Working Paper Series, N. , November 2015

Department of Statistical Sciences
University of Padua
Italy

Comparing density forecasts of aggregated time series via bootstrap

Matteo Grigoletto
Department of Statistical Sciences
Via Cesare Battisti 241
35121 Padova
Italy

Abstract: When forecasting aggregated time series, several options are available. For example, the multivariate series or the individual time series might be predicted and then aggregated, or one may choose to forecast the aggregated series directly. While in theory an optimal disaggregated forecast will generally be superior (or at least not inferior) to forecasts based on aggregated information, this is not necessarily true in practical situations. The main reason is that the true data generating process is usually unknown and models need to be specified and estimated on the basis of the available information. This paper describes a bootstrap-based procedure, in the context of vector autoregressive models, for ranking the different forecasting approaches for contemporaneous time series aggregates. Estimation uncertainty and model misspecification will be considered and the ranking will be based not only on the mean squared forecast error, but more in general on the performance of the predictive distribution. The forecasting procedures are applied to the United States aggregate inflation.

Keywords: Aggregate forecasts; Bootstrap; Density forecasts; Evaluation; Inflation

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Matteo Grigoletto
matteo.grigoletto@unipd.it

Comparing density forecasts of aggregated time series via bootstrap

Matteo Grigoletto
Department of Statistical Sciences
Via Cesare Battisti 241
35121 Padova
ITALY

Abstract: When forecasting aggregated time series, several options are available. For example, the multivariate series or the individual time series might be predicted and then aggregated, or one may choose to forecast the aggregated series directly. While in theory an optimal disaggregated forecast will generally be superior (or at least not inferior) to forecasts based on aggregated information, this is not necessarily true in practical situations. The main reason is that the true data generating process is usually unknown and models need to be specified and estimated on the basis of the available information. This paper describes a bootstrap-based procedure, in the context of vector autoregressive models, for ranking the different forecasting approaches for contemporaneous time series aggregates. Estimation uncertainty and model misspecification will be considered and the ranking will be based not only on the mean squared forecast error, but more in general on the performance of the predictive distribution. The forecasting procedures are applied to the United States aggregate inflation.

Keywords: Aggregate forecasts; Bootstrap; Density forecasts; Evaluation; Inflation

1 Introduction

The problem of forecasting aggregated time series has been studied over the last decades. A classical introduction to this subject is presented in Lütkepohl (1987), while Lütkepohl (2010) gives a recent survey. Time series aggregates are important in many fields of Economics and received new attention after the introduction of the euro-zone. When a forecast for a euro-area time series is desired, it is not obvious whether it is preferable to predict the euro-area series directly, or rather predict the time series for the single countries, and then aggregate the forecasts obtained. When considering inflation, in particular, two aggregation dilemmas are present: aggregation over subindexes and over countries (for the euro-zone).

Two strands of research have developed on this topic, the theoretical and the empirical one. Early theoretical results on aggregation versus disaggregation in forecasting can be found in Theil (1954) and Grunfeld and Griliches (1960). Other theoretical contributions include, among others, Lütkepohl (1984, 1987), Granger (1987), Giacomini and Granger (2004), Hendry and Hubrich (2011) and Sbrana (2012).

According to the theoretical literature, when the data generating process (DGP) is known, except under certain conditions, aggregating forecasts for the multivariate process is at least as efficient, in terms of mean squared forecast error (MSFE), as directly forecasting the aggregate or aggregating univariate forecasts. However, in practice the DGP is not known, and a model needs to be specified and estimated. This can be difficult, and especially so as the number of disaggregate series increases. In this case, combining multivariate forecasts may not be preferable, and this is determined by the properties of the unknown DGP. As a consequence, the choice of the best forecast is essentially an empirical issue. Thus, it is not surprising that recent empirical studies do not reach unanimous conclusions regarding the value of disaggregate information in forecasting aggregates (Lütkepohl 2010, page 55).

For example, several aggregate and disaggregate predictors of the euro-zone inflation are considered by Espasa, Senra, and Albacete (2002) and Hubrich (2005). Similarly, Hendry and Hubrich (2011) forecast the United States aggregate inflation using disaggregate sectoral data. Central banks in the Eurosystem are also recently giving attention to the aggregation of forecasts of disaggregate inflation components (see, e.g., Bruneau, De Bandt, Flageollet, and Michaux 2007, Moser, Rumler, and Scharler 2007). Besides inflation, Marcellino, Stock, and Watson (2003) also analyse real GDP, industrial production and unemployment for the euro area, while Fagan and Henry (1998) and Dedola, Gaiotti, and Silipo (2001) focus on money demand, expliciting contributions generated at a national level. Carson, Cenesizoglu, and Parker (2011) analyse the aggregate demand for commercial air travel in the United States, using airport specific data. Conclusions are far from univocal: e.g., Espasa et al. (2002) and Marcellino et al. (2003) provide evidence against the use of aggregate models and prefer forecasts based on information at country level; on the contrary, Bodo, Golinelli, and Parigi (2000) show that area wide models are preferable for forecasting industrial production. Many other examples exist in the literature.

This paper stems from the observation that, in the literature on forecasting aggregated time series, forecasts are compared generally using the MSFE for point forecasts. Lütkepohl (2010), Hendry and Hubrich (2011) and Faust and Wright (2013) are some recent examples. This can be misleading, since the uncertainty of the future is often poorly summarized by point forecasts and is instead better represented by density forecasts, which are suitable for considering a range of possible future outcomes. In other words, many realistic economic loss functions can't be reduced to the comparison of point forecasts, using the MSFE (Diebold and Mariano 2002). Besides, it should be remarked that the ranking of forecasts needs to be well grounded also in small samples, where uncertainty is greater and theoretical conclusions based on the assumption that the model is known are less reliable. These considerations suggest the use of a bootstrap approach. Since, when using bootstrap, many replicates of the future values are generated, it is natural to examine the forecast performance using prediction intervals (Kupiec 1995, Christoffersen 1998), or the whole density forecast (Diebold, Gunther, and Tay 1998). We will see that, when the sample size is sufficiently large and the comparison is based on MSFE, the bootstrap procedure described here will work as predicted by the available results on forecasting aggregates. However, as the sample size decreases (and hence model uncertainty becomes more important), or when a criterion different from MSFE is used, the approach for forecasting aggregates suggested by the bootstrap might be different. Simulations and an empirical data analysis will allow to conclude that evaluating forecasts using solely the MSFE can indeed lead to prefer procedures with worse performance in terms, e.g., of coverages of prediction intervals.

Vector autoregressive models (henceforth: VAR) will be used for forecasting. These models are widely employed and asymptotic and bootstrap procedures have been formulated (see, e.g., Kim 1999, 2004, Grigoletto 2005, Lütkepohl 2005). There also are extensive results on the bootstrap for univariate autoregressive (AR) models (e.g. Masarotto 1990, Thombs and Schucany 1990, Kabaila 1993, Breidt, Davis, and Dunsmuir 1995, Grigoletto 1998, Clements and Taylor 2001). In small samples, bootstrap methods are shown to have better properties than the asymptotic ones. Besides, these methods allow to take into account the uncertainty attributable to model estimation.

The main difficulty encountered in using the procedure described here is that it is computationally heavy. However, the widespread availability of fast computers and the possibility to parallelise bootstrap simulations seem to imply that this is not a serious drawback. It should also be considered that in many cases (e.g., when forecasting inflation: see Section 6) the frequency of the considered time series implies that it is not necessary to compare forecasting procedures repeatedly in short time intervals.

The paper is organised as follows. In Section 2 the competing forecasts for the aggregate are defined. Section 3 describes the bootstrap procedure. The criteria that will be used to compare the predictive performances of the different approaches are illustrated in Section 4. In Section 5 a simulation study is carried out, while Section 6 treats an application to the USA aggregate inflation. Conclusions follow in Section 7.

