International Journal of Forecasting 35 (2019) 1226–1239

Contents lists available at ScienceDirect

International Journal of Forecasting

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

A comprehensive evaluation of macroeconomic forecasting methods

Andrea Carriero[a], Ana Beatriz Galvão[b,*], George Kapetanios[c]

[a] Queen Mary University of London, United Kingdom
[b] University of Warwick, United Kingdom
[c] King's College London, United Kingdom

ARTICLE INFO

Keywords:
Factor models
BVAR models
MIDAS models
DSGE models
Density forecasts
Meta-analysis

ABSTRACT

We employ datasets for seven developed economies and consider four classes of multivariate forecasting models in order to extend and enhance the empirical evidence in the macroeconomic forecasting literature. The evaluation considers forecasting horizons of between one quarter and two years ahead. We find that the structural model, a medium-sized DSGE model, provides accurate long-horizon US and UK inflation forecasts. We strike a balance between being comprehensive and producing clear messages by applying meta-analysis regressions to 2,976 relative accuracy comparisons that vary with the forecasting horizon, country, model class and specification, number of predictors, and evaluation period. For point and density forecasting of GDP growth and inflation, we find that models with large numbers of predictors do not outperform models with 13–14 hand-picked predictors. Factor-augmented models and equal-weighted combinations of single-predictor mixed-data sampling regressions are a better choice for dealing with large numbers of predictors than Bayesian VARs.

1. Introduction

Forecasting is one of the major aims of economic and econometric analysis, along with modelling the foundations of economic phenomena. As a result, academic work has made considerable efforts to lay the foundations and to build tools for efficient forecasting.

The literature on macroeconomic forecasting can be divided into two broad categories. The first aims to produce models that attempt to explain the economy first and then provide forecasts only as a byproduct of their main aim. In principle, this is optimal, in the sense that a model which can explain the economy successfully should be able to forecast well. Nevertheless, the complexity of the economy and of the models that are needed for its full explanation implies that such forecasts might not be accurate in-sample, let alone out-of-sample.[1] The second strand of research considers models that do not attempt full structural modelling, but simply a reduced-form statistical description. These models frequently have superior forecasting performances, but their reduced-form nature makes it harder to provide economic storytelling to support forecasts. This characteristic is classified as a relevant disadvantage by many economists and policymakers.

This has not stopped either the proliferation of reduced-form models or a rapid increase in their sophistication. Recent trends in this literature include the modelling of structural changes and the efficient use of

* Correspondence to: Prof. Ana Galvão, Finance Group - Warwick Business School, University of Warwick, Coventry, CV4 7AL, United Kingdom.

E-mail address: ana.galvao@wbs.ac.uk.

[1] For example, Chauvet and Potter (2013) and Faust and Wright (2013) conclude their reviews on the forecasting performances of structural and reduced-form models for predicting inflation and output growth by arguing that structural models do not have better forecast accuracies than univariate time series models.

https://doi.org/10.1016/j.ijforecast.2019.02.007
0169-2070/© 2019 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

increasingly large datasets. The former has been driven by the widespread recognition that structural change is a leading cause of forecast failure. A number of approaches of varying degrees of sophistication are being used to accommodate structural change. These range from time-varying coefficient models to methods that allow for the time-varying estimation of standard econometric forecasting models. In this context, as is common with forecasting in general, these increases in sophistication have been found not to necessarily be correlated closely with superior forecasting performance.[2] The second trend of considering large datasets has been spurred on by their use in many economic analyses, given their availability in central banks and other policy-making institutions.[3]

The above developments set the scene for the current paper. Our goal is to provide a comprehensive, state of the art evaluation of recently proposed model classes for forecasting output growth and inflation, giving special attention to model classes that are able to deal with large numbers of predictors. The aim of the paper is to strike a balance between being comprehensive and producing clear messages. This requires us to consider a wide range of models, but to be selective in some dimensions so as to make the evaluation exercise feasible and informative. Furthermore, it requires an evaluation across a number of different countries and sample periods. Finally, we aim to compare and contrast reduced-form models and structural models, which have traditionally been considered inferior for forecasting purposes. This latter aspect of our analysis is found less commonly in forecasting evaluations.[4]

Forecasting comparisons in the literature normally focus on data from only a single country or a small subset of countries (US, UK and euro area).[5] Instead, we will use data from seven economies: US, UK, euro area, Germany, France, Italy and Japan. For these seven economies, we compute forecasts for output growth and inflation using three classes of state-of-the-art reduced-form forecasting models: factor-augmented distributed lag (FADL) models, mixed data sampling (MIDAS) models, and Bayesian vector autoregressive (BVAR) models.[6] These model classes are useful for exploring the predictive information contained in a large number of indicators. As a consequence, we build a dataset with a large number of monthly indicators for each country and assess the importance of employing large (100 predictors) datasets in comparison with medium-sized (a dozen predictors) and small datasets for macroeconomic forecasting. We also consider one class of structural models: a medium-sized dynamic stochastic general equilibrium model (DSGE). We compare the DSGE model's performance with that of reduced-form models for forecasting output growth and inflation in the US, the UK and the euro area.

We have some knowledge of the relative point forecasting performances of DSGE models relative to Bayesian VARs (as, for example, Smets and Wouters (2007)), of FADL relative to factor-augmented MIDAS models (Andreou et al., 2013), and of Bayesian VARs relative to dynamic factor models (Bańbura et al., 2010). This paper advances further by comparing the out-of-sample forecasting accuracies for point and density forecasts of output growth and inflation for the following classes of models: BVAR, FADL, MIDAS and DSGE models.

The design of our forecasting comparison with the elements described above implies that we are evaluating the forecasting performances of 13 reduced-form model specifications for predicting two quarterly macroeconomic time series over horizons from one to eight quarters ahead. We perform this comparison for seven different countries and consider four different subperiods of five years over a 20-year out-of-sample period. In order to get clear messages from our empirical exercise, we develop evaluation methods that pool forecasting performances across countries, model classes, forecasting origin periods and dataset sizes.

Our meta-analysis method employs a regression of the relative performance of each multivariate reduced-form

[2] For example, Faust and Wright (2013) provide evidence that time-varying vector autoregressive models with stochastic volatility do not improve point forecasts of inflation in comparison with a univariate benchmark, although there is stronger evidence that stochastic volatility improves density forecasts of inflation (Clark, 2011). Chauvet and Potter (2013) consider Markov-switching models for the prediction of output growth, and find gains only during recessions and only at short horizons. Based on data for a set of countries, Ferrara, Marcellino, and Mogliani (2015) show that nonlinear models rarely improve on the forecasts of their linear counterparts.

[3] The paper by Stock and Watson (2002) is influential in supporting the use of large datasets for forecasting macroeconomic variables. Other more recent contributions, all pointing towards the importance of using medium–large dataset for macroeconomic forecasting, include the studies by Bańbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015) and Koop (2013).

[4] Density forecasts of DSGE models are evaluated by Del Negro and Schorftheide (2013) and Diebold, Schorftheide, and Shin (2017), but when DSGE models are compared with a large set of statistical models by Faust and Wright (2013) and Chauvet and Potter (2013), only point forecasts are considered. Note also that the set of forecasting models used for predicting inflation by Faust and Wright (2013) differs from the models considered by Chauvet and Potter (2013). While Faust and Wright (2013) consider horizons up to one year ahead, Chauvet and Potter (2013) choose to look at horizons up to two quarters only, whereas Del Negro and Schorftheide (2013) evaluate horizons up to two years ahead.

[5] Those by Kuzin, Marcellino, and Schumacher (2013) and Stock and Watson (2003) are exceptions, in that they consider data from seven countries when designing their forecasting exercises. Ferrara et al. (2015) evaluate models for 19 countries, but they use only a relatively small set of predictors.

[6] Time-varying vector autoregressive models, exploited as forecasting models by D'Agostino, Gambetti, and Giannone (2013), and vector autoregressive models with stochastic volatility, the forecasting performances of which were evaluated by Clark (2011), are two classes of models that are excluded from this forecasting comparison. The main reason for this is that these two classes cannot be adapted easily to large datasets. The approach proposed by Koop and Korobilis (2013) for large datasets considers a VAR with 25 variables as "large", whereas this paper uses datasets up to 155 variables. We also use data from countries with shorter time series where structural changes are harder to identify. This paper considers just one class of mixed frequency models. Mixed frequency specifications are popular for nowcasting, as was shown in the survey by Banbura, Giannone, Modugno, and Reichlin (2013), as well as the recent contribution by Schorfheide and Song (2015). Because we aim to evaluate forecasting performances from nowcasting up to long horizons (two years), we select just one class of mixed frequency models that has a relatively good nowcasting performance (Andreou, Ghysels, & Kourtellos, 2013; Kuzin et al., 2013).

model on a set of characteristics. The relative performance is measured using the root mean squared forecast error for point forecasts and logscores for density forecasts. The performance is measured with respect to the autoregressive model for the same variable and horizon. The method allows us to assess the statistical significance of forecasting horizon, geographical source (country), model class, evaluation period and number of predictors (dataset size) in explaining forecasting performance.

