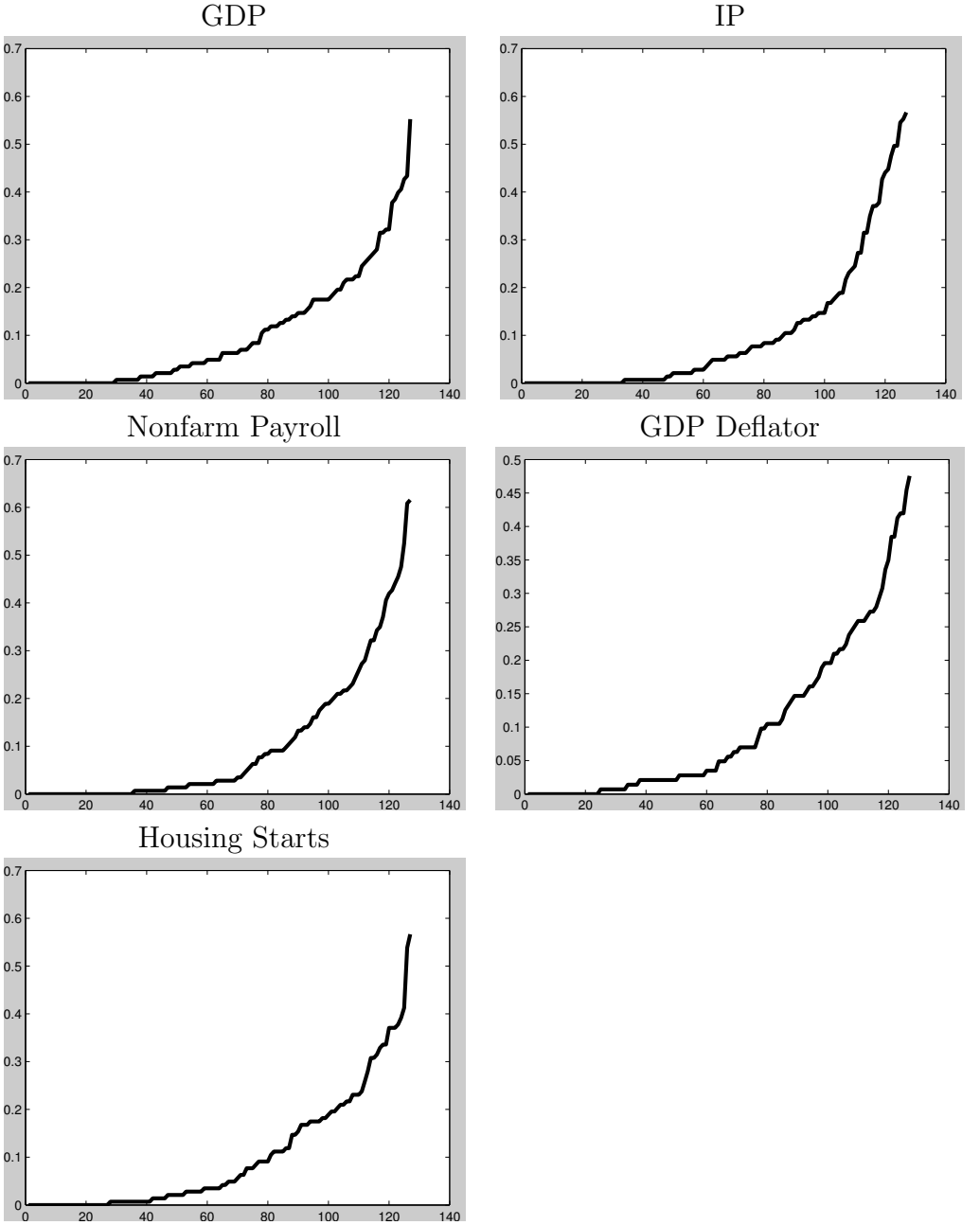
Notes: This figure shows the MSPE ratio of trimmed to non-trimmed forecast combinations with equal weights (y-axis) for different fixed shares of models combined (x-axis). Three sets of results are shown in the graphs: (All) for the whole forecasting period, (PGM) for the period pre-Great Moderation, (GM) for the period of the Great Moderation. A ratio smaller than one means that the trimmed-forecast combination has a smaller MSPE than the combination with the full set of models.

