Aggregate Density Forecasting from Disaggregate Components Using Large VARs

Cobb, Marcus P A

February 2017

Aggregate Density Forecasting from Disaggregate Components Using Large VARs

Marcus P. A. Cobb*

February 2017

Abstract

When it comes to point forecasting there is a considerable amount of literature that deals with ways of using disaggregate information to improve aggregate accuracy. This includes examining whether producing aggregate forecasts as the sum of the component's forecasts is better than alternative direct methods. On the contrary, the scope for producing density forecasts based on disaggregate components remains relatively unexplored. This research extends the bottom-up approach to density forecasting by using the methodology of large Bayesian VARs to estimate the multivariate process and produce the aggregate forecasts. Different specifications including both fixed and time-varying parameter VARs and allowing for stochastic volatility are considered. The empirical application with GDP and CPI data for Germany, France and UK shows that VARs can produce well calibrated aggregate forecasts that are similar or more accurate than the aggregate univariate benchmarks.

Keywords: Density Forecasting; Bottom-up forecasting; Hierarchical forecasting; Bayesian VAR; Forecast calibration

JEL codes: C32, C53, E27, E37

*The author is grateful to Andrea Carriero and Marco Mariotti for their valuable comments and support. This research was produced while studying at the School of Economics and Finance, Queen Mary University of London and the author acknowledges and is grateful for their financial support.

Non-technical Summary

Assessing the state of the economy and providing an outlook for where it is heading involves interpreting large amounts of data in a way that is coherent. Macroeconomic aggregates are fundamental to this process. Due to the fact that they are built from disaggregate information there is an ongoing debate on whether and how to incorporate disaggregate information in order to improve aggregate forecasts. In some situations, however, the dynamics underlying an aggregate forecast are required for analysis. In these cases, to produce the forecast scenarios, practitioners usually rely on the bottom-up approach, that is building the aggregate forecast as the sum of its component's forecasts.

The bottom-up approach, in the context of point-forecasts, can present some challenges, but overall is reasonably straightforward. For density forecasts, on the other hand, it is not. This is troubling given that probability forecasting is being used increasingly in both finance and economics. Some efforts have been made to benefit from the disaggregate components in the process of forecasting the aggregate, but these have relied on methods that do not preserve the direct link between the aggregate and its components. As with the case of point-forecasts, however, in some situations a consistent underlying scenario for the aggregate forecast may be required.

In this paper we present a framework that extends the bottom-up approach to density forecasting with the objective of providing a consistent forecast scenario that is comparable to or better than those of direct methods. We do so by using the methodology of large Bayesian VARs to estimate the whole multivariate process and use the appropriate index weights to produce the aggregate forecast. We allow for both fixed and time-varying parameter VARs and for stochastic volatility. Our empirical application uses CPI and GDP data from France, Germany and the United Kingdom. We find that the multivariate methods are capable of producing bottom-up forecasts that are calibrated and perform equally or better than comparable aggregate methods.

1 Introduction

Assessing the state of the economy and providing an outlook for where it is heading involves interpreting large amounts of data in a way that is coherent. Macroeconomic aggregates are fundamental to this process given that they synthesise the information from countless indicators into relatively few figures. Due to the fact that they are built from disaggregate information there is an ongoing debate on whether and how to incorporate disaggregate information in order to improve aggregate forecasts.

For point forecasts there is sufficient evidence that supports the benefits in terms of aggregate accuracy of including disaggregate information in the forecasting process (Brüggemann and Lütkepohl, 2013). Some argue in favour of including disaggregate information in a model that forecasts the aggregate directly (Hendry and Hubrich, 2011). Over the years, however, a lot of attention has been given to whether forecasting the aggregate as the sum of its component's forecasts achieves better results than forecasting the aggregate directly. This may be due to the fact that this bottom-up approach provides a consistent underlying scenario for an aggregate forecast and is, therefore, favoured among institutions producing short-term forecasts (Esteves, 2013; Ravazzolo and Vahey, 2014). In terms of how it performs compared to other methods Lütkepohl (1987) show that it depends on the disaggregate processes and the aggregation matrix of the particular problem. The differing results from the many practical comparisons confirm that whether it is the best method is an empirical matter.[1]

The amount of research for point forecasts contrasts with that for density forecasting. It would seem that making use of disaggregate components has remained a relatively unexplored area. This is odd given that probability forecasting is being used increasingly in both finance and economics to assess the uncertainty surrounding forecasts (Mitchell and Hall, 2005).

Exceptions to this relative scarceness are Bache et al. (2010) and Ravazzolo and Vahey (2014). They use ensemble forecasting, a method adapted from the meteorology literature, where univariate autoregressive models are used for the components and aggregation weights are estimated so as to produce a well calibrated aggregate forecast. They work on the basis that the component models are almost surely misspecified but argue that an approximation to the aggregate can be found by using an appropriate mixture.

[1] Examples of these comparisons are Espasa et al. (2002), Benalal et al. (2004), Hubrich (2005) and Giannone et al. (2014) for inflation in the Euro area; Marcellino et al. (2003), Hahn and Skudelny (2008), Burriel (2012) and Esteves (2013) for European GDP growth; and Zellner and Tobias (2000), Perevalov and Maier (2010) and Drechsel and Scheufele (2013) for GDP growth in specific industrialized countries.

Their approach is partly motivated by the fact that practitioners commonly rely on univariate models to generate forecasts for components because of the difficulties involved in modelling their dependencies. This may be the case, but discarding the original weights means that, for the purpose of analysis, there is no way to link the aggregate forecast to the expected paths of the components. Espasa and Mayo-Burgos (2013) and Esteves (2013), among others, have raised their concerns regarding evaluating a disaggregate method solely based on aggregate accuracy and, in particular, argue that for the formulation of useful economic policies the dynamics of the underlying component's forecasts may be more important than the aggregate itself.

With those considerations in mind, it seems desirable to retain the original weights and a way of doing this is to model the whole multivariate process. Fortunately, in recent years Bayesian methods for dealing with large multivariate processes have been developed and have generated a lot of interest because of their good performance (Carriero et al., 2009; Banbura et al., 2010; Koop, 2013). In this paper we, therefore, use the methodology of large Bayesian VARs to extend the bottom-up approach to density forecasting with the objective of providing forecasts that are comparable to or better than those of direct methods. To do this, we implement different specifications that relax the constraints of the univariate framework. This includes considering both fixed and time-varying parameter VARs and allowing for stochastic volatility.

The rest of the paper is organized as follows. Section 2 presents the methodology. Section 3 presents an empirical implementation using GDP and CPI data for France, Germany and the United Kingdom. Section 4 summarizes the conclusions.