2 The model and the competing forecasts

Let us consider the VAR(p) model for a k-dimensional vector $y_t = (y_{1t}, \ldots, y_{kt})'$:

$$y_t = A_0 + A_1\, y_{t-1} + A_2\, y_{t-2} + \ldots + A_p\, y_{t-p} + \varepsilon_t \ , \tag{1}$$

where A_0 is a $k \times 1$ vector of constants, A_i for $i = 1, \ldots, p$ are $k \times k$ parameter matrices and ε_t is a $k \times 1$ vector of innovations. Innovations are i.i.d. with $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{E}(\varepsilon_t\, \varepsilon_t') = \Sigma_\varepsilon$, where Σ_ε has finite elements and is positive definite. The model is assumed to be stationary, i.e., all roots of the characteristic equation $\det(I_k - A_1\, z - \ldots - A_p z_p)$ lie outside the unit circle. In the forecasting procedure described in Section 3, the order p will be selected using the corrected AIC criterion by Hurvich and Tsai (1993).

Model (1) can be written in backward form as (Kim 1997, 1998)

$$y_t = H_0 + H_1\, y_{t+1} + H_2\, y_{t+2} + \ldots + H_p\, y_{t+p} + \nu_t \ , \tag{2}$$

where H_0 is a $k \times 1$ vector of constants, H_i for $i = 1, \ldots, p$ are $k \times k$ coefficient matrices, $\mathrm{E}(\nu_t) = 0$ and $\mathrm{E}(\nu_t\, \nu_t') = \Sigma_\nu$, where Σ_ν has finite elements and is positive definite. Bootstrap samples will be generated from the backward representation (2). This will ensure that the last observations in the pseudo-datasets are the same as in the original sample, consistently with the property that VAR forecasts are conditional on past observations (see Thombs and Schucany 1990, in a univariate AR context and Kim 1999, 2004, and Grigoletto 2005 for VAR forecasting).

The iterated h-step-ahead predictors $\hat{y}_T(h)$ for y_{T+h}, $h = 1, \ldots, H$, will be obtained from (1) using the ordinary least squares estimators $\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_p$:

$$\hat{y}_T(h) = \hat{A}_0 + \hat{A}_1\, \hat{y}_T(h-1) + \ldots + \hat{A}_p\, \hat{y}_T(h-p) \ ,$$

where $\hat{y}_T(j) = y_{T+j}$ for $j \leq 0$. In an extensive comparison of direct and iterated multistep forecasts for AR models, Marcellino, Stock, and Watson (2006) find that the iterated forecasts typically outperform direct forecasts.

For bootstrap forecasts, bias-corrected estimators of $A_0, A_1, \ldots, A_p$ will be used to compensate for the bias in least squares estimators (the bootstrap-after-bootstrap procedure proposed by Kilian 1998b, will be employed):

$$\hat{y}^c_T(h) = \hat{A}^c_0 + \hat{A}^c_1\, \hat{y}^c_T(h-1) + \ldots + \hat{A}^c_p\, \hat{y}^c_T(h-p) \ ,$$

where $\hat{A}^c_0, \hat{A}^c_1, \ldots, \hat{A}^c_p$ are the bias-corrected estimators and $\hat{y}^c_T(j) = y_{T+j}$ for $j \leq 0$. See Section 3 for further details on the bootstrap-after-bootstrap bias correction.

We will focus on the following linear transformation of y_t:

$$x_t = F_t\, y_t \ , \tag{3}$$

where F_t is an $m \times k$ aggregation matrix of full rank m. The matrix F_t can be time-varying (as in the application in Section 6) and unknown (as in DGP$_6$ in Section 5). While the considered methods can be applied in the $m > 1$ case (e.g., by considering prediction regions instead of prediction intervals), henceforth discussion will be essentially constrained to $m = 1$.

It should be noted that, while a linearly transformed VARMA(p, q) process has a finite order VARMA representation, an analogous property does not hold for finite order VAR processes (Lütkepohl 1987). Therefore, x_t will in general be in the VARMA class and will only be approximated by a VAR model. It is well known that, in sharp contrast with the ease of identification and estimation of VAR models, the nonuniqueness of a VARMA representation makes identifying and estimating VARMA models quite difficult (e.g. Lütkepohl 2005). Besides, the use of VARMA models, while implying a more parsimonious representation of the underlying process, is far from guaranteeing a significant improvement in forecasts: see Athanasopoulos and Vahid (2008). For these reasons, for our forecasting purposes we will represent x_t with equations analogous to (1) and (2). See also Lewis and Reinsel (1985) and Lütkepohl (1985), who consider the MSFE when the true DGP, which may be of the VARMA type, is approximated by a finite order VAR. While it is rather unusual to use VMA models in a forecasting context, results concerning these models can be found in Sbrana and Silvestrini (2009) and Sbrana (2012).

For predicting the aggregated time series x_t, three alternative approaches are considered (Lütkepohl 2010):

1. Forecasting the disaggregated multivariate model and then aggregating the forecasts:
$$_d\hat{x}_t^c(h) = F_t \, \hat{y}_t^c(h) \; .$$

2. Forecasting the individual disaggregated variables based on univariate models and aggregating the forecasts. The bias-corrected h-step-ahead predictor for the j-th, $j = 1, \ldots, k$, component of y_t will be denoted by $\hat{y}_{j,t}^c(h)$. The corresponding forecast of x_t will be
$$_u\hat{x}_t^c(h) = F_t \, (\hat{y}_{1,t}^c(h), \ldots, \hat{y}_{k,t}^c(h))' \; .$$

3. Forecasting the aggregate x_t directly. The bias-corrected version of these forecasts will be denoted by $\hat{x}_t^c(h)$.

Notably, the univariate processes $y_{j,t}$, $j = 1, \ldots, k$, will belong to the ARMA class (Zellner and Palm 1974 and, e.g., Franses 1998, p. 198). The same is true for x_t when $m = 1$. Of course, identification and estimation of ARMA models is not as involved as it is for their multivariate counterpart. Besides, bootstrap procedures have been proposed that allow resampling from ARMA models: see

Pascual, Romo, and Ruiz (2004). The procedure proposed by these authors, however, differs from the one described here, also because it does not use bias correction, which is an important step in bootstrap prediction (Clements and Taylor 2001, discuss bias correction in detail). Also, the results in Kim (2002) show that, even for ARMA processes, building the bootstrap on the univariate equivalents of (1) and (2) yields good prediction performances. Therefore, here forecasts will be based on AR models also in the univariate case.

The theoretical literature on VARMA processes shows that, when the DGP is known and the MSFE is employed as evaluation criterion, ${}_d\hat{x}^c_t(h)$ is preferable to $\hat{x}^c_t(h)$ and ${}_u\hat{x}^c_t(h)$, while no unique ranking exists between $\hat{x}^c_t(h)$ and ${}_u\hat{x}^c_t(h)$. In practice, however, the DGP must be estimated and, especially when the number of disaggregate variables is large, ${}_d\hat{x}^c_t(h)$ might well produce inferior results with respect to the other approaches. Also, ${}_d\hat{x}^c_t(h)$ can be expected to yield an improved performance (in terms of MSFE) only if the disaggregate series are intertemporally related and heterogeneous. When, on the contrary, the component series are described by similar univariate models, it should be desirable to use $\hat{x}^c_t(h)$, i.e. to forecast the aggregate series directly. The forecast ${}_u\hat{x}^c_t(h)$, obtained with individual forecasts of the univariate components of y_t, is likely to become preferable when the component series have weak relations and their marginal DGPs are sufficiently different. In fact, Lütkepohl (2010), p. 46, shows that, when y_t is formed by independent components, summing up the forecasts for the individual components will be strictly more efficient than forecasting the sum directly only if the components have distinct serial correlation structures. See Lütkepohl (2010) and Sbrana (2012) for more details.