A second evaluation method relies on t-statistics from a (Diebold & Mariano, 1995) equal forecast accuracy test over the 20-year evaluation period. We investigate the empirical distribution of t-statistics using an autoregressive model under the null. We use this approach to complement the results of the meta-analysis when comparing the point and density forecasting performances of specifications that use a large set of predictors with those of specifications that use a smaller set. We also use the empirical distributions of equal-accuracy t-statistics against an AR benchmark to evaluate how the forecasting accuracy of structural models compares with that of reduced-form models.

We find no support for the use of large datasets (100 predictors) rather than medium-sized ones (a dozen predictors). However, we provide evidence that the factor model and an equal-weighted combination of single-regressor MIDAS models are the best specifications for dealing with large datasets, since they perform better than Bayesian VARs on average. We find that DSGE models have relatively good performances for forecasting US and UK inflation at forecasting horizons of longer than one year.

The empirical results provide only limited support for the use of mixed frequency models, which exploit current-quarter information on monthly series, for improving nowcasts of output growth. The reason for this is that there is a large degree of cross-country variation in the nowcasting performances of mixed-frequency models. The results also suggest changes in the relative forecasting performances of forecasting models. The relative performances of reduced-form multivariate models are at their peak in the 1993–1997 period for inflation and in the 2008–2011 period for output growth.

We describe the classes of forecasting models in Section 2. Section 3 provides a summary of the datasets that we employed, which are reported fully in our online appendix. Section 4 describes the key elements of the design of our forecasting exercises, including the statistical tests employed. Section 5 explores the key determinants of the point and density macroeconomic forecasting performances of multivariate statistical models relative to those of AR models using meta-analysis regressions and the empirical distribution of equal-accuracy t-statistics. Evaluations of the point and density forecasting accuracies of structural models in comparison to reduced-from models are discussed in Section 6. Section 7 concludes.

2. Forecasting methods

This section describes the forecasting methods compared in this paper. In contrast to the recent evaluations of the forecasting of output and inflation by Chauvet and Potter (2013) and Faust and Wright (2013) , respectively, we use the same set of forecasting model classes for predicting output growth and inflation. The advantage of this approach is that it allows us to evaluate whether we need different forecasting models for output and inflation. The disadvantage is that we do not evaluate forecasting methods that were designed for certain specific features of each variable, such as the UCSV models for inflation (Stock & Watson, 2007) and Markov-switching models for output (Chauvet, 1998). Another important feature of our forecasting exercise is that we consider both point and density forecasts. This density forecasting evaluation provides us with insights on the accuracy of forecasting models for the whole predictive distribution. The advantage of considering both point and density forecasts is that we can assess whether the choice of the loss function has an impact on model rankings.

The remainder of this section describes how we compute density forecasts of three reduced-form forecasting models: factor models, Bayesian VAR models and MIDAS models. We also describe how we obtain density forecasts using a structural DSGE model and simple univariate models.

In the text below, we use the following notation: Q_t, for $t = 1, \ldots, T$, denotes the raw data; and $q_t = \log(Q_t)$ denotes the time series in log-levels. The variable in first differences is $\Delta q_t = 100*(q_t - q_{t-1})$. The forecast horizon is h, and the maximum forecast horizon is $h_{\max}$.

2.1. Univariate models

We compute forecasts from univariate autoregressive (AR(p)) models. The autoregressive order is selected using the Schwarz information criterion (SIC) and assuming a maximum order of four. We compute the predictive density by bootstrap as per Clements and Taylor (2001). First, we get a full bootstrapped time series $\Delta q^*_{p+1}, \ldots, \Delta q^*_T$ by using the OLS estimates, initial values $\Delta q_1, \ldots, \Delta q_p$ and a $T - p$ bootstrapped time series from the residuals. Using the bootstrapped time series, we estimate an AR(p) model with the same autoregressive order as the original model. Then, we compute forecasts by iteration for $h = 1, \ldots, h_{\max}$, including a bootstrap draw from the residuals for each horizon. This bootstrap procedure will deliver sequential draws as $\Delta\hat{q}^{(i)}_{T+1}, \ldots, \Delta\hat{q}^{(i)}_{T+h_{\max}}$ for each time we reestimate the model on a new bootstrapped sample.

2.2. Factor models

We forecast with factors using the following FADL(p, k) equation for each horizon h:

$$\Delta q_t = \beta_0 + \sum_{i=0}^{p-1} \beta_{i+1} \Delta q_{t-h-i} + \sum_{j=1}^{r} \sum_{i=0}^{k-1} \gamma_{j,i+1} f_{j,t-h-i} + \varepsilon_t, \quad (1)$$

where r counts the number of factors f.

Factors are estimated by principal components applied to either a medium-sized (around 14 variables) or large (around 100 variables) dataset of predictors of q_t. Before performing factor estimation, we decide whether to transform raw data to log-levels as is described in the "log vs. level" column in Tables B2 and B3 in the online appendix. Then we apply ADF unit root tests to define the order of differentiation of each variables. Next, principal components is applied to standardized data to compute the factors. We follow Groen and Kapetanios (2013) in choosing the number of factors. We first choose the autoregressive order p in a univariate regression using the SIC, then we set $k = 1$ to choose the number of factors using the modified SIC of Groen and Kapetanios (2013), assuming a maximum number of factors of four. We have also tried to choose r and k jointly using the modified SIC, and normally $k = 1$ is the choice indicated, with the impact on the average forecasting performance being negligible even when k should be larger.

We compute density forecasts from the FADL model by a fixed regressor bootstrap. We choose this specific approach because it takes into account both parameter and forecasting uncertainties when computing density forecasts, and because we will apply a similar approach, based on (Aastveit, Foroni, & Ravazzolo, 2016), for computing density forecasts with MIDAS models. This implies that we fix the variables on the right-hand-side (RHS) of the regression to their data values, and use bootstrapped values from the residuals to get a full bootstrapped time series $\Delta q^*_{p+1}, \ldots, \Delta q^*_T$ for the left-hand-side (LHS).[7] We then re-estimate the ADL regression using the bootstrapped LHS values and the fixed RHS values. Using bootstrapped coefficients, we compute a forecast draw $\Delta\hat{q}^{(i)}_{T+h}$, conditional on observed values for ..., Δq_{T-1}, Δq_T, and using a bootstrap draw from the reestimated regression residuals. Note that this bootstrapping procedure will deliver the density for one specific forecasting horizon. Our factor modelling approach requires the estimation of a forecasting model for each horizon.

2.3. MIDAS models

The economic predictors in our dataset, summarized in Table 2, are sampled monthly. The factor approach described above requires the aggregation of monthly data into quarters. We exploit monthly information directly by employing an ADL-MIDAS model. The model is written as:

$$\Delta q_t = \beta_0 + \sum_{i=0}^{p-1} \beta_{i+1} \Delta q_{t-h-i} + \gamma \sum_{i=0}^{km-1} w(\theta, i) x_{t-mh-i+l} + \varepsilon_t,$$

where m is the difference in sampling frequencies between q_t and x_t, and $w(\theta, i)$ are the weights for each high frequency lag, which are functions of the parameters θ. In our applications, $m = 3$, since x_t is sampled monthly while q_t is sampled quarterly. The autoregressive order in quarters is denoted by k, and km is the autoregressive order in months such that lags of x are counted in months. The number of lead months is represented by l (named as per Andreou et al. (2013), though it was first employed for macroeconomic forecasting by Clements and Galvão (2008)). The intuition on the use of leads is that forecasts for current and future quarters are computed conditional on monthly observations of economic indicators during the current quarter. In the forecasting exercise, we set $l = 2$ for all h. This implies that we are considering typical nowcasting horizons if $h = 1$. This utilization of monthly data is the main advantage of the MIDAS approach for macroeconomic forecasting (Andreou et al., 2013; Clements & Galvão, 2008; Kuzin et al., 2013).