2 Disaggregate Forecasting Methodology

Over the last decade there has been a growing interest in Bayesian methods for policy analysis and forecasting. As pointed out in Carriero et al. (2015), much attention has concentrated on using Bayesian vector autoregressions (BVARs) with large datasets for point and density forecasting. The idea behind the BVAR is that prior information is imposed on the VAR coefficients to avoid overparametrization.

For practitioners that are affected by the limited feasible size of traditional VARs, such an alternative is probably specially welcome. In spite of the remarkable increase in computational power, however, some approaches remain technically and computationally demanding. This could be a stumbling block for their adoption in contexts where the production of forecasts is subject to very tight time constraints, but fortunately alternatives that avoid the more intensive simulation are available.

The implementation suggested by Banbura et al. (2010) has received considerable attention since it was first presented. They suggest a relatively simple way of using Bayesian shrinkage to overcome the dimensionality problem in traditional VARs. In their empirical application, they find that their BVARs perform at least as well as the popular factor methods. They do, however, only contemplate using constant coefficients and homoskedastic errors. Koop and Korobilis (2013) take it a step further and develop a methodology that also allows implementing time-varying parameters and stochastic volatility without increasing computational demands. Because of this extra flexibility and other convenient features of their implementation, we use their model in our framework to produce bottom-up density forecasts.

2.1 Large Time-varying parameters VARs

Koop and Korobilis (2013) formulate the problem in state-space form:

$$\begin{aligned} y_t &= X_t\beta_t + \varepsilon_t \qquad \varepsilon_t \sim \text{i.i.d.} N(0, \Sigma_t) \\ \beta_{t+1} &= \beta_t + u_t \qquad u_t \sim \text{i.i.d.} N(0, Q_t) \end{aligned} \tag{1}$$

where ε_t and u_s are independent from one another for all s and t. y_t for $t = 1, \ldots, T$ is an $M \times 1$ vector containing observations on M time series and X_t is an $M \times k$ matrix defined so that each TVP-VAR equation contains an intercept and p lags of each of the M variables.

They argue that even for relatively small problems the computational burden could be quite significant. Therefore, instead of proceeding in a standard Bayesian way by using MCMC methods they suggest replacing Q_t and Σ_t with estimates. To achieve this, while still retaining time-varying parameters and stochastic volatility, they use forgetting factors to produce their approximation at each point in time. This means estimating empirically the desired parameters, but in a way that downplays to a chosen degree the contribution of less recent data.

In regards to the time-varying parameters, they start by noting that Q_t only appears in one place in the Kalman filtering process, particularly in the prediction step. Then, following the forgetting factors approach, they replace Q_t for $(\lambda^{-1} - 1)V_{t-1|t-1}$, where $V_{t-1|t-1}$ is the variance of $\beta_{t-1}|y_{t-1}$, resulting in $V_{t|t-1} = \frac{1}{\lambda}V_{t-1|t-1}$. The forgetting factor λ is restricted to be strictly positive and less than one being the constant coefficient specification achievable by setting $\lambda = 1$.

Similarly, to avoid using a posterior simulation algorithm to model volatility, they use an Exponentially Weighted Moving Average estimator for the measurement error covari-

ance matrix. This is done by making $\hat{\Sigma}_t = \kappa\hat{\Sigma}_{t-1} + (1-\kappa)\hat{\varepsilon}_t\hat{\varepsilon}_t'$ with $\hat{\varepsilon}_t = yt - \beta_{t|t}X_t$. Here the forgetting factor κ is also restricted to be between zero and one.

In regards to the estimation of the coefficients of the BVAR, that is the β's, they use a Normal prior. Given their choice of variable transformation, for β_0 they set the prior mean to zero and the covariance matrix to be diagonal. Specifically, for var$(\beta_0) = \underline{V}$, with $\underline{V}_i$ being its diagonal elements, they define $\underline{V}_i = \gamma/r^2$ for coefficients on the r-th lag and for the intercepts use a noniformative prior. This results in having a single hyperparameter γ control the shrinkage of the coefficients. In this case $0 \leq \gamma < \infty$.

2.2 Empirical parameter selection

The model proposed by Koop and Korobilis (2013) is relatively simple and capable of incorporating many features, despite being governed by three parameters. These parameters, however, have to be provided by the researcher.

In regards to the values governing the time-varying parameters and stochastic volatility, Koop and Korobilis provide values that would be consistent with previous literature in those areas. They do acknowledge, however, that a method that determines them from the data would be very appealing and, therefore, go on to develop one based on dynamic model selection methods (DMS).

They set the problem up as one of selecting one model definition from a set of models that are the same in terms of explanatory variables, but differ in terms of parameter values.[2] Their criterion is to choose the specification with the highest probability of being the appropriate one for forecasting at any given time. They estimate this probability by implementing a recursive algorithm developed by Raftery et al. (2010) that, conveniently, can be run within the normal Kalman filtering process used to produce the forecasts.[3]

In this context, the prediction step is extended slightly with the additional equation:

$$\pi_{t|t-1,j} = \frac{\pi_{t-1|t-1,j}^{\alpha}}{\sum_{l=1}^{J} \pi_{t-1|t-1,l}^{\alpha}} \tag{2}$$

were $\pi_{t|t-1,j}$ is the probability that model j should be used to forecast at time t given the information up to $t-1$, α is a forgetting factor and J is the number of specifications

[2] Koop and Korobilis go on to extend the approach to also allow for differing explanatory variables.

[3] The algorithm by Raftery et al. (2010) is explained in detail in Section 2.3 of Koop and Korobilis (2013).

being considered, and the updating step by:

$$\pi_{t|t,j} = \frac{\pi_{t|t-1,j} p_j \left(y_t \mid y^{t-1}\right)}{\sum_{l=1}^{J} \pi_{t-1|t-1,l} p_l \left(y_t \mid y^{t-1}\right)} \tag{3}$$

were $p_j \left(y_t \mid y^{t-1}\right)$ is the predictive likelihood.

The idea behind the algorithm is that good performance in the recent past increases the probability of the model being the appropriate one to forecast for the following period. The predictive likelihood serves as the measure of forecast performance and the forgetting factor α to define what is understood as "recent past". In this case, an α close to zero leads approximately to the equal weighting for all time periods while setting α = 1 corresponds to using the marginal likelihood.

The method is sufficiently general so that Koop and Korobilis also use it to estimate the prior hyperparameter which controls shrinkage in large Bayesian VARs, not only the forgetting factors for the time-varying parameters and stochastic volatility.