While these suggestions define useful guidelines, some questions remain unanswered. In practice, it is hard to define clear borderlines between different situations. For example, when are the heterogeneity and intertemporal relations among component series sufficient to guarantee superiority of the disaggregate forecast ${}_d\hat{x}^c_t(h)$? This and similar questions are largely empirical. Also, there are situations for which no guidance is provided by the theory: it is unclear what would be a desirable approach when there are many component series with a strong intertemporal relation. Finally, no guidance is provided when the loss function underlying the MSFE is inappropriate, as when interval forecasts are desired. The bootstrap procedure described here aims at giving a practical answer to these questions.

3 The bootstrap procedure

The bootstrap procedure adopted here will use the residual (nonparametric) method. Since bootstrap generation is based on the backward representation (2), forecasts computed are conditional on the last p observations in the original observed sample. Bootstrap-after-bootstrap is used for bias-correction

(Kilian 1998b). The bias correction is useful to improve the small sample properties of bootstrap density forecasts. The improvement yielded by bias correction is found to be particularly relevant when the VAR model has near unit roots (see Kim 2001, who describes the benefits of bootstrap-after-bootstrap for interval forecasting).

The bootstrap procedure is composed of the following steps:

1. The T observations are used to determine the model order, obtaining $\hat{p}$. Many selection criteria are available (see e.g. Konishi and Kitagawa 2008). The AIC information criterion has proven to have a good performance in VAR order selection, when compared to other consistent criteria (Kilian 1998a, 2001). The order selection criterion used here is a corrected form of AIC (AIC$_C$), which performs better in small samples, counteracting the overfitting nature of AIC. The AIC$_C$ was introduced by Hurvich and Tsai (1993). In their work, these authors remark that consistency can be obtained only at the cost of asymptotic efficiency and that, of the two properties, asymptotic efficiency is the more desirable. This is because, in practice, there will often be no true order. Therefore, efficiency (Shibata 1980) becomes crucial.

2. The parameters of the forward and backward VAR models (1) and (2) are estimated by least squares. The least squares estimators and the residuals for the forward and backward models are indicated by $\hat{A}_0, \dots, \hat{A}_{\hat{p}}$, $\{\hat{\varepsilon}_t\}$ and by $\hat{H}_0, \dots, \hat{H}_{\hat{p}}$, $\{\hat{\nu}_t\}$, respectively. The residuals are rescaled as suggested in Thombs and Schucany (1990).

3. B_0 pseudo-datasets are generated from

   $$y_t^* = \hat{H}_0 + \hat{H}_1\, y_{t+1}^* + \hat{H}_2\, y_{t+2}^* + \dots + \hat{H}_{\hat{p}}\, y_{t+\hat{p}}^* + \nu_t^* \ ,$$

   where ν_t^* is a random draw, with replacement, from $\{\hat{\nu}_t\}$, and the $\hat{p}$ initial values $y_{n-\hat{p}+1}^*, \dots, y_n^*$ are set equal to $y_{n-\hat{p}+1}, \dots, y_n$ (Kim 1999 and Grigoletto 2005).

   For each pseudo-dataset, the parameters of model (2) are estimated, obtaining $\hat{H}_0^*, \dots, \hat{H}_{\hat{p}}^*$. Denoting by $\overline{\hat{H}}_j^{\,*}$ the sample mean of the B_0 replications of $\hat{H}_j^*$, we have $\text{bias}(\hat{H}_j) = \overline{\hat{H}}_j^{\,*} - \hat{H}_j$, for $j = 0, \dots, \hat{p}$. This bias is used to compute the bias corrected estimates $\{\hat{H}_j^c\}$, with the procedure introduced by Kilian (1998b). This procedure also performs a stationarity adjustment, to ensure that the bias correction does not push stationary estimates into the non-stationary region of the parameter space. The residuals $\{\hat{\nu}_t^c\}$ are computed from the bias-corrected estimates $\{\hat{H}_j^c\}$. Analogous steps lead, for the forward model (1), to the computation of $\text{bias}(\hat{A}_j) = \overline{\hat{A}}_j^{\,*} - \hat{A}_j$, of the bias-corrected estimates $\{\hat{A}_j^c\}$ and of the corresponding residuals $\{\hat{\varepsilon}_t^c\}$.

4. B pseudo-datasets are generated from

$$y_t^{*c} = \hat{H}_0^c + \hat{H}_1^c\, y_{t+1}^{*c} + \hat{H}_2^c\, y_{t+2}^{*c} + \ldots + \hat{H}_{\hat{p}}^c\, y_{t+\hat{p}}^{*c} + \nu_t^{*c} \ ,$$

   where ν_t^{*c} is a random draw, with replacement, from $\{\hat{\nu}_t^c\}$, and the $\hat{p}$ initial values $y_{n-\hat{p}+1}^{*c}, \ldots, y_n^{*c}$ are set equal to $y_{n-\hat{p}+1}, \ldots, y_n$. For each pseudo-dataset, the forward model parameters are estimated by least squares, obtaining the estimates $\tilde{A}_j^*$; these estimates are then corrected using bias$(\hat{A}_j)$ computed in step 3, thus obtaining $\tilde{A}_j^{*c}$, $j = 0, \ldots, \hat{p}$.

5. The bootstrap forecast replicates made at time T for the forecast horizon h are defined as

$$\hat{y}_T^{*c}(h) = \tilde{A}_0^{*c} + \tilde{A}_1^{*c}\, \hat{y}_T^{*c}(h-1) + \tilde{A}_2^{*c}\, \hat{y}_T^{*c}(h-2) + \ldots + \tilde{A}_{\hat{p}}^{*c}\, y_T^{*c}(h-\hat{p}) + \varepsilon_{T+h}^{*c} \ ,$$

   where ε_{T+h}^{*c} is a random draw, with replacement, from $\{\hat{\varepsilon}_t^c\}$ and $y_T^{*c}(j) = y_{T+j}$ for $j \leq 0$.

6. The bootstrap replicates for the forecasts ${}_d\hat{x}_t^c(h)$, i.e. the forecasts of the aggregated time series based on the disaggregated model, can now be computed as

$${}_d\hat{x}_T^{*c}(h) = F_t\, \hat{y}_T^{*c}(h) \ .$$

7. Perform steps 1–5 for each univariate series $y_{j,t}$, $j = 1, \ldots, k$. In this case, the aggregate forecast replicates are defined as

$${}_u\hat{x}_T^{*c}(h) = F_t\, (\hat{y}_{1,T}^{*c}(h), \ldots, \hat{y}_{k,T}^{*c}(h))' \ .$$

8. Perform steps 1–5 for the aggregated time series x_t. These aggregate forecast replicates will be indicated by $\hat{x}_T^{*c}(h)$.

4 Comparing the predictive performances

In this section we examine the predictive performance of the three competing forecasts of the aggregated time series. For the sake of simplicity, we will confine ourselves to the $m = 1$ (i.e. unidimensional aggregate) case: this is the situation most often considered in the relevant literature and also analysed here in the simulation study and in the application.

When a prediction for the conditional mean is desired, straightforward evaluation criteria can be computed by simply comparing the realised and forecasted values on the basis of a suitable loss function. If the loss function is quadratic, the evaluation criterion becomes the MSFE, which has traditionally been used in the literature on aggregated time series. However, as remarked by Christoffersen (1998), users should not be content with a point forecast, since such a forecast describes only one (even if important) possible outcome.

The bootstrap procedure described above yields many forecast replicates. In this context, it is easy to propose a range of likely outcomes, therefore allowing users to be thoroughly prepared to future events, even if they differ from what is usually expected. In particular, in the bootstrap framework, computing prediction intervals or density forecasts are simple tasks. It is therefore natural, and useful, to compare different prediction methods assessing their performance in accomplishing these tasks.

It should also be noted that often the appropriateness of point forecasts is evaluated by studying the correlation structure of prediction errors, using the well known result stating that the h-step error should be MA$(h-1)$. This assessment, however, gives no hint on the performance of the forecast distribution as a whole. For example, the prediction error might be MA$(h-1)$ and, at the same time, the coverage of prediction intervals be completely wrong. The evaluation criteria shown below assess the performance of density forecasts, rather than point forecasts.