We measure the impact of the high frequency x_t on the low frequency q_t by first applying the weights $w(\theta, i)$ to all monthly lags, then multiplying by an intercept γ, which is identified because the weights sum to one. We use the beta function to obtain the weights, that is,

$$w(\theta; i) = \frac{f(\theta; i)}{\sum_{j=1}^{K} f(\theta; j)}$$

$$f(\theta; i) = \frac{(j)^{\theta_1 - 1}(1-j)^{\theta_2 - 1}\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}, \qquad j = i/km.$$

The two parameters in θ are estimated jointly with the other parameters by nonlinear least squares. Note that, as in the case of the factor approach, we need to estimate a MIDAS regression for each forecasting horizon.

We compute density forecasts using a fixed regressor, bootstrapped as per Aastveit et al. (2016) and as described in Section 2.2. Our application of the fixed regressor bootstrap to MIDAS models implies that we also fix θ, that is, take $\theta = \hat{\theta}$ from the estimation with observed data, but obtain different values of β_i and γ for each bootstrapped sample. This has a large beneficial impact on our computational burden. Our density computation strategy is still able to capture the impact of parameter uncertainty on a set of parameters while computing forecasts. Note that, as in the case of factor models, the last step in computing $\Delta\hat{q}^{(i)}_{T+h}$ also requires a draw from the residuals of the re-estimated MIDAS regression.

We consider two different types of MIDAS specifications that are able to deal with large datasets. The first one assumes that x is an individual predictor. Because we plan to employ sizeable datasets, we estimate a single regressor MIDAS model for each predictor, then combine their predictive densities using equal weights. We call this model the combination MIDAS (C-MIDAS) model. In this specification, we decide beforehand whether we will be using log, log-levels or quarterly differences for each of the indicators when using our medium dataset. Our choice of data transformation is indicated in Tables B2 and B3 in the online appendix.

The second specification first estimates factors with monthly data by principal components, applying the data transformation based on unit root tests described for FADL models. We then set the number of factors to one in the case of medium datasets and to two in the case of large datasets following Andreou et al. (2013). We call this specification the F-MIDAS model, and the regressors x_t are the factors estimated in a previous step by principal components.

2.4. BVAR models

Our BVAR approach is the benchmark model of Carriero et al. (2015), who provide a summary of the literature on the application of BVARs to forecasting. Define the

[7] As a consequence, this approach does not take into account the uncertainty in the estimation of the factors, but only in the β_s and γ_s.

vector $y_t = (q_{1t}, q_{2t}, \dots, q_{Nt})'$; then, a VAR($p$) is:

$$y_t = A_0 + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t \qquad (2)$$
$$\varepsilon_t \sim N(0, \Sigma),$$

for $t = p+1, \dots, T$.

We elicit a conjugate normal-inverse Wishart prior:

$$\alpha|\Sigma \sim N(\alpha_0, \Sigma \otimes \Omega_0)$$
$$\Sigma \sim IW(S_0, v_0),$$

where $\alpha = vec([A_c, A_1, \dots, A_p]^{'})$, so that the posterior distributions are

$$\alpha|\Sigma, data \sim N(\overline{\alpha}, \Sigma \otimes \overline{\Omega})$$
$$\Sigma|data \sim IW(\bar{S}, \bar{v}).$$

Carriero et al. (2015) describe the closed-form solutions for the posterior and prior means and variances under the assumption that they follow a Minnesota-style prior as Bańbura et al. (2010). We consider the prior means for the first-order autoregressive coefficients as equal to one if the endogenous variables, y_t, are in log-levels as described above. We also consider a specification in differences, using Δy_t, with the prior mean equal to zero.

In the case of the VAR in levels, we also impose the sum of coefficients prior, which expresses the belief that the average of the past values of a given variable provides a good forecast for that variable. The fact that, in the limit, the sum-of-coefficients prior is not consistent with cointegration motivates the use of an additional prior, known as the 'dummy initial observation' prior. This was proposed by Sims (1993) and avoids giving an unreasonably high explanatory power to the initial conditions, a pathology which is typical in nearly nonstationary models (Sims, 2000). These last two priors together tend to improve the forecasts when dealing with data in levels. Hyperparameters governing priors are set as the baseline case by Carriero et al. (2015). The overall prior tightness λ_1 is selected to maximise the marginal likelihood:

$$\lambda_1 = \arg\max_{\lambda_1} \ln(p(Y)),$$

where $p(Y)$ is computed in closed form as per Carriero et al. (2015). The grid has 15 elements [0.01, 0.025, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.75, 1, 2, 5]. In an out-of-sample forecasting exercise, we compute λ_1 each time we re-estimate the model with a longer sample period.

Forecasts are computed by simulation. We use posterior draws of α and Σ to obtain a implied path for $\hat{y}_{T+1}, \dots, \hat{y}_{T+h}$. Assume that $\mathbf{A} = [A_c, A_1, \dots, A_p]^{'}$, which is an $N \times Np + 1$ matrix; then, we obtain a draw j for all autoregressive coefficients using:

$$(\mathbf{A}^{(j)}) = (\overline{\mathbf{A}}) + chol(\overline{\Omega}^{(j)}) * V^{(j)} * chol(\mathbf{\Sigma}^{(j)})',$$

where $V^{(j)}$ is an $(Np+1) \times N$ matrix obtained from a standard normal distribution. Then, for a draw of $\mathbf{A}^{(j)}$ and $\mathbf{\Sigma}^{(j)}$, we take a sequence of h draws from the $N(0, \mathbf{\Sigma}^{(j)})$ in order to compute by iteration a sequence of forecasts $\hat{y}_{T+1}, \dots, \hat{y}_{T+h}$ for the model in Eq. (2). We use a total of 5000 draws, and split the procedure such that we use a small number of draws of $\mathbf{A}^{(j)}$ and $\mathbf{\Sigma}^{(j)}$, then generate many sequences of forecasts for each parameter draw. The point forecast is the median over all draws for each horizon.

We consider specifications both in levels, which we call L-BVAR, and in differences, called D-BVAR. We set $p = 4$. When the target forecasting variable is the quarterly growth rate, we transform the forecasts for the model in levels accordingly.

2.5. DSGE models

The literature provides evidence of the accuracy of the medium-sized (Smets & Wouters, 2007) model (Christoffel, Coenen, & Warne, 2010; Del Negro & Schorftheide, 2013; Edge & Gurkaynak, 2011; Woulters, 2015). We employ the Smets-Wouters DSGE model with seven observables, including output and inflation, as our structural model. We use the specification as Herbst and Schorftheide (2012) and Smets and Wouters (2007), which assume a deterministic trend to productivity.

We use the priors as Herbst and Schorftheide (2012) and Smets and Wouters (2007). The posterior distribution of the structural parameters is obtained using the random walk Metropolis algorithm described by Del Negro and Schorftheide (2011), and we calibrate the spread parameter such that the acceptance rate is in the 20%–40% range for each country's dataset. We compute the predictive density using 5000 equally-spaced draws from the posterior parameter draws generated by the MCMC procedure. For each parameter draw, we also draw from the normal distribution of the disturbances (structural shocks) to get a sequence of forecasts from $h = 1, \dots, h_{\max}$ for each observed variable.

We compute forecasts with DSGE models for only three economies in our dataset: the US, the UK and the euro area. The reason for this is that the assumption in the model that the central bank that sets interest rates based on a Taylor rule, which depends on domestic inflation, is not adequate for countries which are part of the euro area. We also choose not apply it to Japan, again because the Taylor rule may be a very poor approximation of the Bank of Japan's monetary policy over the last 20 years. We apply the model to euro area data by adding an equation that links employment to hours such that we can use the employment time series instead of hours, following the modification proposed by Christoffel, Coenen, and Warne (2008).

3. Data description

We employ data from seven developed economies: US, UK, the euro area, Germany, France, Italy and Japan. Our target variables are the quarterly change in log real GDP and the quarterly change in seasonally-adjusted log CPI, with the data sources being described in Table B1 in the online appendix. Seasonally-adjusted CPI data are not available for the European countries or Japan. As a consequence, we seasonally adjusted the data using the X12 filter.

For each country, we build a medium and a large dataset of economic indicators, sampled monthly. The datasets are summarized in Table 2 and described in detail in Tables B2 and B3 of the online appendix. When quarterly data are required, we use the average over quarter for factor models, so F-MIDAS nest FADL models,[8] and the end of the quarter value for the BVAR, as is popular in the BVAR literature. When possible, we follow the series included in the datasets of Kuzin et al. (2013) . The medium dataset includes 11–14 variables per country. These are a mix of measures of economic activity, including survey data, prices and financial variables. Similar sets of variables were employed by Carriero et al. (2015). These datasets include oil prices as a common variable.

