European Economic Review 81 (2016) 2–14

Contents lists available at ScienceDirect

European Economic Review

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

Dynamic model averaging in large model spaces using dynamic Occam's window ☆

Luca Onorante [a], Adrian E. Raftery [b,*]

[a] Central Bank of Ireland, Ireland
[b] Departments of Statistics and Sociology, University of Washington, United States

ARTICLE INFO

Available online 10 August 2015

JEL classification:
C5

Keywords:
Bayesian model averaging
Model uncertainty
Nowcasting
Occam's window
Econometric Modeling

ABSTRACT

Bayesian model averaging has become a widely used approach to accounting for uncertainty about the structural form of the model generating the data. When data arrive sequentially and the generating model can change over time, Dynamic Model Averaging (DMA) extends model averaging to deal with this situation. Often in macroeconomics, however, many candidate explanatory variables are available and the number of possible models becomes too large for DMA to be applied in its original form. We propose a new method for this situation which allows us to perform DMA without considering the whole model space, but using a subset of models and dynamically optimizing the choice of models at each point in time. This yields a dynamic form of Occam's window. We evaluate the method in the context of the problem of nowcasting GDP in the Euro area. We find that its forecasting performance compares well with that of other methods.

1. Introduction

Short-term forecasting and nowcasting economic conditions are 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 particular attention is paid to nowcasting, an activity of importance because it allows economic decisions to be made and policy actions to be taken with a more precise idea of the current situation.

Cheap computing power and wide availability of data have made forecasts increasingly available over time. At the same time, rich data environments pose new challenges to forecasters. The first and most important is how to distinguish useful variables from noise. Following the financial crisis and the revisions to existing models, the related problem of which regressors should be used at different times has also become crucial in forecasting and modeling, leading to more emphasis on forecast and model uncertainty.

Model averaging, and in particular Bayesian Model Averaging (BMA) is a useful tool to deal with some of these challenges and has consequently become more popular among practitioners (Del Negro et al., 2014). Advantages of BMA include the possibility of using more parsimonious models, which tend to yield more stable estimates, because fewer degrees of freedom are used in individual models. Also, BMA signals relevant regressors, making the results more informative and easier to interpret. It can be used to account for model uncertainty, or as a tool to choose the best indicator to measure a

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

* Corresponding author. Tel.: +1 206 543 4505.
E-mail address: raftery@u.washington.edu (A.E. Raftery).

http://dx.doi.org/10.1016/j.euroecorev.2015.07.013

concept, for example choosing between different measures of slack in a Phillips curve. Finally, it can also be used to account for uncertainty about model structure beyond variable selection (e.g. linear versus nonlinear models, univariate versus multivariate models, fixed versus time-varying parameters). A version of BMA allowing models to change over time, thus also dealing with structural changes, is Dynamic Model Averaging (DMA). DMA was first proposed by Raftery et al. (2010) and allows the weights used in the model averaging to change over time.

This paper deals with the main limit to the use of DMA in macroeconomic forecasting and nowcasting, namely that the computational requirements still limit the use of DMA to at most middle-sized datasets. The reason for this limitation is simple: in a standard regression with J possible predictors, the number of possible regression models (the model space) amounts to 2^J models. This model space grows rapidly with J and quickly becomes too large to be computationally tractable with DMA.

In this paper we propose a way of implementing DMA in large model spaces, called Dynamic Occam's Window (DOW). This allows us to perform DMA without considering the whole model space, but using a subset of models and dynamically optimizing the choice of models at each point in time. It is particularly well adapted to macroeconomic studies and allows the inclusion of substantial information sets. We describe 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 and puts them into the context of other data-intensive forecasting techniques. In Section 3 we describe the Dynamic Occam's Window method. In Section 4 we describe an economic application, the nowcasting of the euro area GDP, and show that our technique yields good results in terms of forecasting performance and computational efficiency. Section 5 gives the results of several sensitivity analyses and robustness checks, and the final section concludes.

2. Forecasting with dynamic model averaging

This section describes Dynamic Model Averaging (DMA), as it was introduced by Raftery et al. (2010), and briefly compares it with other data-intensive forecasting techniques. While this section is more technical in nature, the rest of the article can be understood without it.

BMA (Leamer, 1978; Raftery, 1988; Madigan and Raftery, 1994; Raftery, 1995; Raftery et al., 1997; Hoeting et al., 1999) is increasingly used for forecasting when there is uncertainty about which forecasting model should be used. Bayesian Model Averaging (BMA) deals with the issue by assigning prior probabilities to each model under consideration and by updating these probabilities using Bayes' Theorem and the observed data. Predictive distributions for future observations are constructed as the weighted average (using the probabilities as weights) of the predictive distributions of the individual models. It has become quite popular in economics, thanks in large part to methodological developments to which Eduardo Ley has contributed (Fernández et al., 2001a; Ley and Steel, 2007, 2009). For other developments of BMA in macroeconomics, see Fernández et al. (2001b), Brock and Durlauf (2001), Brock et al. (2003), Sala-i-Martin et al. (2004), Durlauf et al. (2006, 2008), Eicher et al. (2010) and Varian (2014); see Steel (2011) for a survey.

Dynamic Model Averaging is an extension of BMA to allow both the generating regression model and the regression parameters to vary over time. Following Raftery et al. (2010), we assume a population, M, of K candidate regression models, $M=\{m_1,..,m_K\}$, where model m_k takes the form:

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

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

Each explanatory set $x_t^{(k)}$ contains a subset of the potential explanatory variables x_t (x_t can also include lagged values of y_t). This implies a large number of models: if J is the number of explanatory variables in x_t, then there are $K=2^J$ possible regression models involving every possible combination of the J explanatory variables.

DMA averages across models using a recursive updating scheme. At each time two sets of weights are calculated, $w_{t|t-1,k}$ and $w_{t|t,k}$. The first, $w_{t|t-1,k}$, is the key quantity. It is the weight of model k in forecasting y_t given data available at time $t-1$. The second weight, $w_{t|t,k}$, is the update of $w_{t|t-1,k}$ using data available at time t. DMA produces forecasts which average over all K models using $w_{t|t-1,k}$ as weights. Note that DMA is dynamic since these weights can vary over time.

Dynamic Model Selection (DMS) uses the same model weights as DMA, but when it comes to forecasting y_t it uses only the model with the highest value of $w_{t|t-1,k}$. DMS allows for model switching: at each point in time it is possible that a different model is chosen for forecasting.

