Dynamic Model Averaging in Large Model Spaces*

Luca Onorante† Adrian E. Raftery‡

August 2014

Abstract: This paper proposes a method to perform model averaging in situations in which numerous time series of limited length are available, as typically is the case in macroeconomics. Our procedure allows to perform Dynamic Model Averaging without considering the whole model space but using a subset of models and dynamically optimizing the choice of models at each point in time.

We test the model in an empirical application, nowcasting GDP in the euro area. We show that the forecasting performance is satisfactory and that the results compare well with recent literature and with estimations performed on similar data sets. Several robustness checks confirm the validity of our approach.

Keywords: Dynamic Model Averaging, Big Data, Nowcasting

*The views expressed are those of the authors and do not necessarily reflect those of the Central Bank of Ireland...

†Luca Onorante is Head of the Macro Modelling Project and Deputy Head of Research, Central Bank of Ireland.

‡xxxxxxxxx

1 Introduction

The growing abundance of available macroeconomic data should in theory simplify the quest of econometricians for meaningful empirical relationships. Official statistics and internet-based hard and soft data are growing exponentially, and while main macroeconomic indicators such as GDP and labour market indicators are still released with long lags, an increasing number of macroeconomic indicators is available.

These indicators are often short in time and numerous, and their use poses specific problems. A general survey of the issues when using big data can be found in Varian (2014), but here we focus on two important and unavoidable problems. One is the course of dimensionality. The second is to determine which regressors, among the many available, are important for the relationship or model under investigation.

Model averaging, and in particular Dynamic Model Averaging (DMA), is becoming a popular way to deal with the problems above. Estimation of simple models and averaging the results has proven to be an effective way to use efficiently the available information.

Compared to single model techniques

- Estimation of models is more robust, as less degrees of freedom are used in each estimation.
- The results are easy to intepret, and the importance of each variable or group of variables can be investigated. DMA is therefore a useful device in selecting important regressors.
- When compared to large scale models with the same variables, for example large BVARs, there is no need to tighten the shrinkage when the number of variables increases.

other characteristics make model averaging interesting:

- It can be used to compare and discriminate different and alternative frameworks, for example linear vs non linear or univariate vs multivariate models, fixed coefficients vs time varying parameters.
- It can be used to decide between different regressors (e.g.different measures of slack in a Phillips curve)

In general DMA can be seen as a meta estimation technique where any framework or model can be tested and assessed against the available data.

A growing body of empirical literature confirms that the information retrieved using DMA is useful and generally leads to good forecasts. At the same time, there is a price to pay in terms of computation. Proper DMA requires the

complete exploration of the model space, in other words every possible candidate model must be estimated and evaluated at each point in time. The number of models generally growing exponentially with the potential regressors, there is a limit to the number of variables imposed by computing time. It seems like the "course of dimensionality" may be coming back through the window of unacceptably long computation times.

This paper proposes a variation of DMA particularly adapted to macroeconomic studies and allowing the inclusion of big information sets. Our proposal allows to run DMA without an exhaustive exploration of the space of models, and is ispired on the principle of the Occam window. We further propose an application to the difficult problem of nowcasting GDP in the euro area.

The paper is organized as follows. Section 2 briefly reviews Bayesian and Dynamic Model Averaging and Model Selection. In section 3 we explain the principle underlyuing the Occam window and propose and discuss an algorithm to implement it. In section 4 we propose an economic application, the nowcasting of the euro area GDP, and show in section 5 that the results of our technique compare well with the existing literature despite its lighter approach in terms of computation. Section 6 is dedicated to robustness checks, and the final section concludes.

2 Forecasting with Dynamic Model Averaging

A general discussion of Dynamic Model Averaging (DMA) can be found in the seminal paper by Raftery (2010) and in Koop et. (XXXX). Here we just outline the main concept.

Assume a population M of $m_1, .., m_K$ candidate regression models, each taking the form:

$$y_t = x_t^{(k)} \beta_t^{(k)} + \varepsilon_t^{(k)}, \tag{1}$$

where $\varepsilon_t^{(k)}$ is $N\left(0, \sigma_t^{2(k)}\right)$.

Each explanatory set $x_t^{(k)}$ contains a subset of the potential explanatory variables x_t. It can be immediately seen that this implies a large number of models; if J is the number of explanatory variables in x_t there are $K = 2^J$ possible regressions involving every possible combination of the J explanatory variables.