The number of variables included in the large dataset varies across countries due to data availability, as is recorded in Table 2. It varies between 57 (Japan, France) and 155 (US). The large dataset also includes all variables in the medium dataset. Because of the international transmission of business cycle shocks, we include some key US variables in the large datasets of the six remaining economies, including financial variables such as equity prices and Treasury bond rates. We provide descriptions of all variables, including their datastream codes, in Table B3 of the online appendix.[9]

Due to the lack of availability of a real-time dataset for the monthly indicators for our seven countries, we use only data from the currently available vintage, as is generally the case when evaluating forecasts using models for large datasets (such as Kuzin et al. (2013) and Smets and Wouters (2007) , for example).

DSGE models are estimated using the quarterly growth rate in output per capita. They also use inflation as measured by the GDP deflator. As a consequence, when evaluating the forecasts of the DSGE models, we change the target variable to the growth in output per capita and the quarterly GDP deflator inflation. We then reestimate the forecasting models for these modified target variables for a subset of our reduced-form models to enable us to compare the predictions of structural and reduced-form models. Table B4 in the online appendix describes the variables employed in the DSGE estimation, including their required transformations.

The last observation employed in our forecasting exercise is 2013M9. For the US, Japan and the UK, we use data from 1975M1 (with the exception of UK CPI inflation, which is available only from 1980M1), but for other countries, data are only available later, as is described in Table 2. The data for DSGE estimation are from 1984Q1 for the US, the UK and the euro area.

4. Evaluation design

Our first forecast origin is 1993Q1 for the US, the UK, Japan and France; 1998Q1 for Germany and Italy; and 2003Q1 for the euro area. We set the maximum forecast horizon to eight, enabling us to compute measures of the forecast accuracy for forecasts up to 2011Q3; that is, we have 75 observations in our out-of-sample period for the US, the UK, Japan and France, 55 observations for Germany and Italy, and 35 observations for the euro area. For some of our results, we split the out-of-sample period into windows of five years (20 observations), based on the forecast origin date, to verify whether the relative forecasting performance varies over the out-of-sample period. The literature provides evidence that the predictive ability may change over time (Giacomini & Rossi, 2010). In addition, changes in the underlying structure of the economy and the data characteristics may affect the models' relative forecasting performances.

Table 1
Model acronyms.

m	Name	Description
1	AR	Autoregressive model
2	FADL_M	Factor ADL model with a medium-sized dataset
3	FADL_L	Factor ADL model with a large dataset
4	F-MIDAS_M	Factor MIDAS with a medium-sized dataset
5	F-MIDAS_L	Factor MIDAS with a large dataset
6	C-MIDAS_M	Combination MIDAS with a medium-sized dataset
7	C-MIDAS_L	Combination MIDAS with a large dataset
8	L-BVAR_S	BVAR in levels with a small dataset
9	D-BVAR_S	BVAR in differences with a small dataset
10	L-BVAR_M	BVAR in levels with a medium-sized dataset
11	D-BVAR_M	BVAR in differences with a medium-sized dataset
12	L-BVAR_L	BVAR in levels with a large dataset
13	D-BVAR_L	BVAR in differences with a large dataset
14	DSGE	Smets and Wouters' (2007) medium-sized DSGE model

We compute forecasts from models estimated with expanding samples over the out-of-sample period, that is, we re-estimate each model at each forecast origin and use all observations available up to the forecasting origin.

We use two measures of the forecasting performance. The accuracy of the point forecasts is measured using root mean squared forecast errors (RMSFE), while the log predictive score measures the accuracy of the density forecasts. The advantage of using log scores to compare density forecasts is that the maximization of the logscore is equivalent to the minimization of the Kullback–Leibler distance between the model and the true density. We compute log scores by first fitting a Gaussian kernel density to the 5000 predictive density draws over a grid between −15 and 15, then finding the probability at the outturn.

We use the Diebold and Mariano (1995) t-statistic to test for equal accuracy. The variance is computed using the Newey–West estimator with the maximum order increasing with the horizon.

Table 1 provides a short description of each of the forecasting models that we employ in this evaluation. Similarly to Bańbura et al. (2010), we consider BVAR models of three sizes: small, medium and large. We use medium and large datasets for the FADL and MIDAS models, but our only small model is the BVAR. This model has only three variables: real GDP, CPI, and the short-term interest rate.

[8] This implies that F-MIDAS specification nests the FADL if the MIDAS weighting function is flat, that is, $\theta_1 = \theta_2 = 1$.

[9] Some variables were seasonally adjusted by the X12 filter before estimation, and these are marked SA in Table B3.

Table 2
Data summary.

	Country	Sample period	Out-of-sample period	Medium dataset – number of predictors	Large dataset – number of predictors
1	US	1975M1–2013M9	1993Q1–2013Q3	14	155
2	UK	1975M1–2013M9	1993Q1–2013Q3	13	59
3	Japan	1975M1–2013M9	1993Q1–2013Q3	13	57
4	France	1983M1–2013M9	1993Q1–2013Q3	13	57
5	Italy	1990M1–2013M9	1998Q1–2013Q3	12	128
6	Germany	1991M1–2013M9	1998Q1–2013Q3	13	114
7	Euro area	1998M4–2013M9	2003Q1–2013Q3	11	81

Note: Full descriptions of the time series employed, data sources and data transformations are available in the online data appendix. Table B1 describes the target variables, Table B2 describes the monthly medium-sized datasets, Table B3 reports the monthly large datasets, and Table B4 contains the descriptions and transformations of the series employed for estimating the DSGE models.

5. Explaining the forecasting performances of statistical models

We provide acronyms for each of the forecasting models included in this evaluation in Table 1. They comprise 13 reduced-form models, including a univariate model (AR), and one structural model (DSGE). This section explores the relative forecasting performances of the 12 multivariate reduced-form models, listed as models 2 to 13 in Table 1. Forecasting comparisons that include the DSGE model are discussed in Section 6. We measure the impacts of the model class, forecasting horizon, dataset size and data source (country) on the point and density forecasting performances.

5.1. A meta-analysis

Our aim is to investigate how the relative (to the AR model) forecasting performance of each statistical model class (MIDAS, FADL and BVAR) varies with the number of predictors (medium vs. large dataset), the forecasting horizon (nowcasting, short horizon ($h = 2, \ldots, 4$) and medium horizon ($h = 5, \ldots, 8$)), the 5-year subperiod evaluated, and the geographical source of the dataset.

The dependent variable in our meta-analysis regression is a measure of the forecasting performance of a specific forecasting model relative to that of the autoregressive model when predicting one of the target variables (output growth and inflation) for a specific country, horizon and forecasting origin period. The measures of the forecasting performance are based on root mean squared forecast errors (RMSFE) and the median logscore (MLS)[10] computed for a specific target variable that varies across countries, forecasting models, periods and horizons. The measures for point and density forecasting performances are:

$$rMSFE_{m,p,c,h} = \frac{RMSFE_{AR,p,c,h}}{RMSFE_{m,p,c,h}}$$

$$rMLS_{m,p,c,h} = 1 + [(-MLS_{ar,p,c,h}) - (-MLS_{m,p,c,h})],$$

where $m = 2, \ldots, 13$, which are the statistical models numbered 2 to 13 in Table 1. Each measure varies with the set of forecasting origins employed in the computation, $p =$ 1993Q1–1997Q4, 1998Q1–2002Q4, 2003Q1–2007Q4, 2008Q1–2011Q3, 1993Q1–2011Q3; with the source country $c =$ US, UK, EU, FR, IT, GER, JP; and with the forecasting horizon $h = 1, \ldots, 8$.

As consequence, the total number of relative performance observations (given that the forecasting period availability varies across countries as is noted in Table 2) is 2976. By exploiting a large set of forecasting comparisons, we aim to find sources of performance improvements in macroeconomic forecasting that are not constrained by the model class, forecast horizon, country or evaluation period.

The first characteristic that we explore is the country in which the data are sourced. We use two dummy variables to split the country set in Table 2 into three: $D^{EU} = 1$ for euro area countries ($c =$ EU, FR, IT, GER) (and $D^{EU} = 0$ otherwise), and $D^{JP} = 1$ if $c =$ JP. Thus, the benchmark countries are the US and the UK.

The second characteristic is the forecasting horizon. We split the set of forecast horizons into three groups by defining $D^{sh} = 1$ if $h = 2, \ldots, 4$ and $D^{mh} = 1$ if $h = 5, \ldots, 8$. Accordingly, differences in performance over short and medium horizons are assessed against the nowcasting ($h = 1$) benchmark.