Raftery et al. (2010) derived the following updating equation for DMA:

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

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

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

The forgetting factor, α, is specified by the user. Here we use $\alpha = 0.99$, following Raftery et al. (2010). Thus, starting with $w_{0|0,k}$ (for which we use the noninformative choice of $w_{0|0,k} = \frac{1}{K}$ for $k = 1,..,K$), we can recursively calculate the key elements of DMA: $w_{t|t,k}$ and $w_{t|t-1,k}$ for $k = 1,..,K$.

Other techniques use large datasets in order to forecast quarterly GDP growth. Here we review three of them, to put DMA into context. This paper is closely related to the optimal predictive pool proposed by Geweke and Amisano (2010, 2011). More recently, Del Negro et al. (2014) considered dynamic prediction pools and developed a methodology for estimating time-varying weights in optimal prediction pools. This technique presents the important advantage that it does not operate under the assumption that one of the models considered is correct. It is, however, computationally intensive and so can be applied only to a limited number of models.

Dynamic factor models are based on the idea that a few unobserved common factors are able to capture the essential information in a large cross-section of time-series, and they play a major role in forecasting/nowcasting quarterly real GDP growth; see for example Stock and Watson (1989) and Stock et al. (2012). They are also used to nowcast GDP growth in real time, as new releases of data become available; see for example Giannone et al. (2008). The results are sometimes harder to interpret than in the case of DMA, where each variable (or group of variables) can be singled out and its effect and importance analyzed individually. The two approaches are not incompatible, as factors can be used as potential regressors in DMA.

Large Bayesian vector autoregressive models have also been successfully applied to large datasets and are routinely used in policy institutions (Banbura et al., 2010). They have the advantage of easy estimation and computation of conditional and unconditional forecasts. However, they normally require very tight priors when used on large datasets, and specifying the priors is still something of an art despite recent progress (Giannone et al., 2015).

All these techniques have advantages and disadvantages, and the choice should depend on the forecast of interest. Model averaging emphasizes model uncertainty and structural change and may be preferred when assessing which variables are important in a regression and how this importance evolves over time. In the rest of the paper our benchmark will be DMA, to emphasize the role of the Dynamic Occam's Window.

3. Dynamic Occam's window

When many potential regressors are considered, the number of models is too large to be tractable. However, typically most of the models contribute little to the forecast, as their weights are close to zero. These include for example highly misspecified models, which must be kept while carrying out DMA, despite their poor performance, only to calculate Eq. (2).

We propose a heuristic aiming at eliminating most of these low probability models from the computation, while being able to "resurrect" them when needed. This is an extension of the Occam's Window method of Madigan and Raftery (1994), in which model averaging is based only on the models whose posterior model probability is greater than some multiple C of the highest model probability. (Madigan and Raftery, 1994 used $C = 1/20$, while subsequent implementations have used lower values.)

We now extend Occam's Window to the dynamic context. Our Dynamic Occam's Window (DOW) method is based on two 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 relatively close (in an appropriately defined "neighborhood") to those at the preceding time.

We believe these assumptions are reasonable in typical problems of macroeconomic analysis. If verified, they allow the exploration of the space of models in a parsimonious and efficient way.

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

We propose to implement Occam's 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 $M_0(t)$, and renormalize their current weights, $w_{t|t,k}$, so that they sum to 1 over the current set of models, i.e. so that $\sum_{k:m_k \in M_0(t)} w_{t|t,k} = 1$. After choosing a threshold $C \in (0,1]$, we keep for future use the models $m_k \in M_0(t)$ that fall in Occam's window, namely those in the set

$$\left\{ m_k : m_k \in M_0(t), w_{t|t,k} \ge C * \max_{\ell : m_\ell \in M_0(t)} w_{t|t,\ell} \right\}. \tag{4}$$

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

Initialization:

1. Divide the sample $1,\ldots,T$ into an in-sample period $1,\ldots,T_r$ and a pseudo out-of-sample period $T_r + 1,\ldots,T$.
2. Start with an initial population of models $M_0(T_r)$ and an initial set of weights $w_{T_r|T_r,k}$.

For $t = (T_r) + 1,\ldots,T$:

1. (Forecast) Use the models in $M_0(t-1)$ and the weights $w_{t|t-1,k}$ to perform model averaging as in Raftery et al. (2010), obtaining the forecast distribution $p(y_t|y_{1:t-1}, M_0(t-1))$. We call this the reduced DMA forecast, or DMA-R, because it is obtained with the smaller model population inherited from the previous period.
2. (Expand) Expand $M_0(t-1)$ to a larger population of models $M_1(t)$, including all $m_k \in M_0(t-1)$ and all their "neighboring" models. We define neighboring models as models derived from any model $m \in M_0(t-1)$ by adding or removing a regressor, but different definitions are possible. We call this the expanded DMA forecast, or DMA-E, because it is obtained with the expanded model population. Both DMA-R and DMA-E are based on past information and do not include current information.
3. (Assess) Upon observing y_t, compute the weights $w_{t|t,k}$ for all $m_k \in M_1(t)$, and normalize them to sum to 1 over $M_1(t)$.
4. (Reduce) Define the final population of models for time t, $M_0(t)$, as those in $M_1(t)$ that are in Occam's Window, namely $M_0(t) = \{m_k \in M_1(t): w_{t|t,k} \geq C*\max_{\ell:m_\ell \in M_1(t)} w_{t|t,\ell}\}$.

3.2. Computational issues

We now explain why Dynamic Occam's Window allows the exploration of large model spaces that would not be possible otherwise.

We define, somewhat 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. Using this loosely defined but quite general metric, we compare the DOW method with a DMA that exhaustively explores the space of models. Let J be the number of candidate explanatory variables, T be the number of time points for which we have data, and N be the number of models in Occam's window (a subset of the K possible regression models).

DMA with all models has approximate computational cost

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

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

The DOW 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. Its computational cost is thus approximately that of estimating about

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

different models, where N is the average number of models evaluated at each time point. The role of the average number of models N is explored in Section 5. In particular, we will consider versions of the DOW method where the number of models at each time point is constrained not to exceed a given upper limit, in which case N is effectively specified by the user.

The DOW method allows gains in speed when $\mathrm{NUC_{DOW}} < \mathrm{NUC_{DMA}}$, or

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

To illustrate, consider the case where $T=45$, $J=30$, and $N=10{,}000$. Then

$$\frac{\mathrm{NUC_{DOW}}}{\mathrm{NUC_{DMA}}} = \frac{10.350.000}{48.318.382.080} = 0.02\%, \tag{8}$$

and so the DOW method is about 5000 times faster in this case.