2.3 Aggregate Density Forecasts from Component Forecasts

As pointed out by Ravazzolo and Vahey (2014) practitioners often rely on univariate models because of the difficulties involved in modelling the dependencies among components. Ignoring these dependencies however means that using a traditional bottom-up approach could yield poor aggregate density forecasts. Ravazzolo and Vahey (2014) acknowledge that by assuming that the disaggregate forecasting equations are misspecified and propose approximating the unknown true specification by estimating appropriate aggregation weights.

On the other hand, if the multivariate process is modelled well, using the index weights would be appropriate and should produce well calibrated aggregate forecasts. Determining the distribution of a sum of random variables, however, is generally quite complicated, but in this case, the task is simplified greatly by the fact that the densities for the components are produced using a sampling algorithm. As any given draw describes the whole multivariate process, the aggregate forecast for that draw, can be produced simply by summing the component forecasts using the appropriate index weights. Doing this for all draws provides the aggregate bottom-up density forecast.

3 Empirical Application

The success of the proposed method depends on two factors. The first is whether it performs well in circumstances where the univariate bottom-up approach fails to produce a well calibrated aggregate forecast. The second, and maybe more relevant in practical settings, is how it performs relative to other methods that do produce well calibrated aggregate forecasts. The extent to which this can be measured depends fundamentally on the properties of the data that is used. For this reason, we consider using more than one dataset to have a broader assessment. In particular, we perform a out-of-sample forecasting exercise using GDP and CPI data from Germany, France and United Kingdom. We use different specifications for the BVARs and evaluate the calibration of the aggregate forecast densities using a series of tests and their relative performance using log predictive density scores.

Regarding the forecast horizon, we restrict the scope of this exercise to the one-step-ahead. The reason being that in the context of this exercise, as the series considered are produced using either a fixed-base or annual overlap chain-linking method, the definitive weights for the one-step-ahead forecast are always available at the time of forecasting. For longer horizons, however, they are not. This means that for longer horizons the weights would also need to be forecasted. One option would be to use the previous period's weights as practitioners often do (Ravazzolo and Vahey, 2014), but Lütkepohl (2011) and Hendry and Hubrich (2011), among others, discuss the problems that arise from imposing weights to be unchanging and emphasise that, if the actual weights change through time, forecast performance can deteriorate quickly with longer horizons being affected the most.

3.1 Data

For the exercise we use GDP from the production approach and CPI for France, Germany and the United Kingdom. The data is quarterly and seasonally adjusted, spanning from 1991 to 2015 and available from the OECD statistics database.[4]

The breakdown of the aggregates is the following:

[4] For the United Kingdom the production data on the OECD database starts in 1995. The first four years of the sample are obtained by splicing backwards the historical reference tables available from the Office for National Statistics. No inconsistencies arise from the seasonal adjustment given that the aggregates are adjusted indirectly, that is as the sum of the seasonally adjusted components.

Table 1: Components Breakdown

GDP		
	1. Agriculture, forestry and fishing	7. Financial and insurance activities
	2. Manufacturing	8. Real estate activities
	3. Industry and energy, excluding manufacturing	9. Professional, administrative and support service activities
	4. Construction	10. Public adm., defence, social security, education and health
	5. Trade, transport, accommodation and food services	11. Other service activities
	6. Information and communication	12. Taxes less subsidies
CPI		
	1. Food and non-Alcoholic beverages	7. Transport
	2. Alcoholic beverages, tobacco and narcotics	8. Communication
	3. Clothing and footwear	9. Recreation and culture
	4. Housing, water, electricity, gas and other fuels	10. Education
	5. Furnishings, household equipment and maintenance	11. Restaurants and hotels
	6. Health	12. Miscellaneous goods and services

3.2 BVAR specifications

The evaluation exercise is performed over the 2001-2015 period leaving the first years of data to estimate the models. It is set up in a quarterly rolling scheme using a ten year window where in each period the models are re-estimated and a density forecast is generated. For this we use different specifications for the BVARs.[5] Firstly we use a homoskedastic VAR that is obtained by setting both λ and κ equal to one and in which case $\hat{\Sigma}_t$ is estimated by $\frac{1}{1-t}\sum_{i=1}^{t-1}\hat{\varepsilon}_i\hat{\varepsilon}_i'$. Secondly a homoskedastic TVP-VAR with $\lambda = 0.99$ that is a value that Koop and Korobilis (2013) argue is equivalent to what has previously been used in the relevant literature and for which, for quarterly data, observations five years back receive approximately 80% as much weight as last period's observation.[6] They argue that such a value leads to a gradual change in coefficients and stable models. Based on this, the third model is a heteroskedasic VAR with $\kappa = 0.99$. Finally, to allow for both features, the fourth model is a heteroskedasic TVP-VAR with both λ and κ equal to 0.99.[7] In regards to setting the value for the overall shrinkage of the coefficients we use the parameter selection algorithm described in section 2.2 over a wide grid for all specifications.[8]

Koop and Korobilis argue that the TVP-VARs are well-suited for modelling gradual evolution of coefficients. To accommodate more sudden changes they advocate using dynamic model selection over a whole array of model specifications. Given that the sample includes the years of the financial crisis, allowing for abrupt changes in parameters could

[5] For all we use four lags.

[6] Setting $\lambda = 1$ is equivalent to using the marginal likelihood. The closer to zero the less consideration is given to older information.

[7] Koop and Korobilis also choose λ and κ empirically over a grid. We follow their implementation but find the results are not significantly different from those obtained from setting both parameters to 0.99.

[8] Specifically, we set $\gamma = e^i$ and select i from {-7, -6, ... , -1}.

be particularly relevant. Therefore, as a final approach, we produce a series of forecasts using, at each point in time, the model out of the previous four with the highest probability of being appropriate, $\pi_{t|t-1,j}$, according to the aforementioned algorithm.

As benchmarks for the forecasting exercise we use aggregate univariate AR models and a bottom-up forecast using univariate AR models for the components. For these we contemplate from one to four lags.

3.3 Forecast Evaluation

A popular way of assessing the calibration of the density forecasts is testing the sequence of probability integral transform (PIT) values. These are defined as $p_t = F_t(x_t)$, where F_t is the predictive cumulative distribution functions and x_t the observed realization. If F_t coincides with the true data generating process, the PITs are uniform $U(0, 1)$ for any forecast horizon and i.i.d. for one-step-ahead forecasts (Diebold et al., 1998). Geweke and Amisano (2010) describe this approach as comparing the distribution of the observed data with the distribution that would have resulted if the model under consideration had being used to generate the data.