We are also interested in finding differences among the three model classes. We set $D^{MIDAS} = 1$ if $m = 4, 5, 6, 7$ and $D^{BVAR} = 1$ if $m = 8, 9, 10, 11, 12, 13$, based on the description in Table 1. The benchmark model class is the FADL ($m = 2, 3$). The impact of the number of predictors is evaluated using $D^{small} = 1$ if $m = 8, 9$ and $D^{large} = 1$ if $m = 3, 5, 7, 12, 13$, implying that the benchmark dataset size is the medium one.

Finally, the impact of the evaluation period is assessed by creating a dummy variable for each of the four five-year out-of-sample subperiods. As a consequence, performance improvements are relative to the full out-of-sample period ($p =$ 1993Q1–2011Q3).

We also consider interactions among the dummy variables described above. We consider interactions between horizon and model class dummies, between D^{large} and model class dummies, and between D^{large} and evaluation period dummies.

[10] We use the median rather than the mean logscore in order to minimize the impact of outliers in our analysis. Outlier values are more frequent with logscores than with squared forecast errors.

Table 3
Explaining relative forecasting performances by country, forecasting origin period, horizon, model class and dataset size.

	rMSFE		rMLS	
	Output growth	Inflation	Output growth	Inflation
const (FADL_M, $h = 1$	**1.037*****	**1.017*****	**1.041*****	**1.057*****
1993–2011, US+UK)	(0.026)	(0.015)	(0.042)	(0.054)
Japan	**0.016***	**0.068*****	0.011	−**0.022**
	(0.009)	(0.004)	(0.028)	(0.057)
Euro area	−**0.038***	−0.012	−**0.100**	−**0.066**
(Euro, GER, IT, FR)	(0.015)	(0.028)	(0.064)	(0.068)
1993Q1–1997Q4	−**0.068**	**0.077**	−0.001	**0.068**
	(0.043)	(0.047)	(0.025)	(0.044)
1998Q1–2002Q4	−**0.037*****	**0.022**	**0.022**	−0.016
	(0.008)	(0.036)	(0.031)	(0.027)
2003Q1–2007Q4	0.003	0.006	0.001	−0.005
	(0.009)	(0.032)	(0.024)	(0.046)
2008Q1–2011Q3	**0.039****	0.013	**0.068**	−0.008
	(0.020)	(0.025)	(0.056)	(0.039)
$h = 2,\ldots,4$	−**0.040***	−0.012	−0.002	−0.028
	(0.023)	(0.025)	(0.030)	(0.019)
$h = 5,\ldots,8$	−**0.028**	−**0.035**	−**0.065**	−**0.057***
	(0.028)	(0.039)	(0.063)	(0.035)
MIDAS ($h = 1$)	0.029	−0.002	0.011	0.012
	(0.029)	(0.039)	(0.025)	(0.047)
BVAR ($h = 1$)	−0.013	−0.011	0.004	−**0.098***
	(0.017)	(0.031)	(0.039)	(0.057)
$(h = 2,\ldots,4)$*MIDAS	−0.021	−0.033	−0.035**	−**0.051**
	(0.027)	(0.032)	(0.018)	(0.043)
$(h = 5,\ldots,8)$*MIDAS	−**0.060****	−0.006	−0.048**	−0.024
	(0.026)	(0.025)	(0.021)	(0.049)
$(h = 2,\ldots,4)$*BVAR	**0.058*****	−0.013	−0.025	−0.007
	(0.020)	(0.026)	(0.024)	(0.037)
$(h = 5,\ldots,8)$*BVAR	0.024	0.006	−0.019	0.028
	(0.026)	(0.028)	(0.057)	(0.055)
Small	−**0.031***	**0.027**	−0.005	0.002
	(0.019)	(0.025)	(0.015)	(0.034)
Large	−0.003	−**0.048*****	−0.011	−0.019
	(−0.003)	(0.018)	(0.007)	(0.016)
Large* BVAR	−**0.029***	−**0.070***	−**0.163*****	−**0.201*****
	(0.017)	(0.042)	(0.055)	(0.041)
Large*MIDAS	0.001	0.021	−0.013	−0.030*
	(0.014)	(0.023)	(0.026)	(0.018)
Large*(1993–1997)	−**0.031**	**0.127**	**0.037**	**0.121*****
	(0.024)	(0.081)	(0.038)	(0.041)
Large*(1998–2002)	−**0.021****	−**0.062**	0.007	−0.040
	(0.010)	(0.048)	(0.024)	(0.027)
Large*(2003–2007)	−0.002	0.012	−0.010	−0.021
	(0.007)	(0.014)	(0.015)	(0.015)
Large*(2007–2011)	0.008	−0.003	0.009	0.001
	(0.014)	(0.021)	(0.022)	(0.026)
R^2	0.154	0.126	0.156	0.198
No. of obs.	2976	2976	2976	2976
Mean of dep. var.	0.978	0.986	0.929	0.888

Note: Values larger than one imply that model improves over the AR. All explanatory variables are dummy variables. The regressions are estimated by OLS. The values in brackets are standard errors clustered by country. Values in bold denote estimates that are statistically significant at the 10% level if we use heteroscedasticity-consistent (White) standard errors.

*Indicate rejection of the null of no statistical significance at 10% level.

**Indicate rejection of the null of no statistical significance at 5% level.

***Indicate rejection of the null of no statistical significance at 1% level.

Table 4
Additional regressions.

A: C-MIDAS vs F-MIDAS with observations for MIDAS specifications only				
	rMSFE		rMLS	
	Output growth	Inflation	Output growth	Inflation
C-MIDAS	**0.048***** (0.006)	**0.078**** (0.034)	**0.041***** (0.011)	**0.129***** (0.028)
R^2	0.031	0.037	0.006	0.048
No. of obs.	992	992	992	992
Mean of dep. var.	0.971	0.991	0.941	0.940
B: D-BVAR vs L-BVAR with observations for BVAR specifications only				
	rMSFE		rMLS	
	Output growth	Inflation	Output growth	Inflation
D-BVAR	**−0.037*** (0.021)	**−0.050** (0.056)	−0.008 (0.070)	**0.069** (0.093)
R^2	0.022	0.015	0.001	0.020
No. of obs.	1488	1488	1488	1488
Mean of dep. var.	0.981	0.977	0.904	0.824

Note: The regressions are estimated by OLS. The values in brackets are standard errors clustered by country. Values in bold denote estimates that are statistically significant at the 10% level if we use heteroscedasticity-consistent (White) standard errors.
*Indicate rejection of the null of no statistical significance at the 10% level.
**Indicate rejection of the null of no statistical significance at the 5% level.
***Indicate rejection of the null of no statistical significance at the 1% level.

The meta-analysis regression is then:

$$\begin{aligned} rLoss_{m,p,c,h} = {} & \beta_0 + \beta_1 D^{JP} + \beta_2 D^{EU} \\ & + \beta_3 D^{9397} + \beta_4 D^{9802} + \beta_5 D^{0307} + \beta_6 D^{0811} \\ & + \beta_7 D^{sh} + \beta_8 D^{lh} + \beta_9 D^{MIDAS} + \beta_{10} D^{BVAR} \\ & + \beta_{11} D^{MIDAS} * D^{sh} \\ & + \beta_{12} D^{MIDAS} * D^{lh} + \beta_{12} D^{BVAR} * D^{sh} \\ & + \beta_{14} D^{BVAR} * D^{lh} \\ & + \beta_{15} D^{small} + \beta_{16} D^{l\arg e} + \beta_{17} D^{l\arg e} * D^{BVAR} \\ & + \beta_{18} D^{l\arg e} * D^{MIDAS} \\ & + \beta_{19} D^{l\arg e} * D^{9397} + \beta_{20} D^{l\arg e} * D^{9802} \\ & + \beta_{21} D^{l\arg e} * D^{0307} + \beta_{22} D^{l\arg e} * D^{0811} \\ & + \varepsilon_{m,p,c,h}. \end{aligned} \qquad (3)$$

$$\begin{aligned} \text{for } m &= 2, \ldots, 13;\; p = 1993\text{–}1997,\ 1998\text{–}2002, \\ & 2003\text{–}2007,\ 2008\text{–}2011,\ 1993\text{–}2011; \\ h &= 1, \ldots, 8; \\ c &= \text{US, UK, JP, FR, IT, GER, EU}; \end{aligned}$$

where $rLoss_{m,p,c,h}$ is either $rMSFE_{m,p,c,h}$ or $rMLS_{m,p,c,h}$.

Note that β_0 measures the relative (to the AR model) performance of the FADL medium model ($m = 2$) for $h = 1$ over the full sample period ($p =$ 1993–2011) with US and UK data ($c = 1, 2$). As a consequence, all other coefficient estimates are measures of gains/losses relative to this benchmark.