DMA (and the closely related DMS) average across models using a recursive updating scheme. At each time two sets of weights are calculated, $w_{t|t,k}$ and $w_{t|t-1,k}$. The latter is the key quantity. It represents the weight of model k in nowcasting y_t, at time t, computed using data available at time $t-1$. The first is the update of $w_{t|t-1,k}$ using data available at time t. DMA uses forecasts which average over all $k = 1, .., K$ models using $w_{t|t-1,k}$ as weights. Note that DMA is dynamic since these weights can vary over time. DMS is similar but it involves selecting the model with the highest value for $w_{t|t-1,k}$ and using it for

forecasting y_t. A peculiarity of DMS is that it allows for model switching: at each point in time it is possible that a different model is chosen for forecasting.

Raftery et al (2010) derive the following updating equation:

$$w_{t|t,k} = \frac{w_{t|t-1,k} L_k\left(y_t|y_{1:t-1}\right)}{\sum_{l=1}^{K} w_{t|t-1,l} L_l\left(y_t|y_{1:t-1}\right)} \tag{2}$$

where $L_k\left(y_t|y_{1:t-1}\right)$ is the predictive likelihood, or the predictive density for y_t for model m_k evaluated at the realized value for y_t. The algorithm then produces the weights to be used in the following period by using a forgetting factor, α, normally set to 0.99 following Raftery et al (2010):

$$w_{t|t-1,k} = \frac{w_{t-1|t-1,k}^{\alpha}}{\sum_{l=1}^{K} w_{t-1|t-1,l}^{\alpha}}. \tag{3}$$

Thus, starting with $w_{0|0,k}$ (for which we use the noninformative choice of $w_{0|0,j} = \frac{1}{N}$ for $k = 1, .., K$) we can recursively calculate the key elements of DMA: $w_{t|t,k}$ and $w_{t|t-1,k}$ for $k = 1, .., K$.

3 Occam window explained

When many potential regressors are considered the number of models is too high to be tractable. However, it is a known fact in the model averaging literature that the great majority of models does not really contribute to the forecast, as their weights are zero or very close to zero. These include for example highly misspecified models, which are kept despite their poor performance only to calculate equation (2) and because they might become useful in the future.

We propose an heuristic aiming at eliminating most of these useless models from the computation, while being able to "resurrect" them when needed. Our "Occam window" method is based on two on implicit assumptions:

1. We dispose at the initial time of a valid population of models
2. Models do not change too fast over time: the relevant models at each time are close enough (in a "neighborhood" opportunely defined) to those of the preceding time.

We believe this assumption is reasonable in macroeconomic analysis.
`Some reference on how and when it's used? Adrian? In other disciplines?`

This assumption, if verified, allows the exploration of the space of models in a parsimonious and efficient way.

3.1 Forecast, Expand, Assess, Reduce: the FEAR algorithm

Under the assumptions above, we propose to implement the Occam window on currently used models and keep for future use only those that perform sufficiently well relative to the best performer. Call the current set of models M and their predictive likelihood (or any other chosen performance indicator) L_m. After choosing a threshold C, we keep for future use the models $m \in M$ such that they pass the Occam window:

$$m : m \in M, L_m \geq C * max(L_m)) \tag{4}$$

The FEAR algorithm iterates four steps: Forecasting, Expanding the set of models, Assessing them, and Reducing the model set via the Occam window.

Initialization

1. Divide the sample $1..T$ in a training sample $1..T_r$ and a forecasting/evaluation sample $T_{r+1}..T$
2. Start with a random population of models $M_0(T_r)$ and an arbitrary set of weights $W_0(t_r)$

For every $t = T_r + 1..T$

1. (Forecast) Use the models in $M_0(t-1)$ and the weights $W_0(t-1)$ to perform Model Averaging, e.g. as in Raftery (2010), obtaining the forecast $Pr(y_t \mid M_0(t-1), W_0(t-1))$
2. (Expand) Expand $M_0(t-1)$ into a larger population $M_1(t)$ including all $m \in M_0(t-1)$ and all their neighboring models (all models derived from any model $m \in M_0(t-1)$ by adding or subtracting a regressor)
3. (Assess) Upon observing y_t compute for all $m \in M_1(t)$ the predictive likelihood $L_m(t) = Pr(m \mid y_{1..t})$
4. (Reduce) Let $L_{max} = max_m(L_m(t))$. Define, for an arbitrary Occam threshold C, the final population $M_0(t)$ (and initial for next period) as $M_0(t) = (m : m \in M_1(t), L_m(t) \geq C * L_{max})$

End for

3.2 Computational issues

This section explains why the Occam window approach allows the exploration of very vast models spaces that would not be possible otherwise.

We define, rather imprecisely but as a rough reference, a Notional Unit of Computation (NUC) as a basic operation of estimation. Since we are concentrating on computability, we consider broadly equivalent (one NUC) one OLS

estimation, one period estimation of a Kalman filter and in general each operation involving at least a matrix inversion. On this loosely defined but quite general metric we compare the Occam method with a DMA exhaustively exploring the space of models. Let J = number of candidate explanatory variables, T = length of data in time, and N = population of models in the Occam window (a subset of the K possible regression models).