Fig. 1 shows the relationships (5) and (6). The computational complexity of the DOW method grows quadratically with the length of the available series T, while that of DMA grows only linearly in T but increases exponentially in the number of regressors J. Above 15-20 regressors the DOW method is always more efficient computationally. This is particularly true when the time series are relatively short, since longer series imply a higher number of estimations for each model in the case of Occam's window.

4. An economic application: forecasting GDP growth in the Euro area in the great recession

4.1. Data

We apply the DOW method to nowcasting GDP growth 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 Table 1.[1] Abstracting

[1] We use quarterly data or data at higher frequency. Higher frequency stock data are converted to quarterly data by taking the last observation (for example the last month in a quarter); as a robustness check we also experimented with averages, obtaining very similar results. Flow data are always averages.

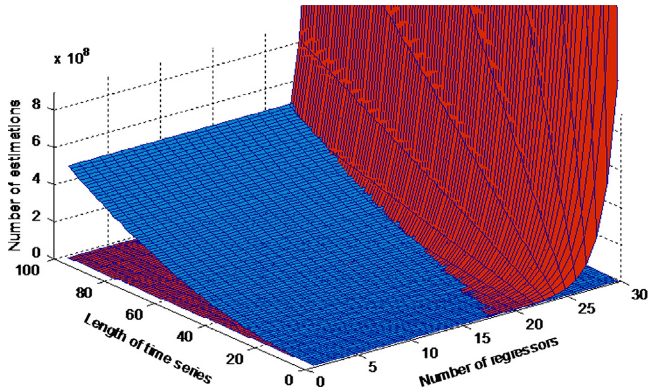

Fig. 1. *Computing time*: The number of Notional Units of Computation, or NUC (vertical axis), plotted against the data length and the number of regressors. The blue area refers to the DOW method, the red area to DMA. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 1
Variables used in the Euro Area GDP growth forecasting area application.

Code	Variable	Transformation	Data source
hicp	Euro area – HICP, Eurostat	Annual growth rate	Eurostat
hicpx	Euro area – HICP excl. unprocessed food an energy, Eurostat	Annual growth rate	Eurostat
ppi	Euro area – Producer Price Index, domestic sales, Total Industry (excluding construction) – Neither seasonally nor working day adjusted	Annual growth rate	Eurostat
unemp	Euro area – Standardized unemployment, Rate, Total (all ages), Total (male and female), percentage of civilian workforce	Level	Eurostat
ip	Euro area – Industrial Production Index, Total Industry (excluding construction) – Working day adjusted, not seasonally adjusted	Annual growth rate	Eurostat
l1 to l5	Euro area – Gross domestic product at market price, Chain linked, ECU/euro, Working day and seasonally adjusted	Annual growth rate	Eurostat
spfgdp2	Survey Professional Forecasters; Forecast topic: Real GDP growth; Point forecast	Annual growth rate	ECB
spfgdp2_var	Survey Professional Forecasters; Forecast topic: Real GDP growth; Variance of Point forecast	Variance	ECB
pmi_e	PMI employment	Level	Markit
pmi_ord	PMI new orders	Level	Markit
pmi_y	PMI output	Level	Markit
x_oil	Brent crude oil 1-month Forward – fob (free on board) per barrel – Historical close, average of observations through period - Euro	Annual growth rate	ECB
x_rawxene	Euro area, ECB Commodity Price index Euro denominated, import weighted, Total non-energy commodity	Annual growth rate	ECB
x_neer	Nominal effective exch. Rate	Annual growth rate	ECB
xusd	Exchange Rates; Currency: US dollar; Currency denominator: Euro; Exchange rate type: Spot	Annual growth rate	ECB
i_short	Euro area – Money Market – Euribor 1-year – Historical close, average of observations through period - Euro	Level	Reuters
i_long	Euro area – Benchmark bond – Euro area 10-year Government Benchmark bond yield – Yield – Euro	Level	ECB
m3	Euro Area – Monetary aggregate M3, All currencies combined – Working day and seasonally adjusted	Annual growth rate	ECB
spread	Government bond, nominal yield, all issuers whose rating is triple A (less) Government bond, yield nominal, all issuers all ratings included	Level	ECB
stress	Composite Indicator of Systemic Stress; Euro area ; Systemic Stress Composite Indicator	Level	ECB
risk_eb	Euro area, Financial market liquidity indicator: Foreign currency, equity and bond markets	Level	ECB
risk_tot	Euro area, Financial market liquidity indicator: Composite indicator	Level	ECB
risk_glob	Euro area, Financial market liquidity indicator: Global risk aversion indicator	Level	ECB
risk_mon	Euro area, Financial market liquidity indicator: Money market	Level	ECB
stox	Dow Jones Eurostoxx 50 Index – Historical close, average of observations through period	Annual growth rate	Reuters
domcred	Euro area, Loans [A20] and securities [AT1], Total maturity, All currencies combined – Working day and seasonally adjusted	Annual growth rate	ECB

from minor differences in publication dates, there are two main nowcasts that a forecaster may perform, depending on whether or not the preceding quarter figure for GDP is available. For simplicity of exposition, we focus on the case when the past quarter is already available. Our nowcasts will be based on an information set comprising past GDP growth 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 two-year-head 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 relevance of uncertainty to the macroeconomic developments included in our sample, we also 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 rates, 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 result, 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. The possibility of adding regressors but discarding them adaptively if they are unnecessary is one of the advantages of our methodology.

In order to concentrate on the effects of the proposed DOW method, we simplify the method of Raftery et al. (2010) slightly, and estimate each model recursively but with fixed parameters. We choose this setup because Koop and Korobilis (2012) have shown that DMA is a good substitute for time varying parameters, and we want to concentrate on the advantages of the DOW method alone in accounting for model changes. DMA is performed as in Raftery et al. (2010), using a discount factor set at $\alpha = 0.99$.

4.2. *Forecasting performance*

Fig. 2 shows that the DOW method had a satisfactory nowcasting performance overall, even in the presence of turning points. The accuracy of the method, as expected, increased with the amount of available data. The 95% prediction intervals take into account both the within and between model uncertainty.

The difficult episode of the recession in 2008–2009 was well captured by DMA. The forecast slightly underpredicted in the trough, but it immediately recovered and became quite accurate in the aftermath of the crisis.

Table 2 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. The pseudo out-of-sample period ranges from 2003q1 until 2014q1.

We use as evaluation metrics: the Root Mean Squared Forecast Error (RMSE), the Mean Average Forecast Error (MAE) and the Maximum Forecast Error (MAX), expressed as ratios relative to the random walk benchmark; the average log score differential (LLIK); and the average continuous ranked probability score differential (ACPRS) (Gneiting and Raftery, 2007). For RMSE, MAE and MAX, smaller is better, while for LLIK and ACRPS, bigger is better.