5.2. Meta-analysis results

Table 3 presents estimates of the regression in Eq. (3) with standard errors clustered by country, implying that we consider country-specific effects. The table columns describe the results for each performance measure (*rMSFE* and *rMLS*) and target variable (output growth, inflation). Cases in which the null hypothesis that the coefficient is equal to zero is rejected are indicated with stars for the 10%, 5% and 1% significance levels. Values in bold show that the estimates are statistically significant at the 10% level when using heteroscedasticity-robust standard errors instead of the country-clustered standard errors displayed in Table 3.

The characteristics considered in Eq. (3) explain between 13% and 20% of the forecasting performance, depending on the target and the type of performance measure. As a consequence, idiosyncratic variation plays an important role in explaining forecasting performances across this large number of forecasting exercises. The following analysis will consider characteristics that have a statistically significant role in explaining the forecasting performance, as indicated in Table 3.

The estimates of the regressions' intercepts are all larger than one, implying that the FADL_M improves on the AR on average when nowcasting US and UK variables. The gains for output growth are larger, and imply a 4% improvement in RMSFE. The estimates for β_1 and β_2 suggest that the benefits of employing multivariate models for predicting output growth rather than AR models are larger with Japanese data but smaller with European data.

The estimated coefficients on the evaluation period dummies point to changes in the statistical performances over time, but the estimates are statistically significant with country-clustered standard errors only when evaluating output growth point forecasts. We find that multivariate models perform relatively better for output growth during the turbulent 2008Q1–2011Q3 period, but relatively worse in the 1998Q1–2011Q3 period.

The estimated coefficients on the forecasting horizon dummies are all negative, implying that the performances of multivariate models relative to the AR model deteriorate with the horizon. This deterioration is statistically significant for point forecasts of output growth and inflation when the horizon is iterated with the MIDAS model dummy variable. This decline in MIDAS forecasting performance with the horizon is compensated in part by the fact that MIDAS models' RMSFEs improve on the benchmark by 3% on average when nowcasting output growth, albeit the estimate of β_9 is not statistically significant. For predicting output growth, BVAR models do relatively better at medium horizons ($h = 5, \ldots, 8$) and are significantly better at $h = 2, 3, 4$. These results suggest that although MIDAS models may deliver accurate nowcasts of output growth for some countries, the performance of this class of models deteriorates rapidly with the forecast horizon, meaning that a BVAR specification may be a more accurate choice in some cases.

The estimated coefficients on the dataset-size dummies indicate that BVARs with only three variables, including both targets, are significantly worse for predicting output growth than models with a moderate number of indicators. For predicting inflation, models with either a small or a medium set of indicators perform significantly better than large datasets.

The interactions between the dataset size and model class clearly indicate that large BVAR models lead to a deterioration in forecasting performance. These results suggest that models with factors (FADL and F-MIDAS) or forecasting combinations (C-MIDAS) are more adequate than BVAR models if the aim is to exploit the information in a large number of predictors (more than 55 indicators) for forecasting output growth and inflation. However, there is no evidence that the use of a large number of predictors instead of a dozen picked variables (medium dataset) improves macroeconomic forecasting. By evaluating the estimates for the iterations between D^{large} and the sample period, we find that using a large set of predictors worsens the output growth point forecasting performance in the earlier periods when the sample sizes employed in the estimation are shorter (recall that we increase the sample size when estimating models at each forecasting origin).

In summary, we find some time variation in the relative forecasting performances of multivariate statistical models to AR models for forecasting output growth across countries: multivariate models are of particular use during the last four-year period (2008–2011). We find no evidence that models with larger numbers of predictors improve on the performances of models with smaller sets of predictors. If using a large dataset, FADL and MIDAS models are more adequate than BVAR models. We find very limited evidence that MIDAS models improve nowcasts.

5.3. Additional meta-analysis comparisons

We consider two main specification types for MIDAS and BVAR model classes. For MIDAS models, we compute forecasts using both a factor-augmented version (F-MIDAS) and an equal-weight forecasting combination strategy (C-MIDAS). For BVAR models, we use one specification in levels (L-BVAR) and another in growth rates (D-BVAR). This subsection uses relative performance regressions to test whether there are any statistical differences in performance between these specification types that hold across countries, horizons, evaluation periods and numbers of predictors.

Table 4A presents results for the four performance measures in Table 3 (output growth and inflation; *rMSFE* and *rMLS*). These are single regressions that are estimated using performance measures computed only for MIDAS models ($rLoss_{m,p,c,h}$ for $m = 4, \ldots, 7$ and with p, h and c varying as in Eq. (3)). We define the dummy variable D^{CMIDAS} as equal to 1 if $m = 6, 7$. As a consequence, if the estimated coefficient of D^{CMIDAS} is significantly positive, we can conclude that the equal-weighted forecasting combination of single-regressor MIDAS models is a better way of exploiting the information in a set of predictors than using monthly factors. The coefficients are indeed positive and statistically significant with country-clustered standard errors in all columns of Table 4A, so we conclude in favour of the C-MIDAS specifications.

Table 4B computes single regressions with the same performance measures, but for BVAR models only ($rLoss_{m,p,c,h}$ for $m = 8, \ldots, 13$ and with p, h and c varying as in Eq. (3)). We define the dummy variable D^{DBVAR} as equal to 1 if $m = 9$, 11, 13 and zero otherwise. The empirical results can inform us as to whether the BVAR-in-differences improves over the BVAR-in-levels. Recall that the main advantage of using the BVAR-in-levels (L-BVAR) is that it allows for the possibility of cointegration. The results in Table 4B suggest that this BVAR specification choice matters only for point forecasting of output growth: L-BVARs perform significantly better than D-BVARs.

5.4. Evaluating the impact of the dataset size with equal accuracy tests

Our previous results suggest that the use of forecasting models with large sets of predictors may have a negative effect on forecasting performances for both output growth and inflation, particularly if using BVAR models with short samples. This subsection evaluates this research question using the empirical variation of "medium vs. large" equal accuracy tests for point and density forecasts as described in Section 4.

Fig. 1 presents empirical t-statistic distributions for the following models: FADL, F-MIDAS, C-MIDAS, L-BVAR and D-BVAR. The Diebold and Mariano (1995) t-statistics are computed using the specification with a medium dataset under the null and the model with a large dataset under the alternative using the full out-of-sample period ($p = 1993$–2011). The box plots are computed for t-statistics obtained for different horizons ($h = 1, \ldots, 8$) and countries. Negative values imply that the model with a large number of predictors is more accurate than the same model with a medium data set. Using a two-sided 5% test, statistical differences are found when the absolute value of the t-statistic is larger than 1.96.

In general, the t-statistics are between -1.96 and 1.96; that is, models using large and medium datasets deliver statistically similar point and density forecasting

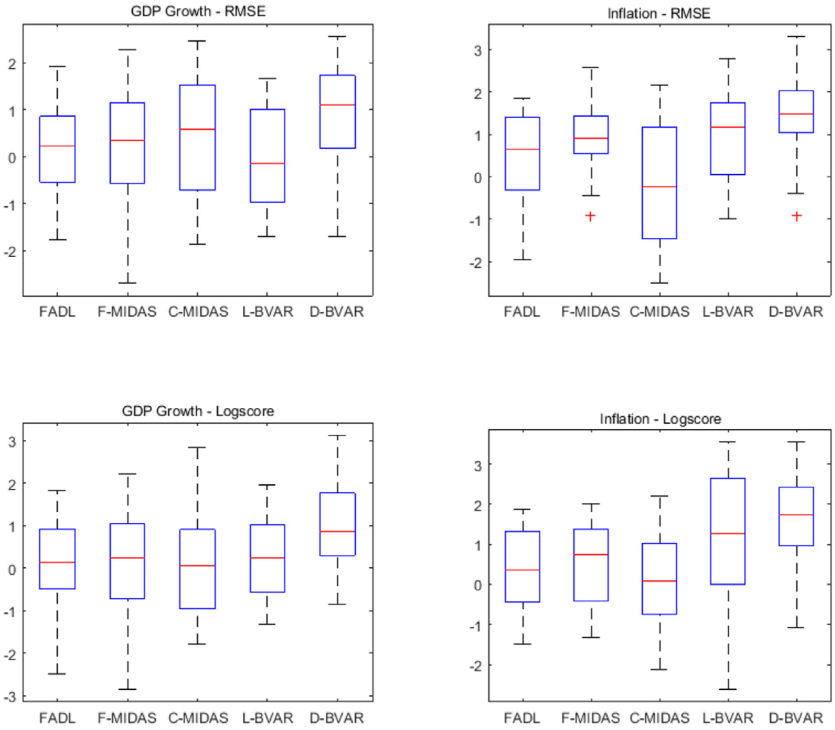

Fig. 1. Box plots of equal accuracy t-statistics for a model with a medium dataset against a large dataset for each indicated forecasting model (aggregated over forecasting horizons (1 to 8) and countries; full out-of-sample period) Note: Negative t-statistics imply that the specification with a large dataset is more accurate than the equivalent with a medium-sized dataset. Each box plot is based on 8 horizons x 7 countries = 56 values. The accuracy test in the first panel is based on MSFEs, while the statistics in the second panel are based on the logscore. "On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually". (Matlab description).

performances. However, based on the median t-statistics, we can say that D-BVARs are worse at handling large datasets than L-BVARs, providing an additional nuance to our results earlier in this section that discouraged the use of BVARs with large datasets. These results also support the use of the C-MIDAS specification instead of the F-MIDAS, in particular when dealing with large datasets for forecasting inflation.