DMA with all models estimates

$$NUC_{DMA} = 2^J * T \tag{5}$$

Because all the potential models need to be estimated once per period.

The Occam method reduces the number of models to be evaluated but changes the population dynamically. It is therefore necessary to re-estimate each model from the beginning each time, therefore it needs to estimate

$$NUC_{OCC} = \frac{(T+1) * T}{2} * N \tag{6}$$

different models. The role of the number of models N is extensively explored in section 6.

The Occam window allows gains in speed when $NUC_{OCC} < NUC_{DMA}$, or

$$N < 2 * \frac{2^J}{(nT+1)} \tag{7}$$

In our test case, N varies, $T = 45$, $J = 25$, then

$$\frac{NUC_{OCC}}{NUC_{DMA}} = \frac{10.350.000}{1.509.949.440} = 0.68\% \tag{8}$$

therefore the Occam window isabout 150 times faster. We represent graphically the relationships (5) and (6) in Figure 1:

Figure 1 shows the number of NUC (vertical axis) against length and mumber of regressors. The blue area refers to the Occam window, the red area to DMA. The computational complexity for the Occam window grows quadratically with the length of the available series T, the one of DMA grows only linearly in T but increases exponentially in the number of regressors J. Above 15-20 regressors the Occam window is always more convenient. This is particularly true when the series are relatively short in time, since long series imply a higher number of recomputation for each model in the case of the Occam window.

4 An economic application: GDP in the euro area

Short term forecasting and nowcasting economic conditions is important for policy makers, investors and economic agents in general. Given the lags in compiling and releasing key macroeconomic variables, it is not surprising that a

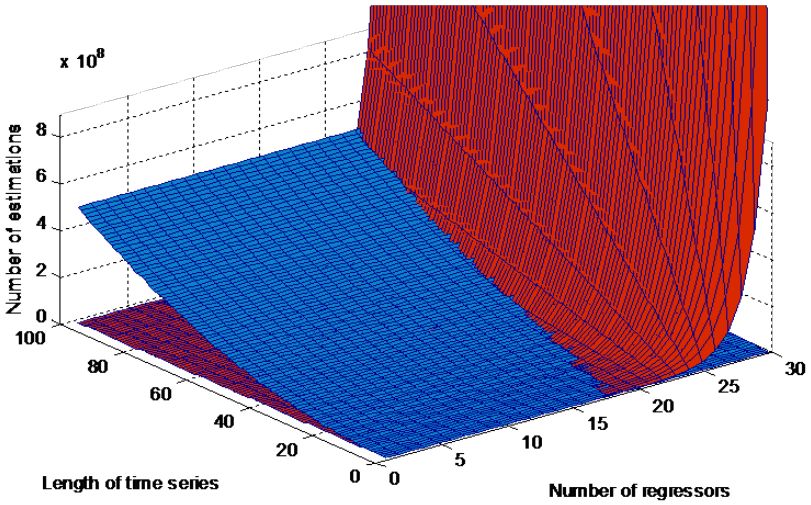

Figure 1: Computing time

particular attention is paid to nowcasting, an activity of particular importance because it allows to set economic decisions and policy actions with a more precise idea of the current situation.

We apply the Occam window method to the nowcast of GDP in the euro area. This problem is particularly difficult because there are many candidate explanatory variables (large J) but most of them cover a short time span (small T). We use quarterly (or converted to quarterly) series available from 1997, and we describe our source data in the appendix `ADD TABLE IN THE END`. Abstracting from minor differences in publication dates, there are two main GDP nowcasts that a forecaster may perform, depending on whether the preceding quarter figure for GDP is available or not. We focus for simplicity of exposition on the case when the past quarter is already available. Our nowcasts will be based on an information set comprising past GDP and current indicators.

The need to use timely indicators largely dictates the choice of potential regressors, but most sectors and economic concepts are well covered. Our indicators include domestic prices (HICP, HICP excluding food an energy and producer prices), cycle indicators (unemployment rate, industrial production, lags of GDP), expectations (mean and dispersion of 2 years-ahead SFP forecasts for GDP, PMI for employment, orders and output), prices of commodities (oil prices, non-energy commodity prices), exchange rates (nominal effective exchange rate, EUR/USD exchange rate), monetary policy variables (short and long interest rates, M3), financial variables (spread between interest rate on bonds of AAA states and average interest rate on bonds, Dow Jones Eurostoxx index, domestic credit). Given the relevant role of uncertainty in the macroe-

conomic developments included in our sample, we include potential macroeconomic risk indicators (Composite Indicator of Systemic Stress, Risk Dashboard data on banking, total, global and monetary factors). All variables are in year-on-year growth rate, with the exception of interest rates and indicators. The target variable in our forecasting exercises is the year-on-year GDP growth rate. As a consequence, at least four lags of the independent variable must be included as potential regressors; we use five to account for potential autocorrelation in the residuals. This may be overcautious, but unnecessary lags will be selected away in the model averaging, and the possibility of adding regressors just to throw them away is after all one of the luxuries of using our methodology.