The DMA-R forecast is based on the smaller population of models M_0 and compares favorably with simple benchmarks. It beats the simple random walk and a standard $AR(2)$ model by a wide margin. We recall that the forecast DMA-R is based on

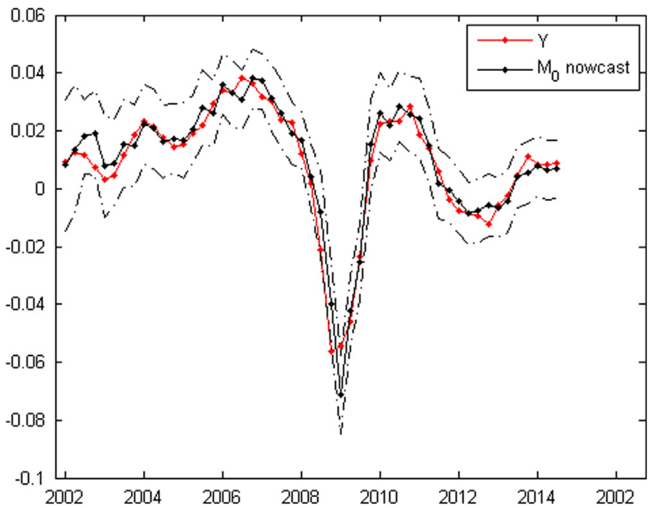

Fig. 2. Forecasting Euro Area GDP growth Using DMA with the DOW Method: nowcasts and 95% prediction intervals.

Table 2
Forecasting performance of different forecasting methods.

Metric	RW	AR2	DMA-R	DMA-E	DMS-R	DMS-E
RMSE	1	0.8755	0.5154	**0.5072**	0.5183	0.5644
MAE	1	0.8922	0.5673	**0.5595**	0.5837	0.6377
MAX	1	1.0721	0.4872	0.4774	**0.4606**	**0.4606**
LLIK	0	0.1884	**0.6682**	0.6622	0.5395	0.4103
ACPRS	0	0.0006	**0.0023**	**0.0023**	0.0022	0.0019

Note: Methods: RW=Random walk model; AR2=second-order autoregressive model; DMA-R=reduced DMA method; DMA-E=expanded DMA method; DMS-R=reduced DMS method; DMS-E=expanded DMS method.

Metrics: RMSE=root mean squared error; MAE=mean absolute error; MAX=maximum absolute error; LLIK=average log score differential; ACPRS=average continuous ranked probability score differential. For RMSE, MAE and MAX, smaller is better, while for LLIK and ACRPS, bigger is better. The best method by each metric is shown in bold font.

past GDP and recent information on the indicator variables.

Forecasts computed using the extended population of models M_1 are reported as DMA-E. The results are very close to those of DMA-R. 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 more quickly 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. DMA-R uses many fewer models than DMA-E and so may be preferred to DMA-E on the grounds of simplicity, ease of interpretation and computational efficiency.

In most cases DMA beats the corresponding forecast computed with DMS, although by a small margin, corroborating the common finding that model averaging can beat even the best model in the pool.

Following Koop and Korobilis (2012), we tried additional benchmarks. These included a single time-varying parameter model including all regressors, and a single Bayesian ordinary least squares model with all regressors. However, these models performed poorly as their estimated paramaters were unstable.

4.3. *Variable posterior inclusion probabilities*

An important value added of model averaging (beyond the good forecasting performance) is the posterior inclusion probabilities of each regressor and their evolution. DMA identifies the importance of single variables and how this varies over time, which helps interpretation. The posterior inclusion probability of a variable at a given time point is calculated by summing the posterior probabilities of the models that use that variable as a regressor. Thus they vary between 0 and 1 and give a measure of the importance of that regressor at the given time point. Their evolution in time is summarized in Figs. 3 and 4.

The posterior inclusion probabilities identify which were the most useful indicators of real activity and how this changed over time. In more detail:

- Lags of GDP were, as expected, important overall. The first lag captures the persistence in GDP, and it remained important even during the crisis, when GDP showed pronounced swings. The fourth and fifth lag capture essentially base effects. Our decision to include lag 5 as a potential regressor turned out to be justified.
- Among the consumer price variables, HICP was an important regressor over most of the sample. This confirms the idea that prices and output are not determined in isolation. Without extending our interpretation to the existence of a European Phillips curve, we notice that this confirms the results for the euro area recently obtained by Giannone et al. (2014).
- Among the early indicators of real activity, industrial production was the most important. This is a well known result in nowcasting, where industrial production is widely used as a timely and already comprehensive subset of GDP. The role of unemployment changed over time, becoming more important in the aftermath of the crisis.
- DMA selected GDP surveys as overall important over the sample, with the exception of the period immediately following the 2008 crisis, which the surveys failed to capture adequately. This confirms the literature on nowcasting inflation (Ang et al., 2007), arguing that surveys have nowcasting power, thereby supporting the importance of expectations in determining macroeconomic outcomes.
- No single external variable alone had a determinant role. This is possibly due to the relative compactness of the euro area. Variables traditionally important in determining prices, such as oil and commodity prices or the exchange rate, appear to have had a limited impact on real GDP. We find this result interesting but not surprising, given that these variables mostly affect prices, and affect GDP only indirectly.
- Among the variables closer to the operation of monetary policy, interest rates progressively lost their importance in the credit constrained post-crisis period, while the monetary variable M3 had an increasing role, possibly highlighting the importance of liquidity in the recent part of the sample.

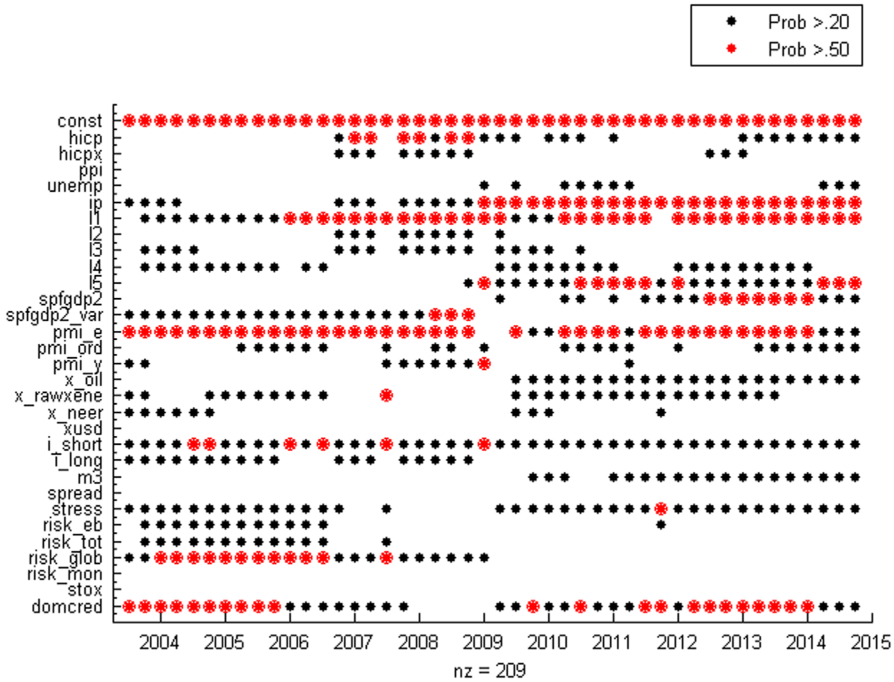

Fig. 3. Euro Area GDP growth: Posterior inclusion probabilities of variables over time: black above 20%, red above 50%. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