Prediction intervals. In most of the literature, evaluation of interval forecasts proceeds by simply comparing the nominal and true coverages. In the following, this amounts to testing an unconditional coverage hypothesis. However, intervals should be evaluated considering that they are based on a time-dependent information set. For example, intervals should be narrow in times of low uncertainty, and wide when uncertainty is higher. Prediction intervals that fail to account for this might be correct on average (and pass the unconditional coverage evaluation), but in any given period the conditional coverage will be incorrect. Since the 1990's a variety of tests have been proposed to measure accuracy of prediction intervals, assessing also conditional coverage. Christoffersen (1998) introduced a model-free approach based on the concept of violation, which occurs when the *ex-post* realisation of a variable does not lie in the *ex-ante* forecast interval. The conditional validity of prediction intervals can be reduced to the problem of assessing whether the following hypotheses are jointly satisfied:

i) *unconditional coverage hypothesis*: the probability of an observation to fall in the corresponding prediction interval must be equal to the coverage rate;

ii) *independence hypothesis*: violations of prediction intervals, observed at different dates, for the same coverage rate, must be independent (i.e. violations should not "cluster").

A test of unconditional coverage was initially proposed by Kupiec (1995), while Christoffersen (1998) proposes a procedure to jointly test the two hypotheses, thus assessing correct conditional coverage (CC). In the following, the

Monte Carlo versions of the CC test will be used, as suggested by Christoffersen and Pelletier (2004), who employ the technique described by Dufour (2006) to overcome the possible scarcity of violations. When the number of available out-of-sample forecasts is not large, this is especially important. The CC test procedure needs to be modified when $h \geq 2$, since in this case optimal forecasts at horizon h are characterised by autocorrelation of order $h-1$. Diebold et al. (1998) recommend using an approach based on Bonferroni bounds. Let us denote by I_T the indicator variable used in the test procedure, which equals 1 when there is no violation (i.e. the forecast interval for horizon h, computed at the forecast origin T, contains the realisation y_{T+h}), and 0 otherwise. The indicator variables are divided in h sub-groups, which are independent under the null hypothesis: $(I_1, I_{1+h}, I_{1+2h}, \ldots)$, $(I_2, I_{2+h}, I_{2+2h}, \ldots)$, $\ldots$, $(I_h, I_{2h}, I_{3h}, \ldots)$. The null hypothesis of correct conditional coverage is then rejected, at an overall significance level bounded by α, when it is rejected for any of the subgroups at the α/h significance level. The use of Monte Carlo p-values, as described above, is particularly recommended as h increases and the number of indicator variables in each subgroup becomes small.

Density forecasts. Diebold et al. (1998) remark that the method proposed by Christoffersen (1998) allows to evaluate whether a series of prediction intervals are correctly conditionally calibrated only at a specified confidence level. This leads to the problem of density forecasts evaluation, which corresponds to the simultaneous conditional calibration of all possible interval forecasts. Diebold et al. (1998) proposed to evaluate density forecast estimates using the probability integral transform (PIT) which, when $h = 1$, is defined as:

$$z_T = \int_{-\infty}^{x_{T+1}} \hat{f}_{T+1}(u|\Omega_T)\, du \ ,$$

where Ω_T is the information available at the forecast origin T and $\hat{f}_{T+1}(\cdot|\Omega_T)$ is a one-step forecast density, which in our case will be one of the three bootstrap based forecasts described in Section 3. Let $f_{T+1}(\cdot|\Omega_T)$ denote the true forecast density. If the forecasting model is correct, Diebold et al. (1998) show that the PIT series $\{z_T\}$ is $i.i.d.$ $U(0,1)$. Evaluating the goodness of estimated forecast densities can therefore be based on the assessment of the uniformity and independence properties of $\{z_T\}$. To this aim, Diebold et al. (1998) employed mainly graphical tools. More recently (e.g. Clements, Franses, Smith, and Van Dijk 2003, Siliverstovs and Dijk 2003) formal tests have been applied: the most commonly employed one is the Kolmogorov-Smirnov test (KS). The KS goodness-of-fit measure has been discussed extensively in the literature (see e.g. Conover 1999); here, the KS test is implemented using the technique developed in Wang, Tsang, and Marsaglia (2003).

Since the KS test assumes independence, as suggested by Diebold et al. (1998) we will test for the presence of serial correlation in the PIT, assuming that the process $\{z_T\}$ has this representation:

$$z_T - \bar{z} = \gamma_1 (z_{T-1} - \bar{z}) + \ldots + \gamma_q (z_{T-q} - \bar{z}) + \varepsilon_T$$

Then, a Lagrange multiplier test (LM) is carried out, with the null hypothesis that $\gamma_i = 0$, $i = 1, \ldots, q$, against the alternative that $\gamma_i \neq 0$ for at least one $i \in \{1, \ldots, q\}$. This test is very widespread, especially in the econometric literature: see e.g. Godfrey (1986), pp. 116–117, or Lütkepohl (2004), pp. 44–45. While several asymptotically equivalent algorithms are available for the LM test, Kiviet (1986) suggests that a simple F version of the test is better behaved in small samples. When $h > 1$, we proceed as described above for prediction intervals, partitioning the PITs in h subgroups: $(z_1, z_{1+h}, z_{1+2h}, \ldots)$, $(z_2, z_{2+h}, z_{2+2h}, \ldots)$, $\ldots$, $(z_h, z_{2h}, z_{3h}, \ldots)$. Then, a test with overall significance level bounded by α is performed. Rejection of the null hypothesis suggests the presence of serial correlation that is not explained by the forecasting model.

Recently, Chevillon (2014) remarked that the method based on non-overlapping subsamples of PITs, described above for horizons $h > 1$, implies that the conditional information sets entering the forecasting models are non-overlapping themselves. The solution proposed by Chevillon (2014) is to fit an MA$(h-1)$ model to the estimated h-step forecast errors, and then compute the PIT for the residuals of this model. This solution is easy to implement in the context of the present paper. However, the estimation of the auxiliary MA$(h-1)$ model injects uncertainty in the density forecast evaluation procedure. As a consequence, the uncertainty (that we would like to take into account with the bootstrap) in the estimation of the main prediction model becomes inextricably confused with the uncertainty in the estimation of the MA$(h-1)$ auxiliary model. Also, Chevillon (2014) does not investigate whether the MA$(h-1)$ correction affects the power of the tests on the PITs. This investigation is especially important here, where test results are used to rank prediction models. For these reasons, while the remark by Chevillon (2014) is important, the solution proposed by this author is not adopted here.

5 Simulation study

The simulation framework is designed to compare the performances of ${}_d\hat{x}_t^c(h)$, ${}_u\hat{x}_t^c(h)$ and $\hat{x}_t^c(h)$ in small samples and when the DGPs are unknown. The experiments suggested by Lütkepohl (1984), Sbrana and Silvestrini (2009) and Hendry and Hubrich (2011) will be considered and extended by also assessing prediction intervals and forecast densities, with the techniques discussed in the previous sections. Differently from the above contributions, here models are estimated recursively, i.e. the available sample until time T is first used to estimate the model and generate forecasts, then the sample up to time $T+1$ is used, and so on. This is meant to represent a real life situation in which

DGP	Coefficients
DGP$_1$	$\alpha_{11} = \alpha_{22} = -0.5$; $\alpha_{12} = \alpha_{21} = 0$
DGP$_2$	$\alpha_{11} = 0.5$; $\alpha_{22} = -0.3$ $\alpha_{12} = -0.66$; $\alpha_{21} = -0.5$
DGP$_3$	$\psi_{11} = \psi_{22} = 0.5$; $\psi_{12} = \psi_{21} = 0$
DGP$_4$	$\psi_{11} = 0.3$; $\psi_{22} = -0.5$ $\psi_{12} = -0.66$; $\psi_{21} = -0.5$
DGP$_5$	$\psi_{11} = 0.7$; $\psi_{22} = 0.3$ $\psi_{12} = 0.2$; $\psi_{21} = 0.32$

Table 1: Definition of the DGPs used in the simulation experiment, where $A_1 = \{\alpha_{ij}\}$ and $C_1 = \{\psi_{ij}\}$.

an out-of-sample period can be used to choose between competing forecasting procedures.