In order to concentrate on the effects of the proposed Occam method, we slightly simplify Raftery (2010) and estimate each model recursively but with fixed parameters. We choose this setup because Koop (quote) has shown that DMA is a good substitute for time varying parameters, and we want to concentrate on the advantages of the Occam method alone in accounting for model changes. Dynamic Model Averaging is performed as in Raftery, using a discount factor set at $\alpha = 0.99$.

The nowcast realized with the Occam Window, along with uncertainty bands, is reported below.

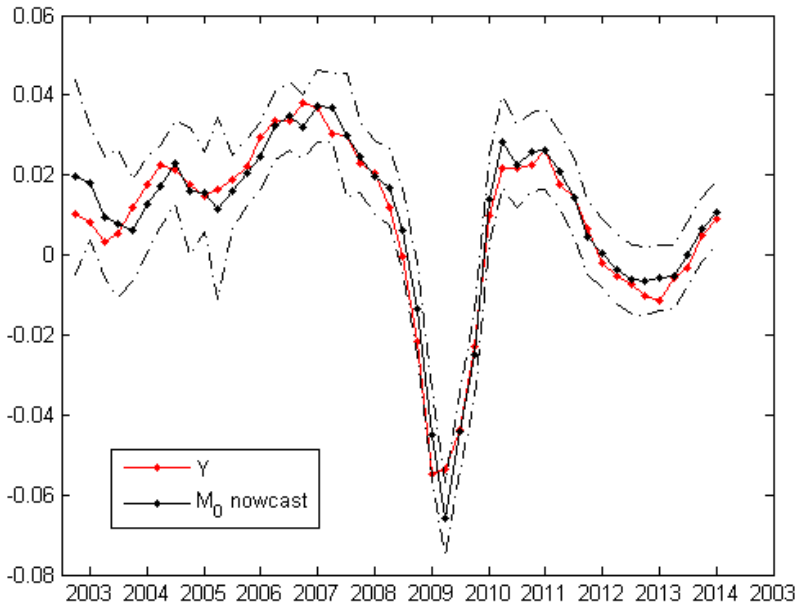

Figure 2: euro area GDP, nowcasting, and uncertainty bands

The chart shows that the DMA with the Occam window has overall a satis-

factory nowcasting performance, even in presence of turning points. The accuracy of the method, as expected, increases with the available data. The 95 per cent error bands take into account the within and between model uncertainty.

The difficult episode of the recession in 2008-2009 is well captured by the DMA. The forecast slightly underpredicts in the trough, but it immediately recovers and becomes quite accurate in the aftermath of the crisis.

The following table compares the forecasting performance in a pseudo-real time exercise. Practically all the indicators we use are seldom or never revised, the main difference with a real time forecasting exercise being the fact that we use the latest available vintage for GDP.[1]

The evaluation sample ranges from 2003q1 until 2014q1.

	RW	$AR2$	$DMA-R$	$DMA-E$	$DMS-R$	$DMA-E$
RMSE	0.0101	0.0088	0.0043	0.0043	0.0048	0.0048
MAE	0.0067	0.0059	0.0033	0.0033	0.0035	0.0035
MAX	0.0332	0.0376	0.0125	0.0126	0.0139	0.0139

Table 1: Forecasting performance

The baseline forecast using MAm_0 compares very favourably with random walk benchmarks. It largely beats both the simple random walk and a standard $AR(2)$. We remind that the forecast MAm_0 is based on past GDP and recent information on the indicator variables.

Forecasts computed using the wider population M_1 are reported for robustness. The results are equivalent to MAm_0. When there are differences in the assessment, these are not sizeable and completely disappear if a sufficient size for population M_0 is allowed. Intuitively, the population M_1 has the advantage of always including all regressors in its models and as a consequence it should react quicker to model changes. On the other hand its forecast is slightly more noisy due to the presence of additional models. The two effects basically cancel out. Each DMA beats the corresponding forecast computed with DMS, although by a small margin, corroborating the common finding that model pooling can beat even the best model in the pool.

We also tried as a further robustness test the DMA of models with time varying parameters. This more solution, more classical in the literature, tends in our case to overreact to the crisis, showing poorer performance. We interpret this result as hinting at the fact that changes in models during the crisis were not due to strong non linearities in the model but to the appreance of new regressors.