In summary, there is no strong evidence that using a large number of predictors provides improved forecasts relative to using a moderate number, but we can provide evidence to support the use of C-MIDAS and FADL specifications instead of BVAR models when dealing with large datasets.

6. Comparing structural vs. reduced-form forecasting models

The previous section investigated common features that explain the relative forecasting performances of reduced-form statistical models across countries, forecasting horizons, forecasting periods and model specifications. This section uses equal accuracy tests, computed as described in Section 4, to compare the performances of reduced-form statistical models (FADL, BVAR, MIDAS) with those of the DSGE model.

Details of the DSGE model employed, including our estimation strategy, were discussed in Section 2.4, while Section 3 described the dataset employed in the estimation of DSGE models. One should note that medium-sized DSGE forecasts are considered only for c = US, UK and EU, and are estimated with output growth per person and GDP deflator inflation. We measure the performance of DSGE models relative to the AR benchmark by recomputing AR forecasts using the same measurements of output growth and inflation as are employed by the DSGE model.

Figures 2 and 3 present box plots of the Diebold and Mariano (1995) t-statistics. The t-statistics are computed for the full out-of-sample period for each country, as listed in Table 2. Negative values mean that the model is more accurate than the AR model. Using a one-sided test, we would reject the null of predictability at the 5% level if the DM t-statistic is smaller than −1.65. The empirical distributions vary with the country and are computed for specific model classes (FADL, MIDAS, BVAR, DSGE). The box plots are presented for three separate horizons ($h = 1, 4$ and 8). Fig. 2 presents results for output growth and inflation, using the quadratic loss function (MSFE) to compute the t-statistics, whereas the plots in Fig. 3 are based instead on the differences in logscore.

The results in Figs. 2 and 3 help us to indicate which model class, including statistical model classes (FADL, MIDAS, BVAR) and the structural model class (DSGE), performs best for each target variable and for a set of forecasting horizons. The median t-statistic in Figs. 2 and 3 can be employed to evaluate how each class of model

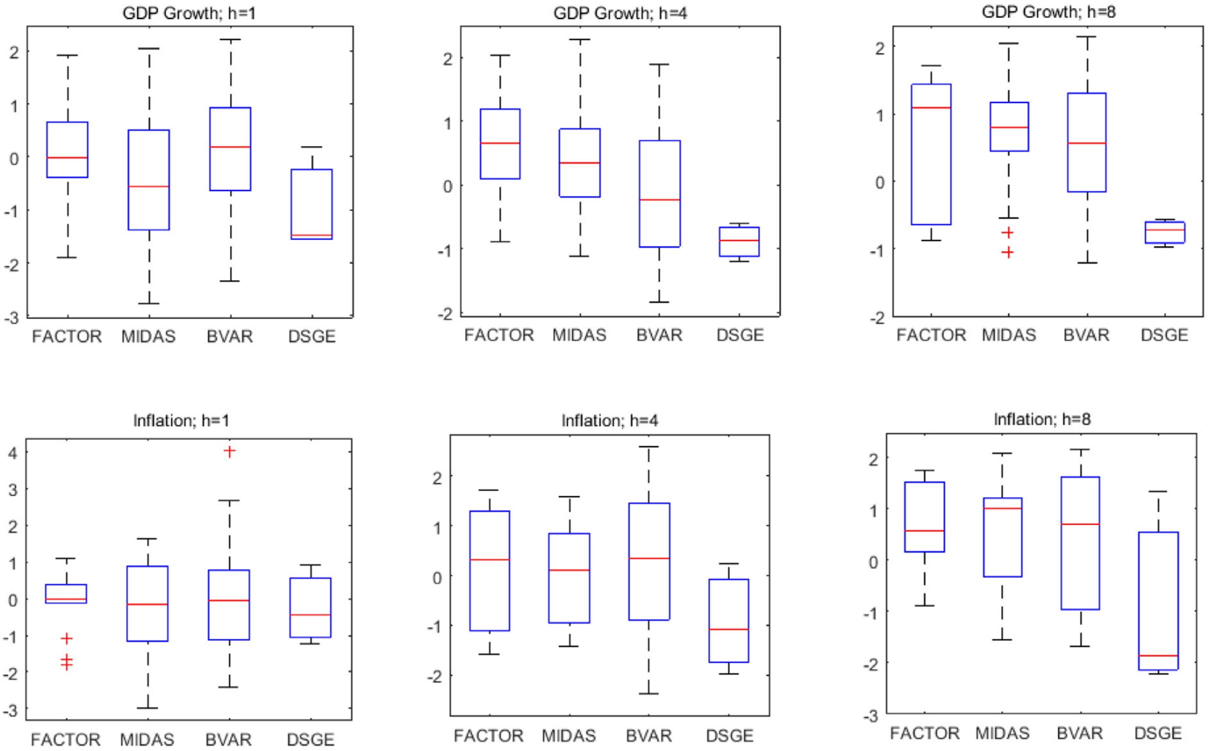

Fig. 2. Box plots of the equal accuracy MSFE t-statistics with the AR model under the null and a forecasting model from the model class indicated under the alternative (aggregated over specifications and countries; full out-of-sample period) Note: The first panel has results for forecasting the GDP growth for horizons 1, 4 and 8. The second panel has equivalent results for forecasting inflation. See Table 1 for the description of the factor (2–3), MIDAS (4–7) and BVAR (8–13) specifications. Note that the DSGE is estimated only for three countries. A list of the seven countries employed is provided in Table 2. Negative values are in favour of the specified multivariate model class. See also the notes to Fig. 1.

performs on average across specifications and countries for each horizon and target variable.

MIDAS models do better at $h = 1$ for output growth, but the distribution of t-statistics has a large spread, suggesting that mixed-frequency models improve output growth nowcasts for the median country but perform poorly for some countries. For $h = 4$, it is clear that BVARs perform better for forecasting output growth. When forecasting inflation, the clear evidence that we have is that DSGE models do better when predicting inflation at $h = 4, 8$ for both point and density forecasts. The results in Figs. 2 and 3 suggest that DSGE models are able to improve AR forecasts of quarterly inflation significantly at $h = 4, 8$.

These results are supported by detailed tables, by country and forecasting horizon, in the online appendix. Table A1 shows the relative performance of the DSGE model against those of the AR and the FADL_M using RMSFEs, while Table A2 shows results using the logscore. The results indicate that the DSGE gains for forecasting inflation are present mainly for the US and the UK, with disappointing results for the Euro area, which is in agreement with the findings of Smets, Warne, and Wouters (2014). The DSGE model performs better in the earlier period (1993–2002) than in the later period (2003–2011), confirming the literature that supports the use of DSGE forecasts during the Great Moderation period (1985–2007) (Del Negro & Schorftheide, 2013).

In summary, we provide evidence that structural (DSGE) models can deliver superior long-horizon forecasts of US and UK inflation.

7. Conclusion

The comprehensive evaluation of macroeconomic forecasting models that is reported in this paper contributes to both the academic literature and the practice of macroeconomic forecasting. By employing datasets for seven developed economies and considering four classes of multivariate forecasting models, we provide new empirical findings, extending and enhancing the evidence that is usually available for US data.

Our multicountry comparison provides a new dimension when comparing structural with reduced-form models in forecasting. The DSGE model specification that we consider (Smets & Wouters, 2007) provides accurate one- and two-year-ahead forecasts of inflation not only for the US, but also for the UK.

This evaluation was designed to look at forecasting horizons from nowcasting up to two years ahead. Our contribution is to consider a large set of model specifications over all of these horizons to enable us to provide evidence that the choice of the best forecasting model class clearly varies with the forecast horizon. We propose meta-analysis regressions in order to draw a small set of clear messages from 2976 relative accuracy comparisons.