Mitchell and Hall (2005) point out that testing in this context is not straightforward given that the impact of dependence on uniformity tests and vice versa is unknown. The empirical literature has relied therefore on using a number of tests simultaneously. Following Mitchell and Wallis (2011) and Ravazzolo and Vahey (2014) we use Pearson's chi-squared test to assess the goodness-of-fit of the PIT histogram to a univariate distribution and the Anderson-Darling test to evaluate the uniformity of the empirical cumulative distribution function of the PITs. We directly test their independence using a Ljung-Box test using autocorrelation of up to four lags. Finally we use the test proposed by Berkowitz (2001) that tests for goodness-of-fit and independence.[9]

A problem with only testing the calibration is that it is quite possible that two or more forecasts can be found to be equally well-calibrated (Gneiting et al., 2007). This is a drawback, specially for practitioners that are looking to choose a single model. An alternative approach is to use scoring rules. These assign a numerical score based on the predictive likelihood and the realization of the variable. Based on the difference in their scores, models can effectively be compared. Following Carriero et al. (2015), we use the log predictive density scores to assess overall calibration of each forecast. In particular, we use the average log score over the sample where the log score for the density forecast f_{it}, is defined as $\log f_{it}(x_t)$.

[9]The test is in fact applied on the inverse normal transform of the PITs to test for normality. We use the three degrees of freedom version that tests against a first-order autoregressive alternative and wrong mean and variance.

Table 2: Tests on PITs for one-step-ahead GDP forecasts

	Germany				France				United Kingdom			
Model	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB
Bottom-Up AR	0.02	0.00	**0.21**	0.01	0.00	**0.28**	**0.37**	**0.16**	0.00	0.00	**0.14**	**0.16**
Direct AR	**0.41**	0.00	0.01	**0.14**	**0.06**	**0.05**	**0.40**	**0.08**	**0.06**	**0.15**	**0.26**	0.02
Homsk. VAR	**0.46**	0.02	**0.13**	**0.91**	**0.12**	**0.57**	**0.72**	**0.59**	0.03	**0.11**	**0.33**	**0.35**
Homsk. TVP	**0.51**	0.01	**0.14**	**0.89**	**0.12**	**0.38**	**0.67**	**0.59**	0.03	**0.06**	**0.92**	**0.45**
Hetsk. VAR	0.00	0.00	0.00	**0.77**	**0.40**	0.02	**0.31**	**0.86**	0.00	0.00	**0.34**	**0.21**
Hetsk. TVP	0.00	0.00	0.00	**0.52**	**0.30**	0.03	0.03	**0.68**	0.00	0.00	**0.33**	**0.21**
DMS	**0.72**	**0.08**	**0.30**	**0.72**	**0.17**	**0.54**	**0.89**	**0.70**	0.01	0.02	**0.11**	**0.29**

Note: P-values for the calibration tests on the probability integral transform (PIT) of the one-step-ahead forecasts for each model for the three countries. The tests are the LR test proposed by Berkowitz (2001) (Bkw.LR), the uniformity tests by Anderson-Darling (AD) and a Pearson's chi-squared (χ^2), the Ljung-Box test (LB) for independence. The models are the bottom-up univariate model (Bottom-up AR), the direct univariate AR model (Direct AR), the homskedastic VAR, the homskedastic TVP-VAR, the heterokedastic VAR, the heterokedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). P-values in bold signify that the null of the respective test are not rejected at 5%. Calculated over the 2001-2015 period.

3.4 Results

3.4.1 GDP forecasts

Table 2 presents the tests on the PITs for the one-step-ahead forecasting exercise for GDP for all three countries. As Ravazzolo and Vahey (2014) put it, well-calibrated forecasts should give high probability values for all four tests. The overall impression from the results, however, is that few specifications pass all four diagnostic tests.[10] For Germany, for example, only the DMS model does so while none does for the United Kingdom. This is not surprising, however, given that the evaluation sample includes the last global financial crisis. The performance of the Direct AR suggests that there is more to it than a generalized shortcoming in the bottom-up approach.

As Mitchell and Hall (2005) point out, doing comparisons based on the tests is not straightforward. The outcome of the PITs tests is binary. Either the forecasts are well-calibrated according to the set of tests, that is, it is not rejected that the PITs can come from a uniform distribution, or they are not. In a practical situation like this, such a judgement seems insufficient. Some assessment on how badly or well-calibrated the forecasts are would probably prove to be useful. With this objective in mind, given that the series of tests evaluate different aspects of the PITs distribution, one might expect that forecasts that fail one test marginally are probably closer to being well-calibrated than those that fail all of them by a mile.[11] Under this premise, the overall reading of the results is that, unsurprisingly, the univariate bottom-up approach would seem to

[10] That is that the null hypothesis of no calibration failure cannot be rejected at the 5% significance level. The tests are conducted on an individual basis which imply a Bonferroni-corrected (joint) p-value of 1.25%.

[11] This idea is related to visually inspecting the histograms and assessing how close they are to a uniform distribution. The relevant literature, in fact, also suggests checking the PITs histograms visually. These are all presented in the Appendix.

Table 3: Log scores for one-step-ahead GDP forecasts

Model	Germany	France	United Kingdom
Direct AR	7.4	22.9	7.0
Homsk. VAR	**12.0**	**26.1**	**10.1**
Homsk. TVP	**15.7**	**25.8**	**10.1**
Hetsk. VAR	2.4	**23.0**	5.2
Hetsk. TVP	4.3	**24.1**	5.3
DMS	**19.3**	**25.9**	**8.3**

Note: Log predictive density scores of the one-step-ahead forecasts for each model for the three countries expressed in terms of the percentage improvement over the bottom-up univariate model (Bottom-up AR). The models are the direct univariate AR model (Direct AR), the homskedastic VAR, the homskedastic TVP-VAR, the heterokedastic VAR, the heterokedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). Log scores in bold denote improvement over the direct univariate model. Calculated over the 2001-2015 period.

provide density forecasts that are less well calibrated than the direct approach. The multivariate bottom-up approaches, however, improve on both the univariate variant and the Direct AR in some cases. Overall the homoskedastic VARs are at least as good as the direct approach for all countries, while the DMS shows improvements only for Germany and France.