Following Koop and Korobilils, we tried additional benchmarks, such as a single TVP model including all regressors, a single B-OLS with all regressors, but these models cannot be estimated or perform very poorly as their estimated paramaters are very unstable.

[1] We could have easily used vintages for the GDP, but we did not see an important value added in this exercise, as forecasters generally try to guess the final numbers.

5 Results, description

This section briefly comments the results of our nowcasting exercise.
The main set of results (and an important value added of using Model Averaging) concerns the inclusion probabilities of each regressors and their evolution. DMA identifies the importance of single variables and how this varies over time, an advantage in terms of easy interpretation and storytelling. Inclusion probabilities of a variable are calculated summing up the weights of the models the use that variable as a regressors. They vary as a consequence between 0 and 1 and give a measure of the importance of that regressor.

Their evolution in time is summarized in the spy plot below and detailed in appendix.

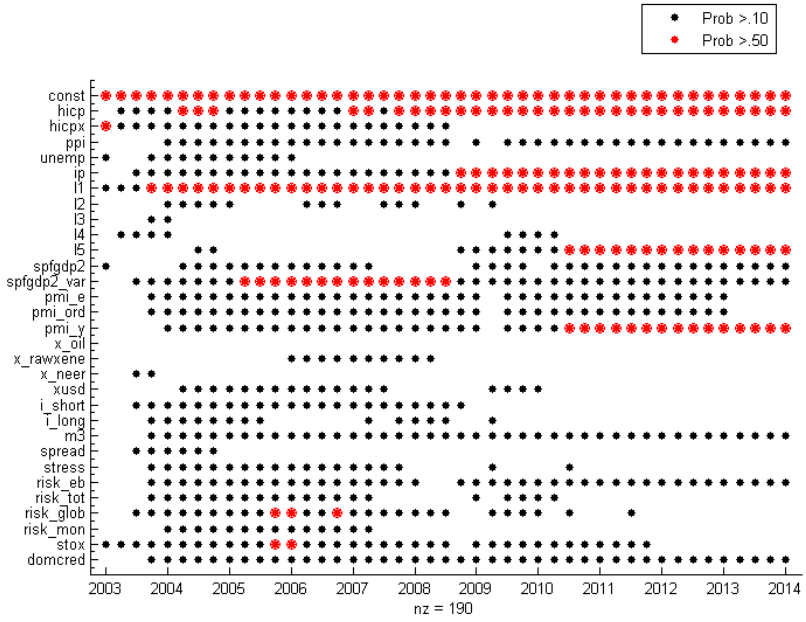

Figure 3: Inclusion probabilities of variables over time: (black) above 10%, (red) above 50%

Overall, and despite being the result of an automatic procedure, inclusion probabilities tell us a coherent story and identify in a small group of variables the most useful indicators of real activity. In greater detail:

- lags of GDP are, as expected, overall important. The first lag captures the

persistence in GDP, and it remains important even during the crisis, when GDP shows pronounced swings. The fourth and fifth lag capture essentially base effects. Our careful approach, including lag 5 in the potential regressors, turns out to be justified.

- Among the consumer price variables, HICP is an important regressors over the whole sample. This confirms the idea that prices and output are not determined in insulation. Without extenfing our interpretation to the existence of a european Phillips curve, we notice that these results confirms the results for the euro are recently obtained by GLMO (2014). Furthermore, the DMA emphasizes the role of producer prices as a forward-looking indicator for nowcasting GDP.

- Among the early indicators of real activity, industrial production is the most important. This is a well known result in nowcasting, where industrial production is widely used as a timley and already comprehensive subset of GDP. The role of unemployment changes over time, becoming less important in the aftermath of the crisis.

- DMA selects almost all GDP surveys as important over the sample, with the exception of the period immediately following the 2008 crisis, which the surveys fail to capture adequately. This results confirms the recent nowcasting literature `LUCA SEARCH FOR REFERNCE` arguing that surveys have a relevant nowcasting power, thereby supporting the importance of expectations in determining macroeconomic outcomes.

- No single external variable, alone, has a determinant role. This is possibly due to the relative closeness of the euro area. Even variables traditionally important in determining prices, such as oil prices or the exchange rate, appear to have a limited impact on real GDP. We find this result interesting but not surprising, given that these variables affect mostly prices and only indirectly GDP.

- Among the variables closer to the operation of monetary policy, interest rates loose progressively importance in the credit constrained post-crisis, while the monetary variable M3 has an increasing role, possibly highlighting the importance of liquidity in the recent part of the sample.