Design of the experiment. We are going to consider two-dimensional and five-dimensional VAR(1) DGPs. If we indicate by α_{ij} the generic element of A_1 in equation (1), the two-dimensional VAR(1) DGPs are defined in Table 1. The elements of the intercept vector A_0 are all set equal to 1. While only autoregressive processes are used to fit and forecast, we are going to take into account possible model misspecification by also employing VMA(1) DGPs, having the following general structure: $y_t = C_0 + \varepsilon_t + C_1 \varepsilon_{t-1}$. Table 1 defines the VMA(1) DGPs adopted, with $\{\psi_{ij}\}$ denoting the generic element of C_1. Also in this case, all the elements of the intercept vector C_0 are set equal to 1. Innovations are Gaussian i.i.d. with $\mathrm{E}(\varepsilon_t) = 0$ and $\Sigma_\varepsilon = I$, except for DGP$_5$, for which the non-diagonal elements in Σ_ε are set equal to 0.3.

DGP$_i$ in Table 1 is from Lütkepohl (1984) for $i = 1, \dots, 4$, and from Sbrana and Silvestrini (2009) for $i = 5$. All DGPs in Table 1 are two-dimensional: four-dimensional VAR(1) DGPs from Hendry and Hubrich (2011) were also considered and did not change the general conclusions. While here only DGP$_5$ has a non-diagonal covariance matrix Σ_ε, Sbrana and Silvestrini (2009) used the same non-diagonal matrix also for other VMA(1) processes. Similar experiments in the present framework did not change the ranking of the different forecasting techniques under study.

For the DGPs in Table 1, we will define the aggregation matrix in Equation (3) as $F_t = F = (1, \dots, 1)$, so that x_t is a simple sum of the components in y_t. It should be noted, however, that in many cases the aggregation weights are time-varying. For instance, there are several proposals to aggregate the euro-area gross domestic product using stochastic weights (see Winder 1997, Beyer, Doornik, and Hendry 2001, and Anderson, Dungey, Osborn, and Vahid 2011). This situation can easily be accommodated in the present framework with the approach by (Lütkepohl 2011). Let us define

$$F_t = \begin{bmatrix} w_t' \\ I_k \end{bmatrix}$$

where $w_t = (w_{1t}, \ldots, w_{kt})'$ is a vector of stochastic weights and I_k is the identity matrix of order k. Then, the first element of $y_t^w = F_t\, y_t$ is the aggregate of interest, and the other elements replicte the series y_t. Hence, considering y_t^w instead of y_t leads to an extension of the information set, since the information in the process w_t is also being incorporated. When stochastic weights are considered, only the predictors based on forecasting the full process y_t^w and on forecasting the aggregate $w_t'\, y_t$ directly are available. In fact, since the weights are unknown, aggregating univariate forecasts of the disaggregate component series $y_{i,t}$ is not possible.

Stochastic aggregation weights will be represented here by DGP$_6$ (from Lütkepohl 2011), a VAR(1) process of dimension $k = 5$, with the intercept vector set to zero and

$$A_1 = \text{diag}(0.9, -0.9, 0.5, -0.5, 0.5)\ .$$

Here, the serial correlation structure differs remarkably from one component to the other, representing cases in which the univariate series have distinct behaviours (e.g., energy and food prices might well have dissimilar DGPs: see Section 6). In order to generate stochastic weights, first a vector $\tilde{w}_t$ is drawn from a k-dimensional Gaussian distribution: $\tilde{w}_t \sim \mathcal{N}((10, 5, 5, 5, 5)', 0.1 \cdot I_k)$, so that the first univariate component has a larger weight on average and the variation in the aggregation weights for adjacent periods tends to be small, as is often the case in empirical time series. The negative elements of $\tilde{w}_t$ are replaced by zero and then the aggregation weights are defined as

$$w_t = \frac{\tilde{w}_t}{\sum_{i=1}^k \tilde{w}_{it}}\ .$$

General results (Lütkepohl 1987, Sbrana and Silvestrini 2009), concerning the MSFE and derived under the assumption that the involved processes are known, allow to describe, for each DGP, the theoretical performances of the forecasting procedures. Since, when the DGP is known, there is no uncertainty originated from model specification and estimation, in the present paragraph the three forecasts will be denoted by ${}_dx_t(h)$, ${}_ux_t(h)$ and $x_t(h)$. For DGP$_i$, $i = 1, 3$, the components are independent and have homogeneous correlation structures; as a consequence, the three forecasts yield the same MSFE. On the contrary, for DGP$_i$, $i = 2, 4, 6$, the MSFE for ${}_dx_t(h)$ will be lower than that for the other two predictors. Finally, DGP$_5$ is interesting since ${}_ux_t(h)$ and $x_t(h)$ have the same one-step MSFE, even if the forecasts do not coincide (Sbrana and Silvestrini 2009). It should be noted that, while MSFEs can be rather dissimilar if the prediction horizon h is small, differences vanish as h increases (see Lütkepohl 1987, Section 4.2.2).

Now, the question we would like to answer is: the above theoretical rankings, based on the MSFE for point forecasts when the DGP is known, are still valid when the evaluation criterion is based on prediction intervals or density forecasts, and the DGP needs to be estimated? The bootstrap procedure, detailed here, is applied as a response to this question.

The results shown are obtained for samples with an initial size $N = 100$. Samples are recursively expanded to include an out-of-sample period having length $S = 100$. This means, e.g. for $h = 1$, that first $y_1, \ldots, y_N$ are used to forecast y_{N+1}, then $y_1, \ldots, y_{N+1}$ are used to forecast y_{N+2}, and so on until the last step, in which $y_1, \ldots, y_{N+S-1}$ are used to forecast y_{N+S}. Simulation experiments with $N = 200$ were also performed and led to analogous qualitative conclusions.

Prediction intervals and density forecasts are computed with $B = 2000$ bootstrap replicates; bias correction is based on $B_0 = 1000$ replicates. Repeating several times the same experiment did not lead to significant changes in the results, implying that the number of bootstrap replicates is sufficient for the analysis.

As discussed above, models are estimated by least squares, while the AIC$_c$ criterion by Hurvich and Tsai (1993) is adopted for order selection, with maximum number of lags $p = 10$.

Comparison of prediction methods. We will consider prediction horizons $h = 1, 2, 3$. To evaluate point forecasts, we will use the root MSFE (RMSFE), which is obviously equivalent to the MSFE, but is used more often in the applied literature (e.g. Hendry and Hubrich 2011) and will also be adopted later in the application. Therefore, the out-of sample period is used to compute the RMSFE, with respect to point forecasts given by the mean of the bootstrap forecast distributions. More formally:

$$\text{RMSFE}_h = \sqrt{\frac{\sum_{T=N}^{N+S-h}[x_{T+h} - \hat{x}_T(h)]^2}{S - h + 1}}$$

where $\hat{x}_T(h)$ is the average of the B values obtained for ${}_d\hat{x}_T^{*c}(h)$, ${}_u\hat{x}_T^{*c}(h)$ or $\hat{x}_T^{*c}(h)$. The RMSFEs are then averaged over $M = 200$ Monte Carlo repetitions.