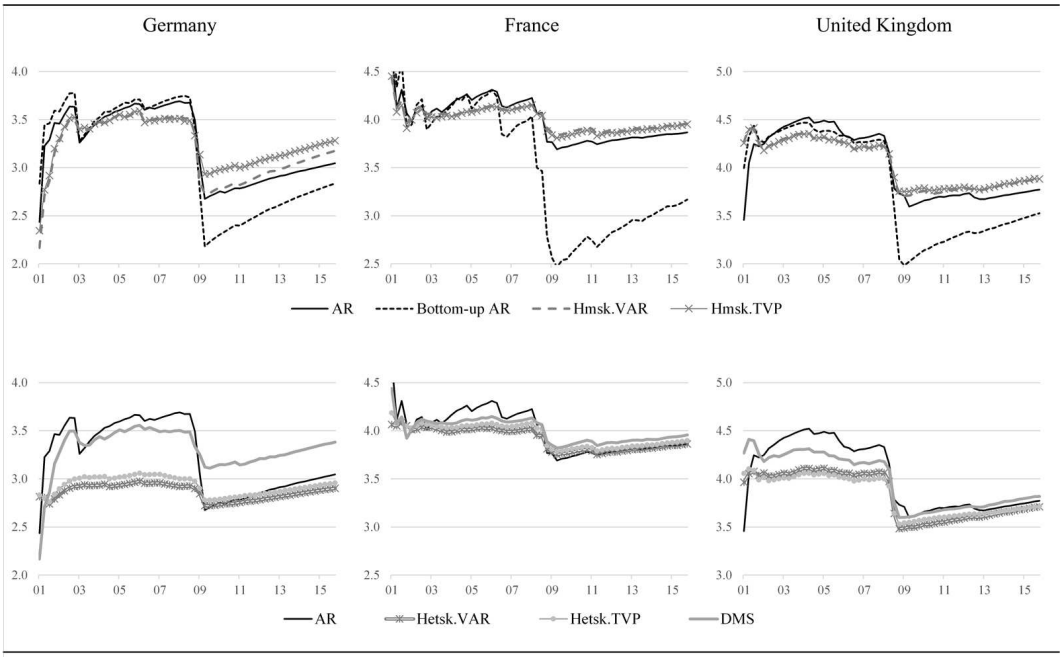
Even if trying to differentiate between models based on the PITs tests were possible, in a context where a practitioner is after the best available model, the results form the PITs tests in this case are of limited value. For example, for France there are four approaches that qualify as well-calibrated where none of the models achieves the highest value in all tests. For United Kingdom, on the other hand, none is well-calibrated and then there are three models that fail only one test. Therefore, to rank the different models, we turn to look at the logarithmic predictive density scores. Table 3 presents the log scores expressed in terms of the percentage improvement over the bottom-up univariate model.

The multivariate models perform better than the univariate bottom-up approach, but the improvements are heterogeneous. Overall, it is the homoskedastic models that show the best performance improving over the aggregate univariate model by as much as eight percentage points.

From the performance of the four different BVARs, it would seem that most of the gains come from allowing the process to be modelled using a multivariate model and that further gains can be obtained by allowing for the coefficients to vary over time. In contrast, incorporating stochastic volatility has a negative effect.

This is consistent with the results from the dynamic model selection (DMS). For Germany it performs very well. It shows the highest accuracy with an improvement of nearly 12% over the aggregate AR. It is also the only model for which uniformity and independence are not rejected by any of the tests. For France it performs virtually the same as the homoskedastic multivariate models both in terms of calibration according

Figure 1: GDP Recursive Log Scores



Note: Recursive log scores calculated over the 2001-2015 period. The models are the aggregate univariate model (Direct AR), the bottom-up forecast using univariate AR models for the components (Bottom-Up AR), the homskedastic VAR, the homskedastic TVP-VAR, the heterokedastic VAR, the heterokedastic TVP-VAR , the heterokedastic TVP-VAR with recursively estimated decay and forgetting factors and the result of dynamic model selection over the five VARs (DMS).

to the PITs tests and log score. For the United Kingdom it performs better than the direct AR but worse than the homoskedastic multivariate models.

The improvements over the bottom-up univariate model seem quite substancial, up to 26% in the case of France, so an obvious question to ask is whether the differences in predictive accuracy are significant or not. To assess whether they are, we consider a Kullback-Leibler information criterion equal predictive performance test (KLIC) as presented in Mitchell and Hall (2005). The test compares two loss differential series in a way that is analogous to the point forecast accuracy test popularized by Diebold and Mariano (1995).

We find that although the improvements seem quite large in magnitude, the differences are not significant according to the test. This could seem odd at first, but the recursive log scores that are presented in Figure 1 provide an answer to why this is the case.[12]

It is immediately obvious that the crisis produces a sharp decline in the scores. Common to all the three countries is that the bottom-up univariate model is significantly and by far the most affected out of all models. The second most affected model, however, is the aggregate AR. Although up until the crisis the univariate models are among the

[12]The homoskedastic models are presented in the top panel and the heteroskedastic models and DMS in the bottom. The aggregate AR is included in both to serve as a point of reference.

Table 4: Tests on PITs for one-step-ahead CPI forecasts

	Germany				France				United Kingdom			
Model	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB
Bottom-Up AR	0.00	**0.08**	0.04	**0.10**	0.00	**0.06**	**0.07**	**0.52**	0.00	**0.93**	**0.79**	0.03
Direct AR	**0.22**	**0.41**	**0.50**	**0.71**	**0.09**	**0.39**	**0.59**	**0.23**	**0.85**	**0.66**	**0.12**	**0.35**
Homsk. VAR	**0.81**	0.01	**0.07**	**0.62**	**0.44**	**0.10**	**0.09**	**0.69**	**0.68**	**0.66**	**0.40**	**0.69**
Homsk. TVP	**0.91**	**0.36**	**0.69**	**0.55**	**0.68**	**0.06**	0.02	**0.84**	**0.49**	**0.66**	**0.55**	**0.42**
Hetsk. VAR	**0.42**	**0.11**	0.04	**0.47**	0.00	**0.18**	**0.41**	**0.08**	**0.09**	**0.71**	**0.43**	**0.12**
Hetsk. TVP	**0.84**	**0.06**	**0.26**	**0.56**	0.00	**0.16**	**0.44**	**0.58**	0.03	**0.85**	**0.07**	**0.15**
DMS	**0.67**	**0.59**	**0.86**	**0.65**	**0.66**	**0.05**	0.04	**0.64**	**0.43**	**0.93**	**0.98**	**0.13**

Note: P-values for the calibration tests on the probability integral transform (PIT) of the one-step-ahead forecasts for each model for the three countries. The tests are the LR test proposed by Berkowitz (2001) (Bkw.LR), the uniformity tests by Anderson-Darling (AD) and a Pearson's chi-squared (χ^2), the Ljung-Box test (LB) for independence. The models are the bottom-up univariate model (Bottom-up AR), the direct univariate AR model (Direct AR), the homskedastic VAR, the homskedastic TVP-VAR, the heterokedastic VAR, the heterokedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). P-values in bold signify that the null of the respective test are not rejected at 5%. Calculated over the 2001-2015 period.

best performers, the multivariate models show falls that are proportionally smaller and therefore, at least in some cases, end up being better over the whole sample.[13]

The performance of both homoskedastic VARs is slightly worse than that of the univariate methods up until the crisis, but the comparatively better reaction to the crisis suggests that the increased uncertainty due to the estimation of additional parameters could be worth while. The opposite seems to be the case with the methods that incorporate stochastic volatility.