When compared with a similar work (KO) on inflation[2], it is apparent that the determinants of GDP growth are somehow similar. The differences are largely intuitive: while cycle variables influence consumer prices, GDP is also well forecasted with producer prices, determined in advance; international prices do not have the same importance on the cycle as they have on inflation. On the other hand, as in KO, variables representing expectations are important predictors, with the exception of the crisis period.

The complete set of inclusion probabilities is reported in appendix.

[2]KO use DMA on a very similar dataset, but they explore the whole space of models, and this limits the number of regressors they can use.

A natural complement to the results above is the average size of the coefficient of each variable. While inclusion probabilities provide important information about which variables should be included in the regressions at each point in time (as a significance test would do), they do not specify the size of their effect, and even a variable with a very high inclusion probability may have a small overall impact on GDP. The coefficients are averages over models at each point in time, and vary of course within the sample. A discussion about each coefficient is out of the scope of the paper; for completeness, a chart with all the estimated coefficients is reported in appendix.

6 Robustness of the Occam method

6.1 Initial conditions

The Occam method assumes that at the initial estimation time the correct population of models is known. This is not true at the beginning of the sample, where models are estimated on less data and the forecaster does not start with a valid pool of models to draw from. Our implementation of the Occam method for example is based on an initial M_0 population consisting of a unique model, the constant. In this section we check the robustness of the forecast to the choice of the initial population of models.

The chart below reports the average number of variables included in the model. Same size of models is a necessary but not sufficient statistic for convergence in populations of models, but it allows an easy graphic exposition of the issue.

It is immediately clear that the Occam window favours models with about 8-10 variables, and that the initial population M_0 is little representative. The spy plot supports this finding by showing that the inclusion probabilities change rapidly at the beginning of the sample.We test the importance of the initial conditions by using 1) a random population of nN models, and 2) an initial population m0 equal to the final population m0(T).

	MAm_1	MAm_0	rw	MSm_1	MSm_0	$AR2$
RMSE	0.0043	0.0043	0.0101	0.0047	0.0047	0.0088
MAE	0.0032	0.0032	0.0067	0.0035	0.0035	0.0059
MAX	0.0127	0.0127	0.0332	0.0139	0.0139	0.0376

Table 2: Forecasting results starting from an initial population of models of average size 4

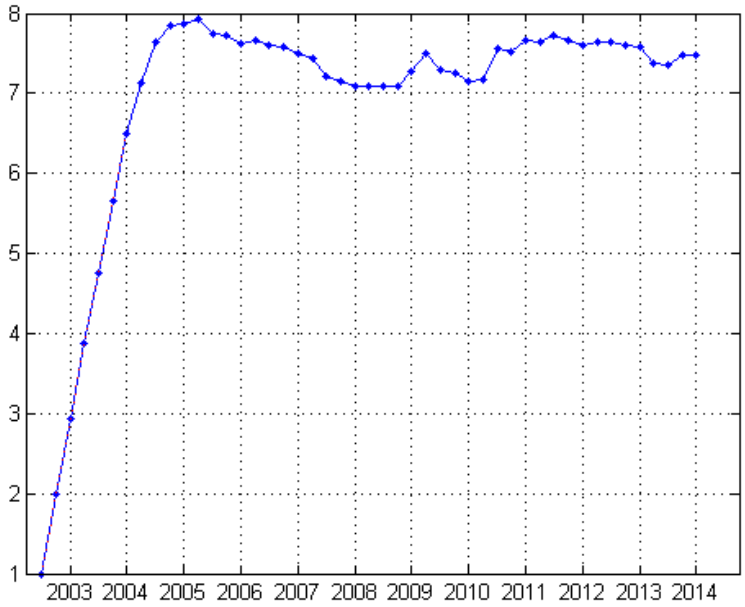

Figure 4: Chart of forecasting results starting from different intial assumptions: one model, an initial population of models of average size 9, or the final population of models

6.2 Maximum number of models

A pure application of the principle of the Occam window and of the FEAR algorithm would keep each model passing condition (4). This would soon lead to a relatively high of models in the wider population M_1, as this population, generated from the E-step of the algorithm, includes all possible neighbours of the preceding population M_0. The latter, however, is comparatively well contained by the following R-step, where condition (4) is applied. Figure 4 shows the evolution of the size of population M_0 over time.