The B values generated for ${}_d\hat{x}_T^{*c}(h)$, ${}_u\hat{x}_T^{*c}(h)$ and $\hat{x}_T^{*c}(h)$ form a bootstrap forecast distribution. Hence, the evaluation criteria described in Section 4 are also available. The outcomes of the conditional coverage (CC), Kolmogorov-Smirnov (KS) and Lagrange multiplier (LM) tests, used to evaluate the performance of the prediction intervals and forecast distributions, are described by their p-values. The performance measure adopted here is the one used in Engle (2002) (p. 343 and Tables 2 to 5):

$$\tilde{p}_\alpha = \frac{\sum_{i=1}^{M} I(p_i < \alpha)}{M}$$

where p_i is the p-value for the test at the i-th Monte Carlo iteration, $I(\cdot)$ is the indicator function and α is the chosen significance level. Hence, $\tilde{p}_\alpha$ is the fraction of α-level tests rejecting. A prediction method is preferable when it yields a smaller value of $\tilde{p}_\alpha$. It must be strongly emphasized that $\tilde{p}_\alpha$ is a performance measure, and not a p-value.

Simulation results. [1] Table 2 shows the RMSFEs ratios, with respect to the RMSFE for $\hat{x}^c_t(h)$, for the different processes, when estimation uncertainty comes into play. When the component univariate processes are independent and homogeneous (DGP$_i$, $i = 1, 3$), the three methods yield similar RMSFEs. On the contrary, for DGP$_i$, $i = 2, 4, 6$, which have dependent and/or inhomogeneous component processes, at the prediction horizon $h = 1$, ${}_d\hat{x}^c_t(h)$ produces smaller RMSFEs. As h increases, differences become less marked. These results are of course consistent with those found in previous literature, as e.g. Lütkepohl (1984, 2011) and Hendry and Hubrich (2011). Also, for DGP$_5$, we see that the RMSFEs observed for ${}_u\hat{x}^c_t(h)$ and $\hat{x}^c_t(h)$ are very similar, especially at $h = 1$, in agreement with the theory in Sbrana and Silvestrini (2009).

Let us now turn to the more general evaluation of the performance of the predictive distribution. Table 3 displays the fraction of 5% tests rejections (i.e. $\tilde{p}_{0.05}$) for the Kolmogorov-Smirnov (KS), LM (no serial correlation of order 1) and Christoffersen (CC, conditional coverage, 95%) tests on the predictive distributions, described in Section 4.

The KS tests show similar performances for the three methods when the component processes are independent and homogeneous (DGP$_i$, $i = 1, 3$). For the DGPs with interrelated component processes, it can be noted that ${}_u\hat{x}^c_t(h)$ yields better, or at least comparable, KS test results for DGP$_i$, $i = 2, 4$, while it performs worse, and rather markedly so, for DGP$_5$. This result is particularly interesting since, as mentioned above, DGP$_5$ is designed to guarantee the same theoretical one-step forecasting performance, in terms of MSFE, for ${}_u\hat{x}^c_t(h)$ and $\hat{x}^c_t(h)$. As can be seen, this does not necessarily translates into comparable performances when an evaluation criterion different from the MSFE, as the KS criterion, is considered. An analogous reasoning applies to DGP$_6$, for which using the disaggregate information yields an evident advantage in terms of RMSFE (notably for $h = 1$), but this advantage vanishes, at least at $h = 1$, when the KS test results are used as ranking criterion (the same is true for the CC test: see below).

It should be noted that, when the component processes are interrelated (DGP$_i$, $i = 2, 4, 5$), the LM test detects the inability of ${}_u\hat{x}^c_t(h)$ to fully represent the linear dependence structure, yielding fractions of rejections $\tilde{p}_{0.05}$ for the LM test which are sensibly larger, especially at $h = 1$, for ${}_u\hat{x}^c_t(h)$ than for the other methods.

[1] R (R Core Team 2014) was used to run the simulations. It should be mentioned that while the necessary computations are rather CPU-intensive, they are easy to parallelise. In fact, the results presented here have been obtained on a computer cluster. The parallelisation was based on the snow (Tierney, Rossini, Li, and Ševčíková 2011) R package.

	DGP_1			DGP_2			DGP_3		
	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	1.006	1.004	1.004	0.843	0.981	0.957	1.008	1.007	1.006
${}_u\hat{x}_t^c(h)$	1.001	1.001	1.001	1.012	0.993	0.991	1.002	1.001	1.000
	DGP_4			DGP_5			DGP_6		
	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	0.892	1.018	1.017	0.996	1.020	1.015	0.787	0.965	0.914
${}_u\hat{x}_t^c(h)$	1.010	0.990	0.994	1.001	0.991	0.992	-	-	-

Table 2: RMSFE ratios, with respect to the RMSFE for $\hat{x}_t^c(h)$, in $S = 100$ out-of-sample steps (average over $M = 200$ Monte Carlo runs). The initial sample size is $N = 100$.

The KS and LM tests concern the whole forecast distributions. However, when prediction intervals of a specified confidence level are of interest, results for the CC test are particularly relevant (results shown in Table 3 concern the 95% confidence level). In this regard, $\hat{x}_t^c(h)$ performs in most cases similarly or better than the other two methods. The preferability of $\hat{x}_t^c(h)$ is particularly marked for DGP_4 and DGP_5, as for these processes the CC test fractions of rejections $\tilde{p}_{0.05}$ are usually rather smaller for $\hat{x}_t^c(h)$, even for $h = 2, 3$, while for the other DGPs differences among the forecasting models are generally less noticeable when $h = 2, 3$. Most notably, for DGP_5, according to the CC test results, the preferability of $\hat{x}_t^c(h)$ with respect to ${}_u\hat{x}_t^c(h)$ is evident (as it is for the KS criterion), even if these procedures are theoretically equivalent in terms of MSFE (at $h = 1$). In summary, when computing confidence intervals, $\hat{x}_t^c(h)$ appears to be generally a safe choice, at least for the DGPs considered. This result partially contradicts the suggestions grounded on the theoretical results recalled above for the MSFE criterion, which therefore should not be the only criterion used to rank the alternative forecasting procedures, at least when decisions are based not on the consideration of a single possible future outcome, but rather on a range of likely outcomes.

6 Application

The techniques illustrated in the previous sections are now applied for predicting aggregate inflation for the all items U.S. consumer price index (CPI). As discussed in the introduction, conclusions on the preferability of aggregate or disaggregate models for forecasting are far from univocal and depend, e.g., on the period and the forecast horizon considered. For example, Hubrich (2005), analysing euro area inflation, finds that neither approach works necessarily better, while results in Bermingham and D'Agostino (2011) are more in favour of aggregating disaggregate forecasts. See Faust and Wright (2013) for a recent

		DGP$_1$			DGP$_2$			DGP$_3$		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
	KS	.050	.105	.135	.030	.075	.150	.060	.070	.150
${}_d\hat{x}^c_t(h)$	LM	.050	.120	.115	.090	.105	.105	.045	.135	.100
	CC	.095	.090	.145	.070	.135	.200	.140	.235	.265
	KS	.070	.135	.160	.015	.040	.055	.045	.045	.140
${}_u\hat{x}^c_t(h)$	LM	.050	.105	.140	.975	.145	.400	.045	.145	.120
	CC	.070	.095	.135	.140	.085	.155	.080	.190	.270
	KS	.050	.115	.125	.115	.085	.170	.055	.055	.155
$\hat{x}^c_t(h)$	LM	.050	.120	.145	.075	.140	.135	.060	.155	.115
	CC	.095	.120	.175	.075	.115	.225	.100	.195	.265
		DGP$_4$			DGP$_5$			DGP$_6$		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
	KS	.015	.065	.110	.070	.080	.130	.060	.095	.085
${}_d\hat{x}^c_t(h)$	LM	.045	.085	.155	.040	.120	.130	.045	.080	.115
	CC	.100	.185	.285	.140	.190	.300	.125	.150	.190
	KS	.010	.090	.130	.115	.100	.160	-	-	-
${}_u\hat{x}^c_t(h)$	LM	.940	.110	.150	.495	.105	.145	-	-	-
	CC	.110	.185	.255	.150	.325	.400	-	-	-
	KS	.030	.060	.105	.045	.095	.115	.050	.145	.110
$\hat{x}^c_t(h)$	LM	.055	.105	.145	.040	.095	.155	.085	.115	.130
	CC	.085	.130	.215	.105	.175	.240	.090	.175	.210

Table 3: Fraction of 5% tests rejecting for the Kolmogorov-Smirnov (KS), LM (no serial correlation of order 1) and Christoffersen (CC, conditional coverage, 95%) tests on the predictive distributions. Results are obtained for $S = 100$ out-of-sample steps and are averaged over $M = 200$ Monte Carlo runs. The initial sample size is $N = 100$.

discussion on inflation forecasting.