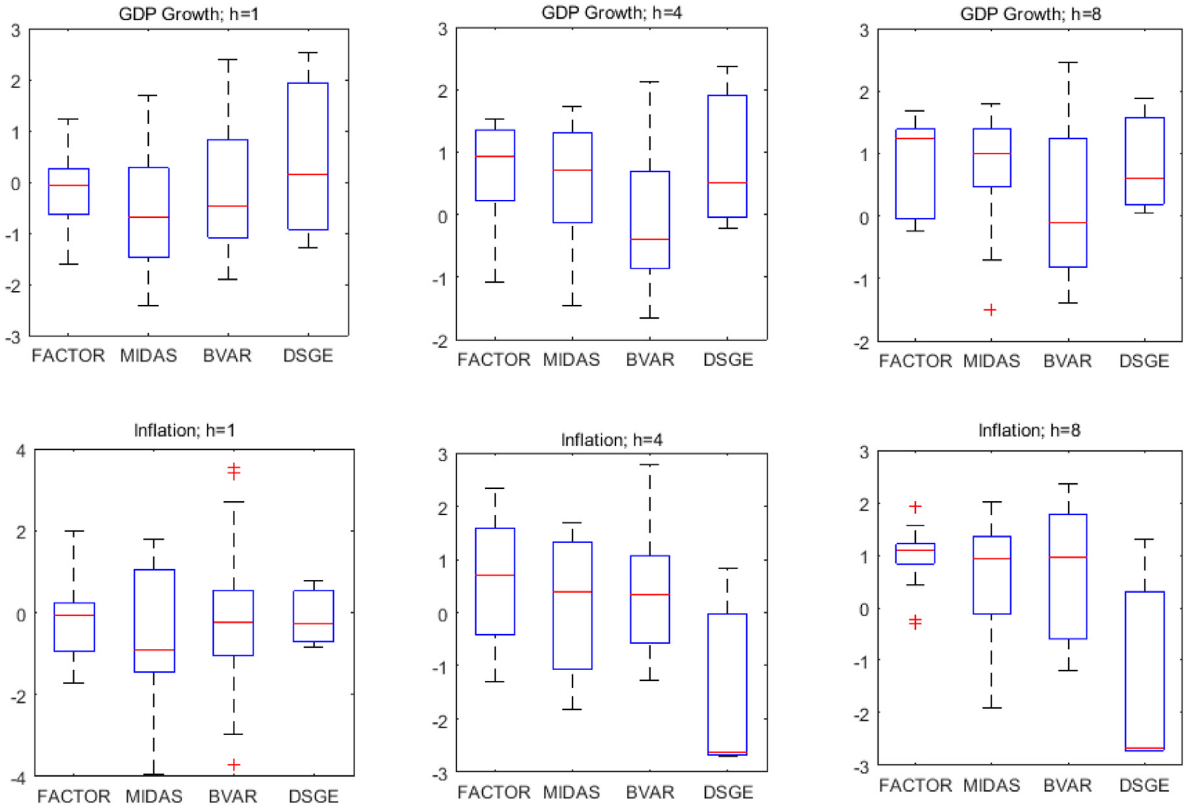

Fig. 3. Box plots of the equal accuracy logscore t-statistics with the AR model under the null and a forecasting model from the model class indicated under the alternative (aggregate over specifications and countries, full out-of-sample period) Note: See the notes to Fig. 1.

We extend previous results based only on Bayesian VARs (Koop, 2013) by showing that the use of a large set of predictors instead of a moderate set does not improve forecasts. Our contribution is to employ five different specifications from three model classes in order to investigate whether it is worthwhile to use large datasets instead of only 10–15 chosen predictors for both point and density forecasting, and we find that a medium dataset does indeed typically suffice. When dealing with large numbers of predictors (more than 50) for estimating a forecasting model over a short time period, we find that factor augmented distributed lag models and equal-weighted combinations of single-predictor mixed-data sampling regressions perform better than BVARs for predicting key macroeconomic variables when considering point and density forecasting.

Acknowledgments

All three authors acknowledge support for this work from the Economics and Social Research Council [ES/K010611/1]. We are grateful to our research assistant Katerina Petrova, who was also funded by the ESRC, for excellent assistance.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2019.02.007.

References

Aastveit, K. A., Foroni, C., & Ravazzolo, F. (2016). Density forecasts with MIDAS models. *Journal of Applied Econometrics*, *32*, 783–801.

Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should macroeconomic forecasters look at daily financial output and how? *Journal of Business & Economic Statistics*, *31*, 240–251.

Banbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. In *Handbook of economic forecasting, Vol. 2A* (pp. 195–237). Elsevier, chapter 4.

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, *25*(1), 71–92.

Carriero, A., Clark, T. E., & Marcellino, M. (2015). Bayesian VARs: Specifications choices and forecast accuracy. *Journal of Applied Econometrics*, *30*, 46–73.

Chauvet, M. (1998). An econometric characterization of business cycle dynamcis with factor structure and regime switches. *International Economic Review*, *39*, 969–996.

Chauvet, M., & Potter, S. (2013). Forecasting output. In *Handbook of economic forecasting, Vol. 2A* (pp. 141–194). Elsevier, chapter 3.

Christoffel, K., Coenen, G., & Warne, A. (2010). Forecasting with DSGE models. ECB Working Paper Series no. 1185.

Christoffel, K., Coenen, G., & Warne, A. (2010). Forecasting with DSGE models. ECB Working Paper Series no. 1185.

Clark, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, *29*, 327–341.

Clements, M. P., & Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. *Journal of Business & Economic Statistics*, *26*, 546–554.

Clements, M. P., & Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting*, *17*, 247–267.

D'Agostino, A., Gambetti, L., & Giannone, D. (2013). Macroeconomic forecasting and structural change. *Journal of Applied Econometrics*, *28*, 82–101.

Del Negro, M., & Schorftheide, F. (2011). Bayesian macroeconometrics. In J. Geweke, G. Koop, & H. van Dijk (Eds.), *The oxford handbook of bayesian econometrics* (pp. 293–389). Oxford University Press.

Del Negro, M., & Schorftheide, F. (2013). DSGE model-based forecasting. In *Handbook of economic forecasting, Volume 2A* (pp. 57–140). Elsevier, chapter 2.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 253–263, Reprinted in Mills, T. C. (ed.) (1999), *Economic forecasting*. The international library of critical writings in economics. Cheltenham: Edward Elgar.

Diebold, F. X., Schorftheide, F., & Shin, M. (2017). Real-time forecast evaluation of DSGE models with stochastic volatility. *Journal of Econometrics*, *201*, 322–332.

Edge, R. M., & Gurkaynak, R. S. (2011). How useful are estimated DSGE model forecasts. Federal Reserve Board, Finance and Economics Discussion Series 11.

Faust, J., & Wright, J. H. (2013). Forecasting inflation. In *Handbook of economic forecasting, Vol. 2A* (pp. 3–56). Elsevier, chapter 1.

Ferrara, L., Marcellino, M., & Mogliani, M. (2015). Macroeconomic forecasting during the Great Recession: the return of non-linearity? *International Journal of Forecating*, *31*, 664–679.

Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, *25*, 595–620.

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economic and Statistics*, *97*, 412–435.

Groen, J. J. J., & Kapetanios, G. (2013). Model selection criteria for factor-augmented regressions. *Oxford Bullettin of Economics and Statistics*, *75*, 37–63.

Herbst, E., & Schorftheide, F. (2012). Evaluating DSGE model forecasts of comovements. *Journal of Econometrics*, *171*, 152–166.

Koop, G. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, *28*, 177–203.

Koop, G., & Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, *177*, 185–198.

Kuzin, V., Marcellino, M., & Schumacher, C. (2013). Pooling versus model selection for nowcasting with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometrics*, *28*, 392–411.

Schorfheide, F., & Song, D. (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, *33*, 366–380.

Sims, C. (1993). A nine-variable probabilistic macroeconomic forecasting model. In *Business cycles, indicators and forecasting* (pp. 179–212). National Bureau of Economic Research.

Sims, C. (2000). Using a likelihood perspective to sharpen econometric discourse: three examples. *Journal of Econometrics*, *95*(2), 443–462.

Smets, F., Warne, A., & Wouters, R. (2014). Professional forecasters and real-time forecasting with a DSGE model. *International Journal of Forecasting*, *30*, 981–995.

Smets, F., & Wouters, R. (2007). Shocks and frictions in US business cycles.. *American Economic Review*, *97*, 586–606.

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

Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, *41*, 788–829.

Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39(Suppl.)*, 3–33.

Woulters, M. H. (2015). Evaluating point and density forecasts of DSGE models. *Journal of Applied Econometrics*, *30*, 74–96.