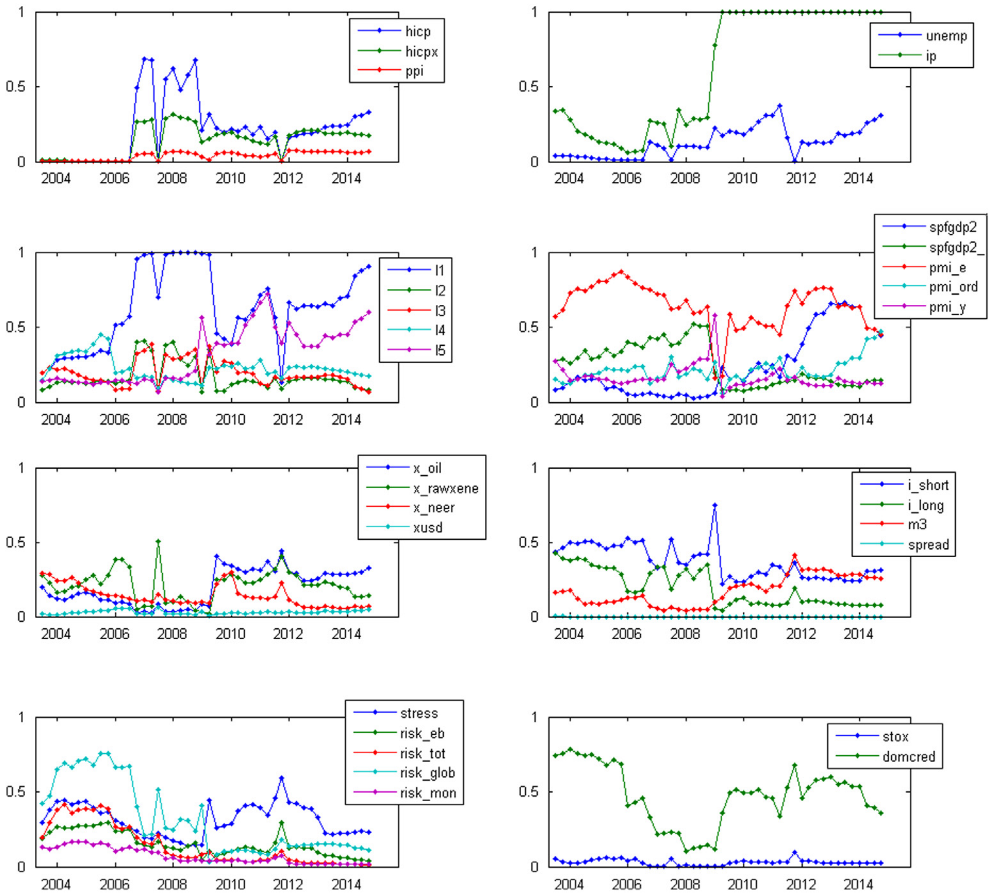

Fig. 4. Euro Area GDP growth: Posterior inclusion probabilities of single variables over time.

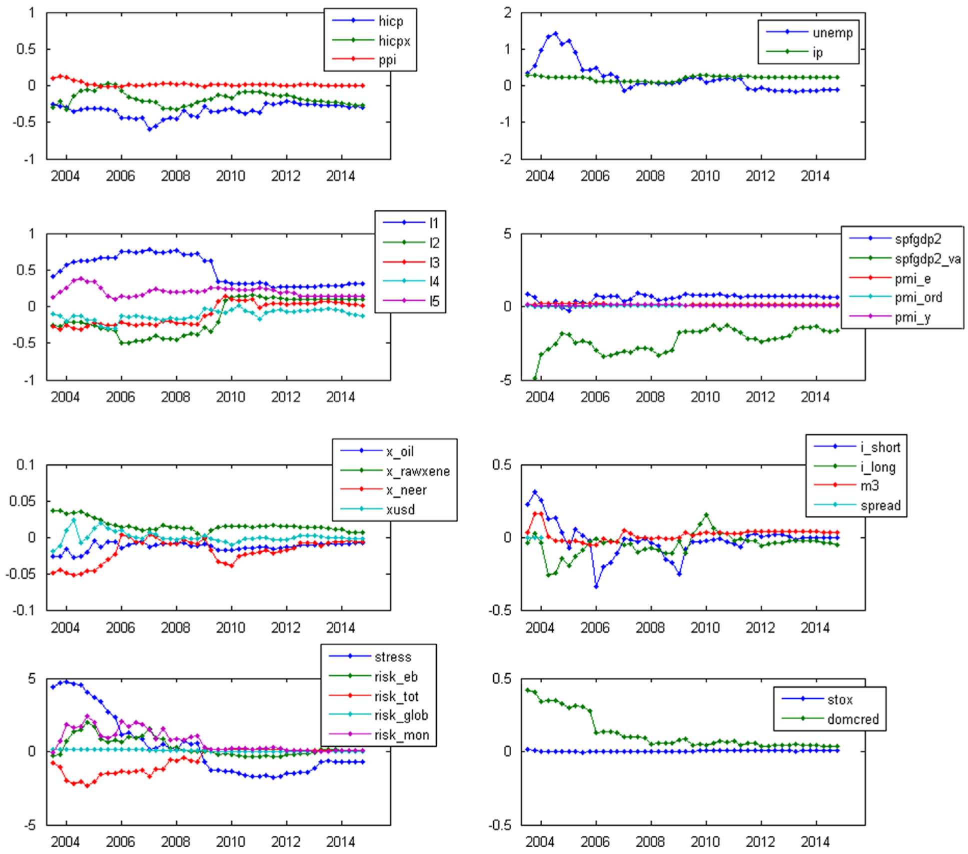

Fig. 5. Euro Area GDP growth: Posterior means of regression coefficients over time under DMA.