Figure 2: Ratio of MCS Trimmed to Untrimmed Forecast Combination - 4Q Forecasts



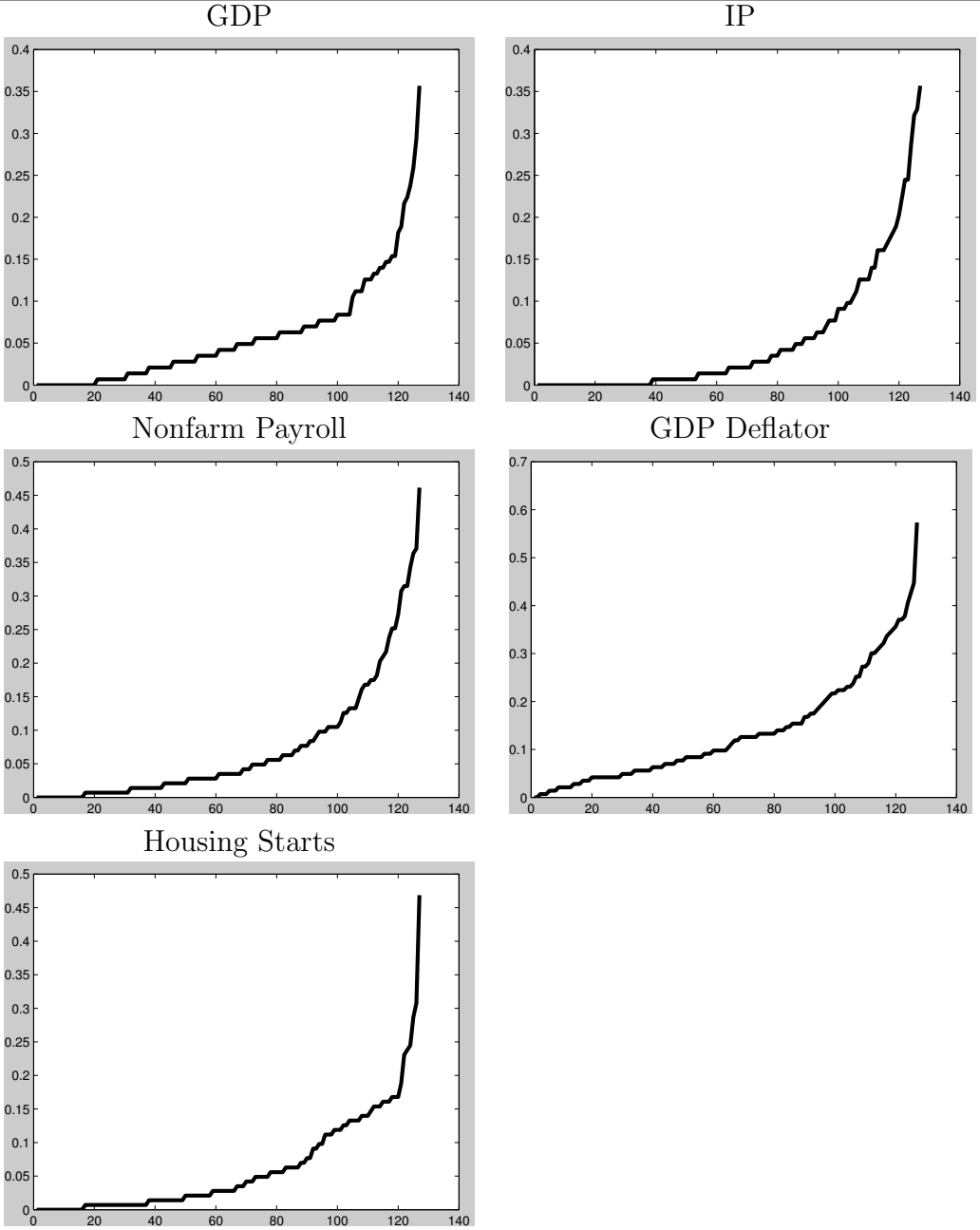
Notes: This figure shows the MSPE ratio of trimmed to non-trimmed forecast combinations with equal weights (y-axis) for different p-values for the MCS (x-axis). Three sets of results are shown in the graphs: (All) for the whole forecasting period, (PGM) for the period pre-Great Moderation, (GM) for the period of the Great Moderation. A ratio smaller than one means that the trimmed-forecast combination has a smaller MSPE than the combination with the full set of models.

Figure 3: Proportion of time each variable is selected for the Fixed set of best forecasts: 1Y Forecasts with practice sample of 5 years



Notes: This figure displays the selection rate of the models to the set of best models selected by the 10% fixed cutoff, sorted from lowest selection rate to highest.

Figure 4: Proportion of time each variable is selected for the MCS set of best forecasts: 1Y Forecasts with practice sample of 5 years



Notes: This figure displays the selection rate of the models to the set of best models selected by the MCS with a pvalue of 50%, sorted from lowest selection rate to highest.