The effort of the algorithm to find a stable population of models at the beginning is reflected in the high number of models retained. It is important to remind that we start our evaluation sample after ten data points, therefore many models are very poorly estimated and their performance varies wildly. After a few periods, a stable population has been found and it is progressivley refined, therefore the size of M_0 decreases.
Once starting from a valid population, as in one of our hypothesis, the FEAR algorithm increases the population site only during turbulent times, for example in 2008-2009. The algorithm automatically increases the population M_0 because the forecast is less accurate and no model is clearly dominating. This leads the

	MAm_1	MAm_0	rw	MSm_1	MSm_0	$AR2$
RMSE	0.0043	0.0042	0.0101	0.0047	0.0047	0.0088
MAE	0.0031	0.0031	0.0067	0.0035	0.0035	0.0059
MAX	0.0127	0.0127	0.0332	0.0139	0.0139	0.0376

Table 3: Forecasting results starting from the final estimated population

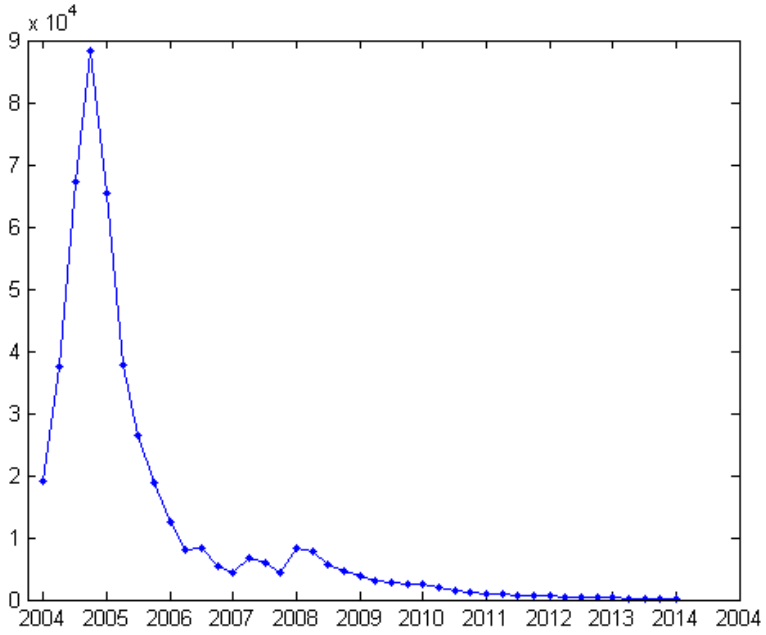

Figure 5: Occam window and number of models over time: M_0

FEAR algorithm to "resuscitate" additional models in the attempt to improve forecast. The preceding figure shows that this attempt is susally successful. Quiet periods are instead characterized by smaller, stable model populations.

Finally, as the sample size increases and models including the best regressors are selected the necessary population site becomes quite small (the last M_0 has size 186). Overall, population M_0, from which the baseline nowcast is generated, never exceeds 10000 models, the wider M_1 can be up to about ten times larger.

In the interest of speed, we introduced the possibility of specifying a maximum number of models N, and our last robustness chaek experiments with this number in order to assess whether it implies a deterioration of the forecast.

Figure 5 reports the nowcasting performance (as measured by the RMSE)

in relation to maximum model size N.

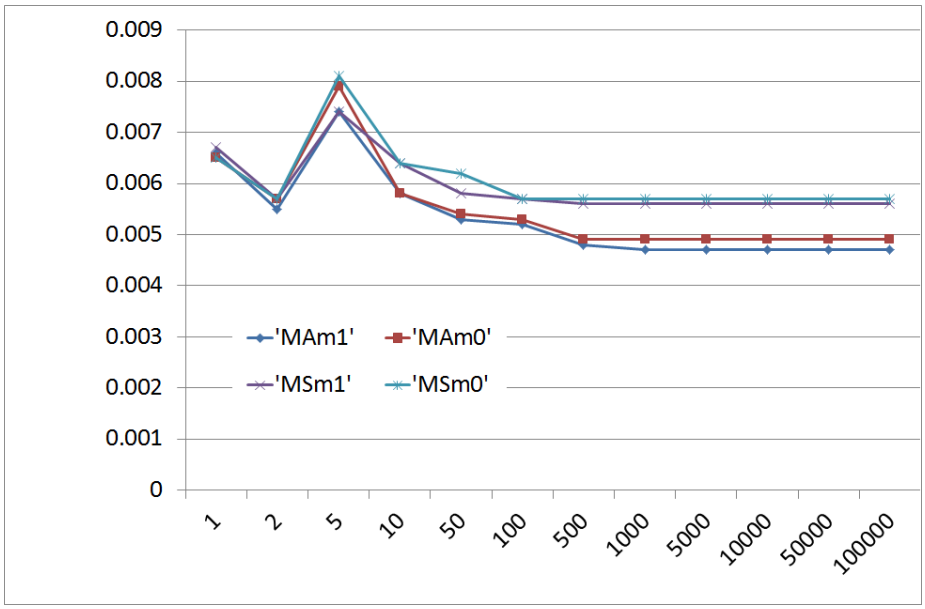

Figure 6: Number of models and RMSE of nowcasting

In our model space of 27 potential regressors the forecasting performance improves until about 10000 models in population M_0. Bigger model populations, as we have seen from the unconstrained estimation, do not lead to any further improvement, and constraints set at 50000 or above on the total population are not binding and thus exactly equivalent to the Occam window without a maximum number of models. We would of course still recommend to keep the maximum number of models as high as possible. The picture also confirms that in our case DMA performs slightly better than DMS for any population size. This is a robust result in the case of macroeconomic variables, but it cannot be generalized. Koop and others (***) and Morgan et al (***), for example, have shown using Google searches as predictors that DMS performs better in contexts where the data are noisy and forecasting benefits from excluding many of them.