- Risk and stress indicators were useful predictors of GDP. Overall risk indicators seemed to matter before the crisis, in particular the indicator of global risk. The stress indicator was important during the crisis.
- Finally, stock market developments did not seem to be a very useful predictor, while (lack of) domestic credit was overall important during the crisis.

Koop and Onorante (2014) carried out a similar analysis of inflation. They used DMA on a similar dataset, but they explored the whole model space, which limited the number of predictor variables they could use. Comparing their results with ours, it is apparent that the determinants of GDP growth were fairly similar to those of inflation. In particular, variables representing expectations are important predictors, with the exception of 2008–2009 at the beginning of the crisis period.

A natural complement to the results above is the posterior mean of the coefficient of each variable. These are shown in Fig. 5. While posterior inclusion probabilities provide important information about which variables should be included in the regressions at each point in time, they do not specify the size of their effect, and even a variable with a very high posterior inclusion probability may have a small overall impact on GDP. The posterior means are averages over models at each point in time, and so vary over time.

5. Sensitivity analysis

We now assess the performance of the DOW method and how the performance changes with different specifications of the algorithm. We assess its sensitivity to different initial conditions, different maximum numbers of models, and finally we use a large and noisy database. We find that the method's performance is not very sensitive to these changes.

5.1. Initial conditions

The DOW method requires the specification of an initial set of models at the first time point. In our implementation we used an initial set consisting of just one-variable models. In this section we check the sensitivity of the forecast to the choice of the initial population of models.

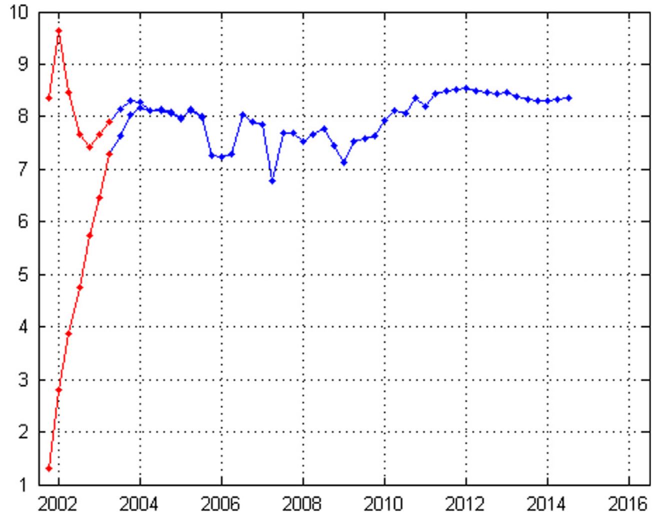

Fig. 6. Euro Area GDP growth: Evolution of average model size starting from different initial model populations: one-variable models or 100 randomly selected models of average size 8. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 3
Euro Area GDP growth: forecasting from an initial population of models of average size 8.

Metric	RW	AR2	DMA-R	DMA-E	DMS-R	DMS-E
RMSE	1	0.8755	0.5283	**0.5085**	0.5644	0.5644
MAE	1	0.8922	0.573	**0.5623**	0.6377	0.6377
MAX	1	1.0721	0.4872	0.4774	**0.4606**	**0.4606**
LLIK	0	0.1884	**0.6671**	0.622	0.4103	0.4103
ACPRS	0	0.0006	**0.0022**	0.0021	0.0019	0.0019

Note: Methods: RW=Random walk model; AR2=second-order autoregressive model; DMA-R=reduced DMA method; DMA-E=expanded DMA method; DMS-R=reduced DMS method; DMS-E=expanded DMS method.
Metrics: RMSE=root mean squared error; MAE=mean absolute error; MAX=maximum absolute error; LLIK=average log score; ACPRS=average continuously ranked probability score. The best method by each metric is shown in bold font.

Fig. 6 reports the average number of variables included in the models. The same average size of models is a necessary (although not sufficient) condition for convergence in populations of models, and it allows an easy graphic exploration of convergence.

It appears that DMA favors models with about 7–8 variables, and thus that the initial population of models M_0 is not representative of the final models selected. Fig. 3 supports this finding by showing that the posterior inclusion probabilities changed rapidly at the beginning of the period.

We test two very different initial conditions: (1) our baseline population of one-variable models only, and (2) a population of 100 randomly selected models with eight variables each. Fig. 6 shows that beyond the few periods of presample (in red) and a few initial points, the two initial populations gave comparable results in terms of model size.

The same holds for forecasting performance. Table 3 reports the forecasting performance of the alternative, random initial population of models with average size 8. The results are very similar to those in Table 2, in which DMA was initialized with one-variable models only. Once again, the DMA methods outperformed the others considered.

5.2. Maximum number of models

A pure application of the Occam's window principle and of the FEAR algorithm would require keeping each model that satisfies condition (4). This may sometimes lead to a relatively high number of models in the wider population M_1, as this population, generated from the Expand step of the algorithm, includes all possible neighbors of the preceding population M_0. The size of the latter, however, is controlled by the following Reduce step, where condition (4) is applied. Fig. 7 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 note that we start our pseudo out-of-sample period after ten data points; as a result many models are poorly estimated at first and their performance varies considerably. After a few periods, a stable population has been found and it is progressively refined; as a result, the size of M_0 decreases rapidly. From this point on, the FEAR algorithm increases the population size during turbulent times, for example during the Great Recession. Whenever the

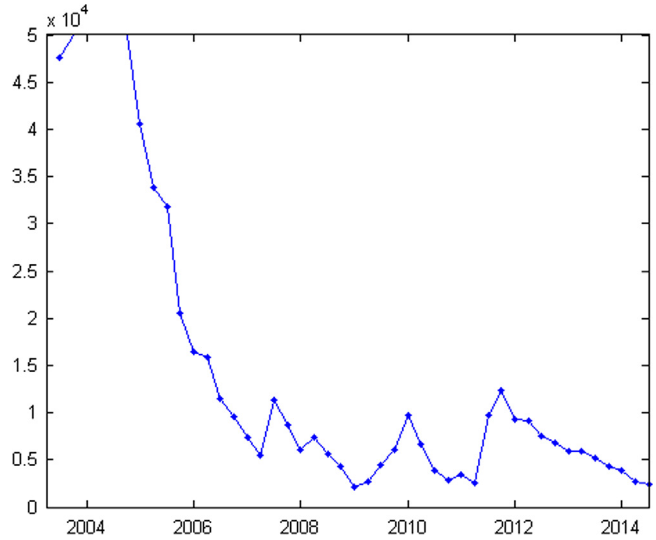

Fig. 7. Euro Area GDP growth: The number of models over time in the reduced set of models, M_0, in the Dynamic Occam's Window algorithm.

forecast is less accurate and no model is clearly dominating, the algorithm "resuscitates" additional models in the attempt to improve the forecast. Fig. 7 shows that this attempt is usually successful. Quiet periods are instead characterized by smaller, decreasing model populations.