3.4.2 CPI forecasts

The results for CPI are less pronounced but have some things in common with those of GDP. Table 4 presents the tests on the PITs for the one-step-ahead forecasts for all three countries. Overall the forecasts from most models are well-calibrated according to the tests. The univariate bottom-up model, however, fails at least one test in each case. In regards to the multivariate models, in this case the models that include stochastic volatility are similarly well-calibrated to those that do not.

In terms of ranking the models by accuracy, as it can be seen from Table 5, the improvements of the multivariate models are smaller than in the case of GDP and heterogeneous between countries. For example, for Germany, the methods improve over the univariate bottom-up approach but are only marginally better than the direct AR if stochastic volatility is included. For the other two countries, there is little difference in accuracy between the univariate methods, but in the case of France the multivariate methods improve by as much as 5% while for the United Kingdom these are below 2%.

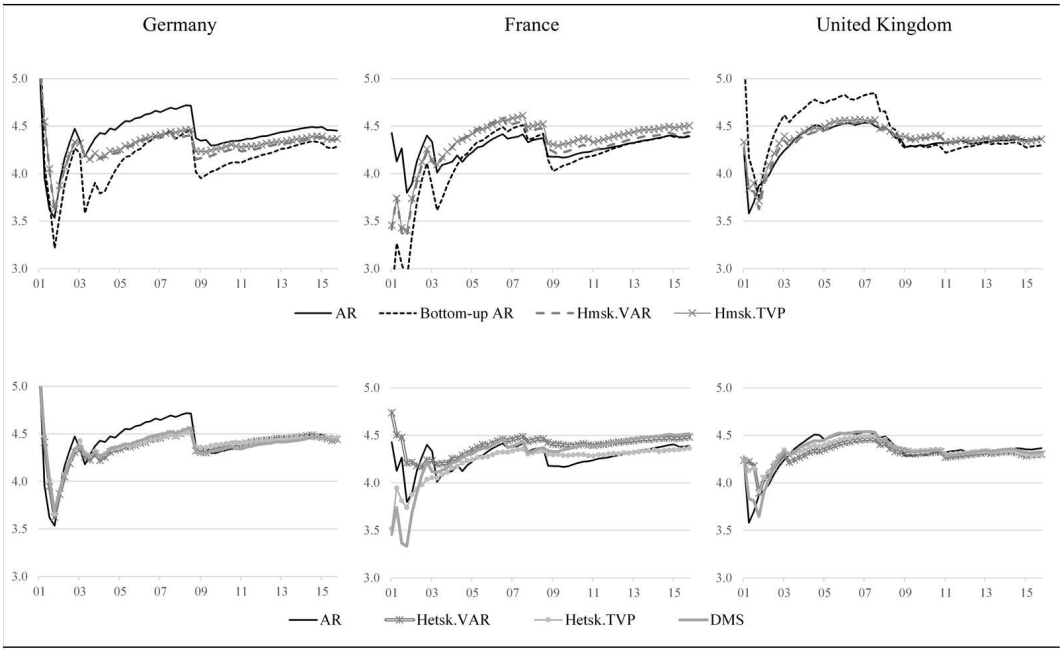
[13]The recursive log scores calculated excluding the crisis years, not reported, show that the univariate models perform very well over the restricted sample.

Table 5: Log scores for one-step-ahead CPI forecasts

Model	Germany	France	United Kingdom
Direct AR	4.7	0.2	-0.4
Homsk. VAR	2.8	**2.9**	**1.8**
Homsk. TVP	3.0	**4.5**	**1.5**
Hetsk. VAR	**4.8**	**4.0**	0.0
Hetsk. TVP	**5.2**	**1.3**	-0.1
DMS	4.6	**4.8**	**0.6**

Note: Log predictive density scores of the one-step-ahead forecasts for each model for the three countries expressed in terms of the percentage improvement over the bottom-up univariate model (Bottom-up AR). The models are the direct univariate AR model (Direct AR), the homskedastic VAR, the homskedastic TVP-VAR, the heterokedastic VAR, the heterokedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). Log scores in bold denote improvement over the direct univariate model. Calculated over the 2001-2015 period.

Figure 2: CPI Recursive Log Scores



Note: Recursive log scores calculated over the 2001-2015 period. The models are the aggregate univariate model (Direct AR), the bottom-up forecast using univariate AR models for the components (Bottom-Up AR), the homskedastic VAR, the homskedastic TVP-VAR, the heterokedastic VAR, the heterokedastic TVP-VAR , the heterokedastic TVP-VAR with recursively estimated decay and forgetting factors and the result of dynamic model selection over the five VARs (DMS).

In regards to how the models are affected by the crisis, Figure 2 presents the recursive log scores for CPI. It is still the case that the multivariate models are proportionally less affected than the univariate models, but the overall impact of the crisis is smaller. With this, performance over the whole sample is not that different between models. This is confirmed by the KLIC test. As opposed to the case of GDP, in this case the added complexities do not seem to pay off.

3.4.3 Overall assessment

Unsurprisingly, the performance of the different methods varies quite significantly depending on the dataset. However, there are a number of things that can be learned from the overall performance. The first thing is that, in line with the statements of Ravazzolo and Vahey (2014), the univariate bottom-up approach produced forecasts that were not well-calibrated in terms of the PITs tests and inferior to those produced using the direct approach in terms of relative performance. Some of the multivariate bottom-up methods, on the other hand, performed similarly or better. These results suggest that multivariate methods can overcome the problems in calibration that result form using univariate models in this context. The varying degrees of success of the different specifications, however, also suggest that the added complexities may not always be justified.

Overall, the homoskedastic fixed-parameter VAR is probably the best performer due to its consistency. Although, in some cases, gains were achieved by allowing time-varying parameters, most of the improvements were attainable in the simpler multivariate setting. This comes as good news for practitioners, as it suggests that the more extended implementation by Banbura et al. (2010) would probably also work well in the same setting.

The differences between the results for GDP and CPI suggest that the strengths of the multivariate methods only emerge if the interactions among variables are prominent enough. The more significant effects of the financial crisis on GDP, both in magnitude and persistence, result in the multivariate methods beating the univariate counterpart. The rest of the time, they were no different. The KLIC tests and the evaluation excluding the crisis years support this view.