As in Hendry and Hubrich (2011), the data set used in the present analysis includes the all items U.S. consumer price index (CPI) as well as four subcomponents, i.e. prices of: 1) food, 2) commodities less food and energy commodities, 3) energy and 4) services less energy services (see Figure 1). The data set can be retrieved from the U.S. Bureau of Labor Statistics (BLS)[2]. The time series employed are monthly and seasonally adjusted (X-12 ARIMA), except for CPI services less energy services, which did not have a seasonal behaviour.

We present results for models estimated using monthly changes in year-on-year inflation. In fact, we found that modelling month-on-month (rather than year-on-year) inflation and/or inflation levels (rather than inflation changes), could lead to a slight reduction in the MSFE, but generally yielded worse forecasting performances when the whole predictive distribution is evaluated.

[2] http://www.bls.gov/cpi/data.htm

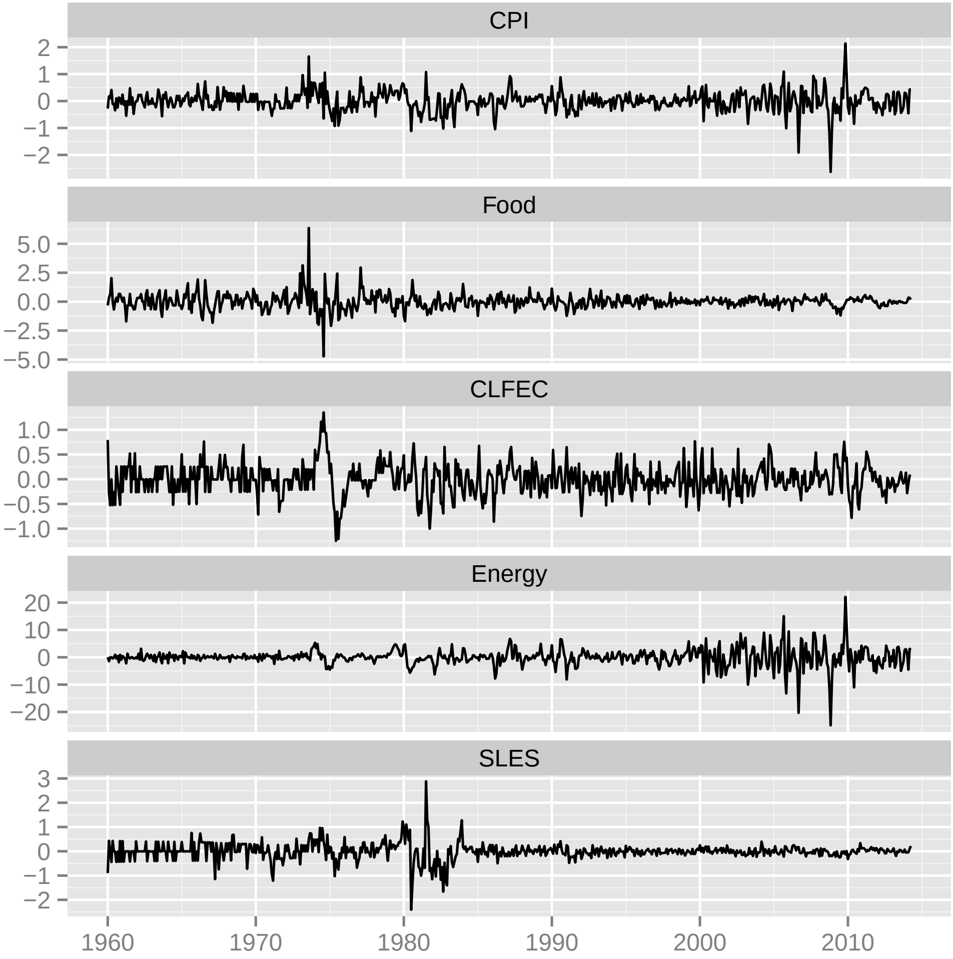

Figure 1: Monthly changes (percent) in year-on-year inflation for U.S. consumer price index (CPI) and its subcomponents: Food, Commodities Less Food and Energy Commodities (CLFEC), Energy, Services Less Energy Services (SLES).

We will compute out-of-sample forecasts for different periods, using the year 1984 for splitting the sample. In fact, in the literature often attention is paid to forecasting inflation in the post-1984 period, which corresponds to the great moderation. This is in line e.g. with the analyses by Stock and Watson (2007), Hendry and Hubrich (2011) and Groen, Paap, and Ravazzolo (2013).

More precisely, the out-of-sample evaluation is based on periods 1970(1)–1983(12), 1984(1)–2004(12) and 1984(1)–2014(4). The period 1984(1)–2004(12) is interesting since it is only affected by relatively small variations of the inflation volatility (making prediction intervals easier to compute), while volatility starts to increase thereafter.

	1970(1)–1983(12)			1984(1)–2004(12)			1984(1)–2014(4)		
	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}^c_t(h)$	0.449	0.725	1.072	0.280	0.484	0.615	0.390	0.684	0.878
${}_u\hat{x}^c_t(h)$	0.424	0.709	1.028	0.256	0.442	0.573	0.380	0.668	0.878
$\hat{x}^c_t(h)$	0.387	0.634	0.919	0.259	0.442	0.563	0.359	0.643	0.843

Table 4: RMSFE, U.S. year-on-year inflation (percentage points).

		1970(1)–1983(12)			1984(1)–2004(12)			1984(1)–2014(4)		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}^c_t(h)$	KS	0.058	0.126	0.179	0.292	0.152	0.335	0.327	0.460	0.408
	LM	0.163	0.027	0.133	0.420	0.063	2e-5	0.849	0.056	2e-5
	CC	5e-4	5e-4	0.007	0.095	0.585	0.227	5e-4	5e-4	0.126
${}_u\hat{x}^c_t(h)$	KS	0.048	0.241	0.146	0.099	0.300	0.268	0.288	0.701	0.620
	LM	0.009	0.003	0.054	0.259	0.341	1.5e-4	0.472	0.493	0.001
	CC	0.031	0.015	0.033	0.504	0.588	0.123	0.033	0.018	0.087
$\hat{x}^c_t(h)$	KS	0.065	0.057	0.073	0.974	0.582	0.128	0.917	0.772	0.332
	LM	0.018	0.087	0.302	0.700	0.004	0.001	0.732	0.035	0.007
	CC	0.009	0.002	0.016	0.492	0.440	0.245	0.085	0.096	0.221

Table 5: Results for U.S. year-on-year inflation: p-values for the Kolmogorov-Smirnov (KS), LM (no serial correlation of order 1) and Christoffersen (CC, conditional coverage, 95%) tests on the predictive distributions.

For estimation, a 10 year rolling sample is employed. We found this to be preferable to a recursively expanding sample, since the volatility of aggregate as well as components of inflation generally changes in time. Therefore using models estimated, for example, during a long period of stable inflation, will lead to poor density forecasts for periods of higher volatility (in particular, observed coverage rates will be smaller than expected). As we did for simulations, models are estimated by least squares, and the AIC$_c$ criterion (with maximum number of lags $p = 10$) is employed for order selection. Besides, prediction intervals and density forecasts are computed with $B = 2000$ bootstrap replicates, while bias correction is based on $B_0 = 1000$ replicates.

Concerning the aggregation matrix F_t in Equation (3), disaggregate forecasts are combined replicating the procedure adopted by the BLS. Since aggregation weights change over time, the current weights available at prediction time are employed, as future weights are unknown to the forecaster. We use this procedure, rather than the one described for DGP$_6$ in Section 5, in order to be able to compute ${}_u\hat{x}^c_t(h)$, and also for better comparability with the results in Hendry and Hubrich (2011).