Finally, as the sample size increases and models including the best regressors are selected, the population size becomes quite small (the last M_0 had size 2350). Overall, the population M_0, from which the baseline nowcast is generated, never exceeded 100,000 models, while the wider M_1 could be up to about ten times larger.

In the interest of computational speed, we introduced the possibility of specifying a maximum number of models N, and we now experiment with this number in order to assess whether it implies a deterioration of the forecast. Fig. 8 reports the nowcasting performance in relation to maximum model size N.

In our model space of 30 potential regressors, forecasting performance improved until the size of the population M_0 reached about 10,000 models. Bigger model populations did not lead to any further improvement, as we have seen from the unconstrained estimation. Constraints set at 50,000 or more on the maximum number of models were mostly not reached, and thus provide results equivalent to Occam's window without a maximum number of models. We would of course still recommend keeping the maximum number of models as high as possible, subject to computational constraints.

Fig. 8 also confirms that in our case DMA performed slightly better than DMS for any maximum number of models. This is a robust result in the case of macroeconomic variables, but it cannot be generalized. Koop and Onorante (2014) and Morgan et al. (2014), for example, have shown using Google searches as predictors that DMS performed better than DMA in situations where the data were noisy and forecasting benefitted from excluding many regressors.

When looking at single regression parameters, we observe that convergence may be slower for parameters with low posterior inclusion probabilities. For some specific parameters and posterior inclusion probabilities there are observable convergence issues up to 50,000 models. The low-probability models containing these parameters do not affect the overall forecast much, but when this more specific information is important, we would suggest increasing the maximum number of models by one (or if possible two) orders of magnitude.

5.3. *Adding many noisy regressors*

Until now we have been using a standard, large-sized policy database, composed of "reasonable" regressors used in the profession. As a robustness check we now introduce a very large database. Despite the large size of the data and the noisiness of the database, the DOW delivers a good performance.

The enlarged database includes 60 variables (including the lags of GDP growth). We collected a large dataset for the Euro Area and we did not select out any variable, thus leaving regressors that have no obvious relationship with output growth either in theory or in forecasting practice.

This leads to two problems. First, a high quantity of "noise" relative to informative data has been shown to reduce the accuracy of forecasts (Morgan et al., 2014 with Google indicators). While this problem is minimized in DMA because the most significant regressors are selected and most others are assigned low weights, the same result needs to be shown to hold for the DOW method. Second, a model space of 2^{60} models can only be sparsely explored even when increasing the maximum number of retained models. We want to show that both factors lead to only a small deterioration of the forecast; for this reason, we keep the number of models in M_0 to 10,000 as in the baseline scenario.

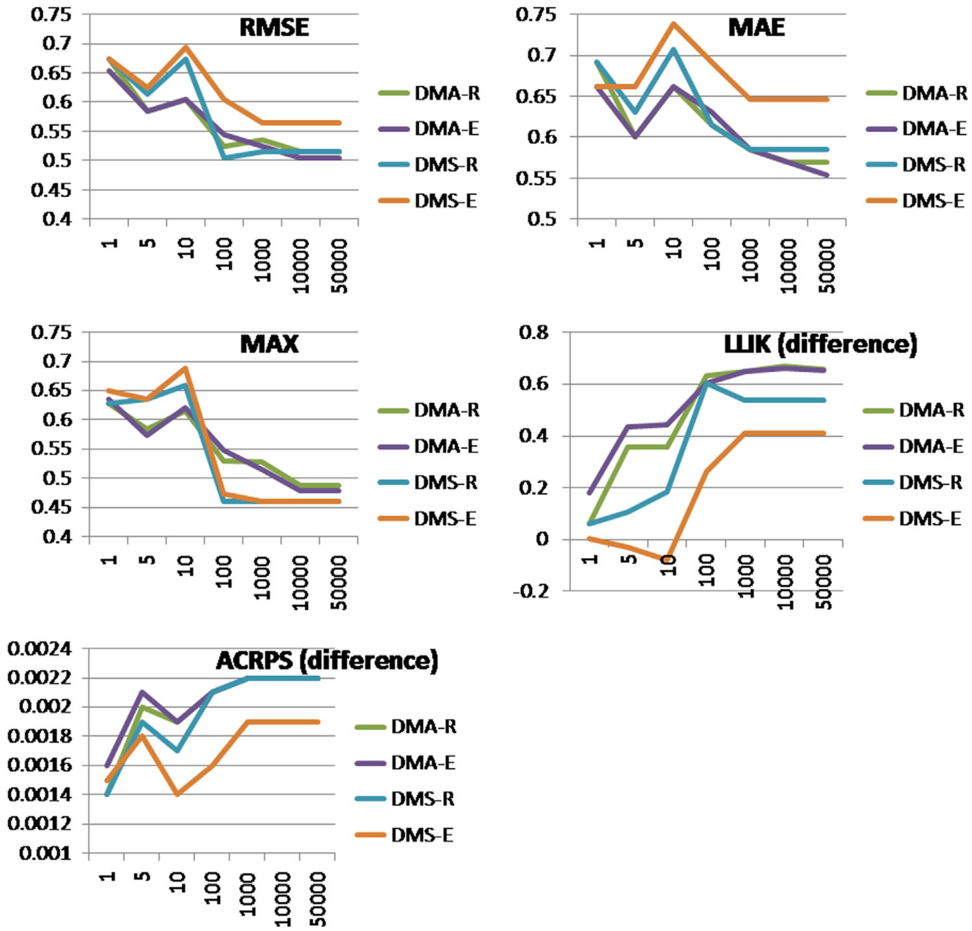

Fig. 8. Euro Area GDP growth nowcasting: Performance metrics relative to random walk, for different maximum numbers of models.

Table 4
Euro area GDP growth: forecasting from a population of 2^{60} models.