4 Conclusions

In this paper we use the information at a component level to produce consistent aggregate and disaggregate density forecasts. To do this we use the methodology of large

Bayesian VARs to extend to a probabilistic setting the bottom-up approach used commonly for point-forecasts. We implement a relatively simple, but flexible, and computationally cheap method to consider both fixed and time-varying parameter VARs and stochastic volatility.

Our motivation follows that of Espasa and Mayo-Burgos (2013) in that, for the purpose of economic analysis, we consider our method to be successful if it produces forecasts that are at least as good as those of a direct method. In regards to this, the empirical application shows that, although the results vary to some extent between countries and series, overall, the multivariate methods are capable of producing bottom-up forecasts that are calibrated and perform equally or better than the aggregate benchmark. The results also suggest that there are additional gains from allowing for time-varying parameters.

In terms of future research, there are many possibilities. One is to produce the estimates for the time-varying and stochastic volatility parameters using alternative methods which includes using a full Bayesian approach and compare the results with those of the approximations. A natural extension would be to couple the method with one to forecast the aggregation weights and use the augmented framework to forecast at longer horizons. A third direction for research could be to incorporate useful economic indicators and other relevant variables into the forecasting process in a way that is similar to Banbura et al. (2010).

Appendix

A Figures and Tables

A.1 PITs Histograms for GDP

Histograms for the probability integral transform (PIT) calculated over the 2001-2015 period.

Figure 3: Germany

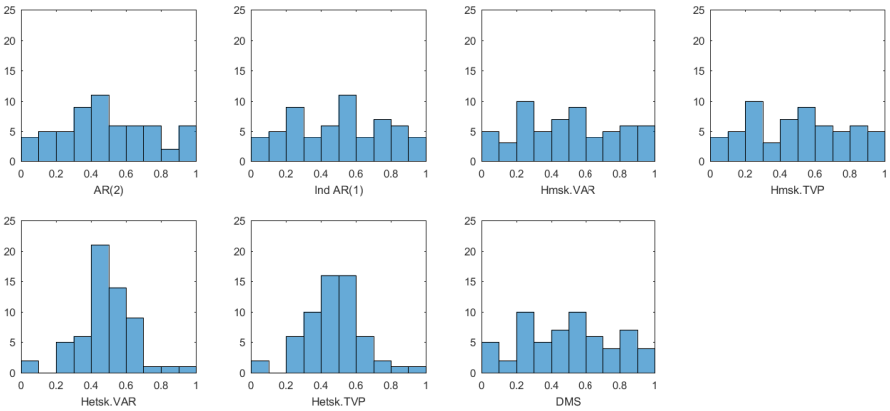


Figure 4: France

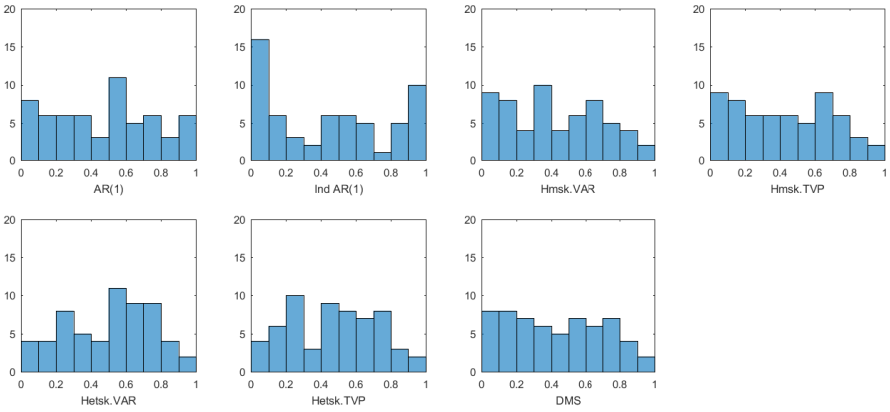
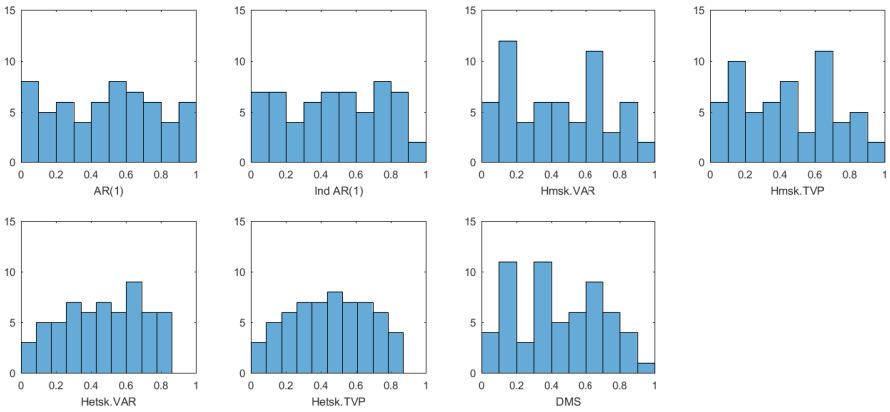


Figure 5: United Kingdom



A.2 PITs Histograms for CPI

Histograms for the probability integral transform (PIT) calculated over the 2001-2015 period.