When looking at specific parameter values we observe that convergence may be slower for those parameters characterized by low inclusion probabilities. For some specific parameters and inclusion probabilities there are observable convergence issues up to 50.000 models. When this more specific information is important, we would suggest increasing the maximum number of models by one (or if possible two) order of magnitude.

7 Conclusions

This paper proposes an innovative method to perform model averaging in presence of very large model spaces. This method, based on the Occam window, is particularly efficient in situations in which numerous time series of limited length are available, as typically is the case in macroeconomics. Our procedure allows to perform Dynamic Model Averaging without considering the whole model space but using a subset of models and dynamically optimizing the choice of models at each point in time.

After explaining the principle of the Occam window and outlining an implementation algorithm, we test the model in an empirical application, nowcasting GDP in the euro area. We show that the forecasting performance is satisfactory compared to common benchmarks and that the results compare well with recent literature and with estimations performed on similar data sets. Several robustness checks confirm the validity of our approach.

References

Raftery 2010

Koop and Korobilis

Artola, C. and Galan, E. (2012). "Tracking the future on the web: Construction of leading indicators using internet searches," Documentos Ocasionales No. 1203, Bank

of Spain.

Appendix

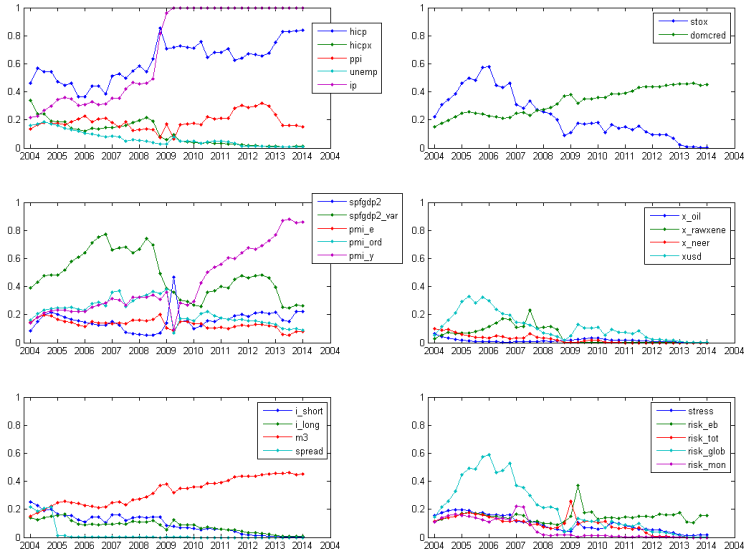

Figure 7: Inclusion probabilities of single variables over time

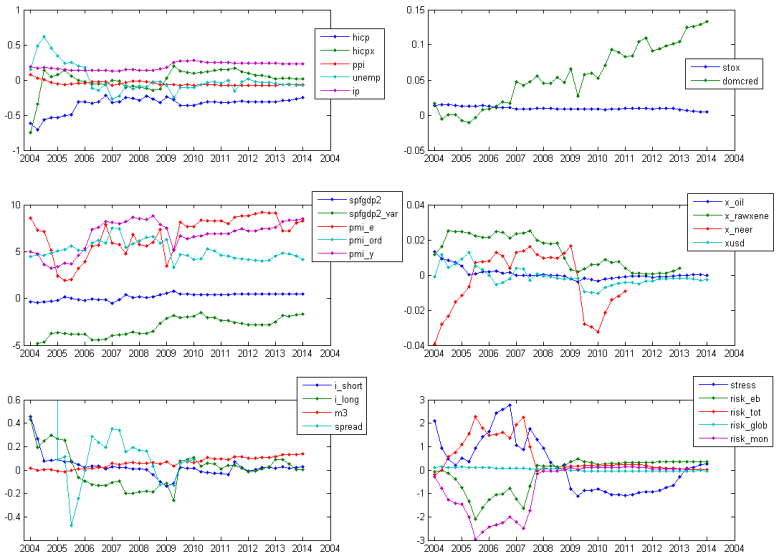

Figure 8: Average coefficients of single variables over time