Metric	rw	AR2	DMA-R	DMA-E	DMS-R	DMS-E
RMSE	1	0.8755	0.5387	0.5317	**0.5172**	0.5635
MAE	1	0.8922	0.5832	**0.5797**	0.5835	0.6375
MAX	1	1.0721	0.5279	0.5163	**0.4606**	**0.4606**
LLIK	0	0.1884	**0.6497**	0.6492	0.5391	0.4108
ACPRS	0	0.0006	**0.0022**	**0.0022**	**0.0022**	0.0019

Note: Methods: RW = Random walk model; AR2 = second-order autoregressive model; DMA-R = reduced DMA method; DMA-E = expanded DMA method; DMS-R = reduced DMS method; DMS-E = expanded DMS method.
Metrics: RMSE = root mean squared error; MAE = mean absolute error; MAX = maximum absolute error; LLIK = average log score; ACPRS = average continuously ranked probability score. The best method by each metric is shown in bold font.

The statistics in Table 4 show that, compared to the DMA where the full set of regressors considered was chosen by the forecaster, there is only a small deterioration in forecasting performance. The performance of DMS also deteriorates due to the frequent changes in the model chosen to forecast but improves slightly in relative terms.

This result suggests that, while it is still preferable to use expert judgement and theoretical guidance in the choice of potential regressors, overall the DOW is quite robust to the introduction of many, noisy regressors, and can be applied to wide datasets even without a previous selection of variables.

6. Discussion

We have proposed a new method for carrying out Dynamic Model Averaging when the model space is too large to allow exhaustive evaluation, so that the original DMA method of Raftery et al. (2010) is not feasible. This method, based on

Occam's window and called Dynamic Occam's Window (DOW), is particularly efficient when many time series of limited length are available, as is typically the case in macroeconomics. Our procedure allows us 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 assessed the model in an important empirical application, nowcasting GDP in the euro area. We showed that the forecasting performance was satisfactory compared to common benchmarks, and the computational burden was substantially lower. Several sensitivity analyses confirm the robustness of our approach to the choice of the user-specified control parameters.

There are several areas we left for future research. Additional sensitivity analyses could be useful, for example using linear and non linear models, or data at higher frequency. Furthermore, a forecast comparison and a deeper analysis of the pros and cons of DMA and other comparable techniques would help put this new method into broader context.

Acknowledgments

The authors are grateful to the editor and two anonymous reviewers for very helpful comments. Raftery's research was supported by NIH grants R01 HD054511, R01 HD070936 and U54 HL127624, and by a Science Foundation Ireland E.T.S. Walton visitor award, grant reference 11/W.1/I2079. Raftery thanks the School of Mathematical Sciences at University College Dublin for hospitality during part of the preparation of this paper.

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.euroecorev.2015.07.013.

References

Ang, A., Bekaert, G., Wei, M., 2007. Do macro variables, asset markets, or surveys forecast inflation better? J. Monet. Econ. 54, 1163–1212.
Banbura, M., Giannone, D., Reichlin, L., 2010. Large Bayesian VARs. J. Appl. Econ. 25, 71–92.
Brock, W., Durlauf, S.N., 2001. Growth empirics and reality. World Bank Econ. Rev. 15, 229–272.
Brock, W., Durlauf, S.N., West, K., 2003. Policy evaluation in uncertain economic environments. Brookings Pap. Econ. Act. 1, 235–322.
Del Negro, M., Hasegawa, R.D., Schorfheide, F., 2014. Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance. Staff Report 695, Federal Reserve Bank of New York.
Durlauf, S.N., Kourtellos, A., Tan, C.M., 2008. Are any growth theories robust? Econ. J. 118, 329–346.
Durlauf, S.N., Kourtellos, A., Tan, C.M., 2012. Is God in the Details? a Reexamination of the Role of Religion in Economic Growth. Technical Report.
Eicher, T., Papageorgiou, C., Raftery, A.E., 2010. Determining growth determinants: default priors and predictive performance in Bayesian model averaging. J. Appl. Econ. 26, 30–55.
Fernández, C., Ley, E., Steel, M.F.J., 2001a. Benchmark priors for Bayesian model averaging. J. Econ. 100, 381–427.
Fernández, C., Ley, E., Steel, M.F.J., 2001b. Model uncertainty in cross-country growth regressions. J. Appl. Econom. 16, 563–576.
Geweke, J., Amisano, G., 2010. Comparing and evaluating Bayesian predictive distributions of asset returns. Int. J. Forecast. 26, 216–230.
Geweke, J., Amisano, G., 2011. Optimal prediction pools. J. Econom. 164, 130–141.
Giannone, D., Lenza, M., Momferatou, D., Onorante, L., 2014. Short-term inflation projections: a Bayesian vector autoregressive approach. Int. J. Forecast. 30, 635–644.
Giannone, D., Lenza, M., Primiceri, G.E., 2015. Prior Selection for Vector Autoregression. Technical Report.
Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: the real-time informational content of macroeconomic data. J. Monet. Econ. 55, 665–676.
Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14, 382–417.
Koop, G., Korobilis, D., 2012. Forecasting inflation using dynamic model averaging. Int. Econ. Rev. 53, 867–886.
Koop, G., Onorante, L., 2014. Macroeconomic Nowcasting Using Google probabilities. Mimeo, August 2014.
Leamer, E.E., 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York.
Ley, E., Steel, M.F.J., 2007. Jointness in Bayesian variable selection with applications to growth regression. J. Macroecon. 29, 476–493.
Ley, E., Steel, M.F.J., 2009. On the effect of prior assumptions in BMA with applications to growth regression. J. Appl. Econ. 24, 651–674.
Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Am. Stat. Assoc. 89, 1535–1546.
Morgan, J., Muzikarova, S., Onorante, L., 2014. Nowcasting European Unemployment Using Internet Search Data. Mimeo, August 2014.
Raftery, A.E., 1988. Approximate Bayes Factors for Generalized Linear Models. Technical Report 121, Department of Statistics, University of Washington.
Raftery, A.E., 1995. Bayesian model selection for social research. Sociol. Methodol. 25, 111–163.
Raftery, A.E., Kárný, M., Ettler, P., 2010. Online prediction under model uncertainty via dynamic model averaging: application to a cold rolling mill. Technometrics 52, 52–66.
Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. 92, 179–191.
Sala-i-Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. Am. Econ. Rev. 94, 813–835.
Steel, M.F.J., 2011. Bayesian model averaging and forecasting. Bull. E.U. U.S. Inflat. Macroecon. Anal. 200, 30–41.
Stock, J.H., Watson, M.W., 1989. New indexes of Coincident and Leading Economic Indicators. NBER Macroeconomics Annual.
Stock, J.H., Watson, M.W., 2012. Dynamic factor models. In: Clements, M.P., Hendry, D.F. (Eds.), The Oxford Handbook of Economic Forecasting, Oxford University Press, Oxford, U.K..
Varian, H.R., 2014. Big data: new tricks for econometrics. J. Econ. Perspect. 28, 3–28.