Figure 6: Germany

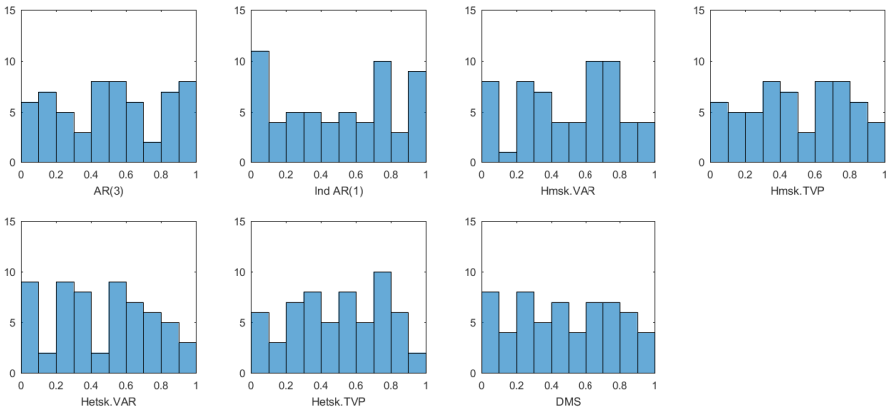


Figure 7: France

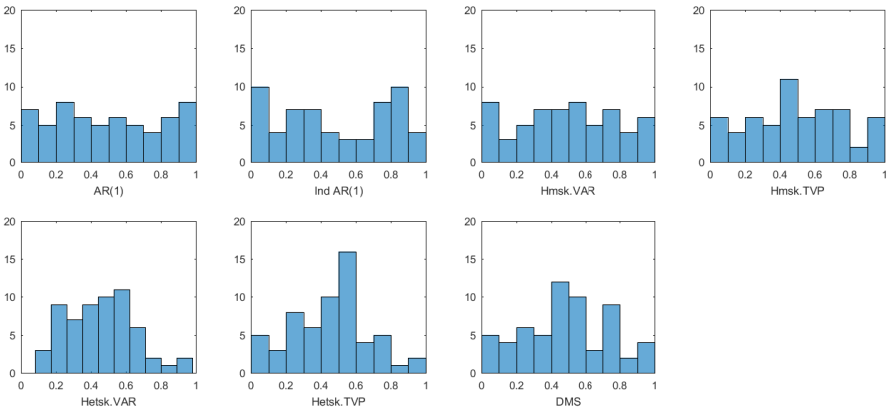
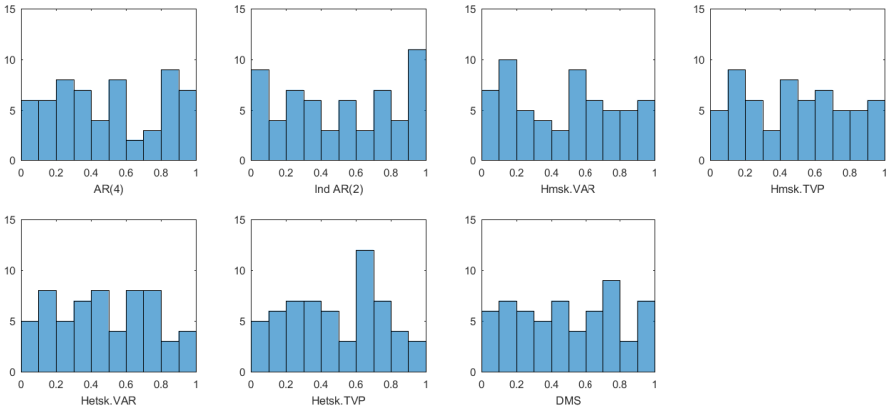


Figure 8: United Kingdom



References

Bache, I. W., J. Mitchell, F. Ravazzolo, and S. P. Vahey (2010). Macro-modelling with many models. *Twenty Years of Inflation Targeting: Lessons Learned and Future Prospects*. Chapter 16.

Banbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics 25*(1), 71–92.

Benalal, N., J. L. Diaz del Hoyo, B. Landau, M. Roma, and F. Skudelny (2004). To aggregate or not to aggregate? euro area inflation forecasting. Working Paper 374, European Central Bank.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics 19*(4), 465–474.

Brüggemann, R. and H. Lütkepohl (2013). Forecasting contemporaneous aggregates with stochastic aggregation weights. *International Journal of Forecasting 29*(1), 60–68.

Burriel, P. (2012). A real-time disaggregated forecasting model for the euro area gdp. *Economic Bulletin*, 93–103.

Carriero, A., T. E. Clark, and M. Marcellino (2015). Bayesian vars: specification choices and forecast accuracy. *Journal of Applied Econometrics 30*(1), 46–73.

Carriero, A., G. Kapetanios, and M. Marcellino (2009). Forecasting exchange rates with a large bayesian var. *International Journal of Forecasting 25*(2), 400–417.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*, 863–883.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & economic statistics* (13).

Drechsel, K. and R. Scheufele (2013). Bottom-up or direct? forecasting german gdp in a data-rich environment. IWH Discussion Papers 7, Halle Institute for Economic Research.

Espasa, A. and I. Mayo-Burgos (2013). Forecasting aggregates and disaggregates with common features. *International Journal of Forecasting 29*(4), 718–732.

Espasa, A., E. Senra, and R. Albacete (2002). Forecasting inflation in the european monetary union: A disaggregated approach by countries and by sectors. *The European Journal of Finance 8*(4), 402–421.

Esteves, P. S. (2013). Direct vs bottom–up approach when forecasting gdp: Reconciling literature results with institutional practice. *Economic Modelling 33*, 416–420.

Geweke, J. and G. Amisano (2010, April). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting 26*(2), 216–230.

Giannone, D., M. Lenza, D. Momferatou, and L. Onorante (2014). Short-term inflation projections: A bayesian vector autoregressive approach. *International Journal of Forecasting 30*(3), 635–644.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 243–268.

Hahn, E. and F. Skudelny (2008). Early estimates of euro area real gdp growth: a bottom up approach from the production side. Working Paper Series 0975, European Central Bank.

Hendry, D. F. and K. Hubrich (2011). Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business & Economic Statistics 29*(2).

Hubrich, K. (2005). Forecasting euro area inflation: Does aggregating forecasts by hicp component improve forecast accuracy? *International Journal of Forecasting 21*(1), 119–136.

Koop, G. and D. Korobilis (2013). Large time-varying parameter vars. *Journal of Econometrics 177*(2), 185–198.

Koop, G. M. (2013). Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics 28*(2), 177–203.

Lütkepohl, H. (1987). *Forecasting aggregated vector ARMA processes*, Volume 284. Springer Science & Business Media.

Lütkepohl, H. (2011). Forecasting nonlinear aggregates and aggregates with time-varying weights. *Jahrbücher für Nationalökonomie und Statistik*, 107–133.

Marcellino, M., J. H. Stock, and M. W. Watson (2003). Macroeconomic forecasting in the euro area: Country specific versus area-wide information. *European Economic Review 47*(1), 1–18.

Mitchell, J. and S. G. Hall (2005). Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and niesr fancharts of inflation. *Oxford bulletin of economics and statistics 67*(s1), 995–1033.

Mitchell, J. and K. F. Wallis (2011). Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics 26*(6), 1023–1040.

Perevalov, N. and P. Maier (2010). On the advantages of disaggregated data: insights from forecasting the us economy in a data-rich environment. Working Papers 10-10, Bank of Canada.

Raftery, A. E., M. Kárnỳ, and P. Ettler (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics 52*(1), 52–66.

Ravazzolo, F. and S. P. Vahey (2014). Forecast densities for economic aggregates from disaggregate ensembles. *Studies in Nonlinear Dynamics & Econometrics 18*(4), 367–381.

Zellner, A. and J. Tobias (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting 19*(5), 457–465.