Tables 4 and 5 describe the root MSFE (RMSFE) for the year-on-year inflation (percentage points) and the p-values[3] for the Kolmogorov-Smirnov (KS), LM (no serial correlation of order 1) and Christoffersen (CC, conditional coverage, 95%) tests on the predictive distributions.

The period 1970(1)–1983(12) was characterised by inflation with evolving volatility. Out-of-sample predictions during this period are therefore particularly difficult using AR models. According to the RMSFE criterion, $\hat{x}_t^c(h)$ performs best and should be adopted in this period. In disagreement with this suggestion, ${}_u\hat{x}_t^c(h)$ appears to be preferable if correct conditional coverage for the 95% prediction intervals is desired, since it passes the CC test (at the 0.01 level of significance), while the other two approaches perform worse in this respect. This happens despite the fact that ${}_u\hat{x}_t^c(h)$ does not appear to be able to fully capture the linear dependence structure in the data (see the p-values for the LM test).

In the period 1984(1)–2004(12) relatively small changes in inflation volatility have occurred, and the forecasting methods perform generally better. In agreement with the results in Hendry and Hubrich (2011), the RMSFEs are smaller than those for the period 1970(1)–1983(12), with ${}_u\hat{x}_t^c(h)$ and $\hat{x}_t^c(h)$ having similar performances and being preferable to ${}_d\hat{x}_t^c(h)$. The three procedures all yield reliable prediction intervals (see the p-values of the CC test), with ${}_u\hat{x}_t^c(h)$ and $\hat{x}_t^c(h)$ generating p-values much larger than that for ${}_d\hat{x}_t^c(h)$ when $h = 1$. The null hypothesis for the LM test is often rejected when $h = 2, 3$, suggesting the presence, at these forecast horizons, of linear dependence that is not accounted for by the models.

The inflation volatility for 1984(1)–2014(4) is initially stable, and then increases towards the end of the period. The performances of the three prediction methods are somewhere in the middle between those for the other two periods. Again, $\hat{x}_t^c(h)$ generates smaller RMSFEs. Here, $\hat{x}_t^c(h)$ is preferable also when evaluated with the CC test. In fact, $\hat{x}_t^c(h)$ always passes the CC test at the 0.05 level of significance, while the same does not hold for the other two prediction methods. This suggests that, for this period, $\hat{x}_t^c(h)$ is a safe choice from several points of view.

7 Conclusions

When considering estimation and specification uncertainty, the best way to predict aggregated time series is unclear and depends on the data at hand. A procedure that ranks the available prediction methods on the basis of an observed time series, and takes uncertainty into account, appears then to be useful. Since forecasters are often interested not in one, but in a range of

[3] While Table 3 shows a measure of performance (given by the percentage of 5% tests rejecting over Monte Carlo iterations), in Table 5 simple p-values are displayed.

possible outcomes (e.g. Rossi and Sekhposyan 2014, or Brandt, Freeman, and Schrodt 2014), it is advantageous to base the ranking of prediction methods on the performance of the whole predictive distribution, as opposed to the evaluation of a single point forecast. The bootstrap technique, and the ranking methods, detailed in the paper allow to fulfill these objectives.

Simulations, and an empirical application to the prediction of U.S. CPI inflation, show that only evaluating point forecasts, as often done in the current literature, can indeed suggest a prediction method that has, e.g., a worse performance in terms of prediction intervals.

It is also found that the ranking of prediction methods is highly dependent on the data generating process. However, when prediction intervals of a specified level are of interest, a univariate prediction of the aggregate seems to be the method of choice, at least within the simulation design considered here.

The empirical results suggest that autoregressive models, while being adequate (in all cases at least one of the models gave reasonable test results), might benefit from the flexibility given from allowing for breaks in the error variance and/or regression parameters (see e.g. Bos, Koopman, and Ooms 2014, who investigate changing time series characteristics of US inflation using long memory processes). This appears to be a promising direction for further research.

References

Anderson HM, Dungey M, Osborn DR, Vahid F (2011) Financial integration and the construction of historical financial data for the euro area. Economic Modelling 28:1498–1509

Athanasopoulos G, Vahid F (2008) VARMA versus VAR for macroeconomic forecasting. Journal of Business and Economic Statistics 26:237–252

Bermingham C, D'Agostino A (2011) Understanding and forecasting aggregate and disaggregate price dynamics. ECB Working Paper No. 1365

Beyer A, Doornik JA, Hendry DF (2001) Constructing historical euro-zone data. The Economic Journal 111:102–121

Bodo G, Golinelli R, Parigi G (2000) Forecasting industrial production in the euro area. Empirical Economics 25:541–561

Bos CS, Koopman SJ, Ooms M (2014) Long memory with stochastic variance model: A recursive analysis for US inflation. Computational Statistics and Data Analysis 76:144–157

Brandt PT, Freeman JR, Schrodt PA (2014) Evaluating forecasts of political conflict dynamics. International Journal of Forecasting 30:944–962

Breidt FJ, Davis RA, Dunsmuir W (1995) Improved bootstrap prediction intervals for autoregressions. Journal of Time Series Analysis 16:177–200

Bruneau C, De Bandt O, Flageollet A, Michaux E (2007) Forecasting inflation using economic indicators: The case of France. Journal of Forecasting 26:1–22

Carson RT, Cenesizoglu T, Parker R (2011) Forecasting (aggregate) demand for US commercial air travel. International Journal of Forecasting 27:923–941

Chevillon G (2014) Multi-step forecast error corrections: A comment on "Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set" by Barbara Rossi and Tatevik Sekhposyan. International Journal of Forecasting 30:683–687

Christoffersen PF (1998) Evaluating interval forecasts. International Economic Review 39:841–862

Christoffersen PF, Pelletier D (2004) Backtesting value-at-risk: A duration-based approach. Journal of Financial Econometrics 2:84–108

Clements MP, Taylor N (2001) Bootstrapping prediction intervals for autoregressive models. International Journal of Forecasting 17:247–267

Clements MP, Franses PH, Smith J, Van Dijk D (2003) On SETAR non-linearity and forecasting. Journal of Forecasting 22:359–375

Conover WJ (1999) Practical nonparametric statistics. John Wiley & Sons, New York

Dedola L, Gaiotti E, Silipo L (2001) Money demand in the euro area: Do national differences matter? Economic Working Paper 405, Bank of Italy, Economic Research Department

Diebold FX, Mariano RS (2002) Comparing predictive accuracy. Journal of Business and Economic Statistics 20:134–144

Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts with applications to financial risk management. International Economic Review 39:863–883

Dufour JM (2006) Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. Journal of Econometrics 133:443–477

Engle R (2002) Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics 20:339–350

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

Fagan G, Henry J (1998) Long run money demand in the EU: Evidence for area-wide aggregates. Empirical Economics 23:483–506

Faust J, Wright JH (2013) Forecasting inflation. In: Elliott G, Timmermann A (eds) Handbook of Economic Forecasting, Elsevier, Amsterdam, pp 2–56

Franses PH (1998) Time series models for business and economic forecasting. Cambridge University Press, Cambridge

Giacomini R, Granger CWJ (2004) Aggregation of space-time processes. Journal of Econometrics 118:7–26

Godfrey LG (1986) Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches. Cambridge University Press, New York

Granger CWJ (1987) Implications of aggregation with common factors. Econometric Theory 3:208–222

Grigoletto M (1998) Bootstrap prediction intervals for autoregressions: Some alternatives. International Journal of Forecasting 14:447–456

Grigoletto M (2005) Bootstrap prediction regions for multivariate autoregressive processes. Statistical Methods & Applications 14:179–207

Groen JJJ, Paap R, Ravazzolo F (2013) Real-time inflation forecasting in a changing world. Journal of Business & Economic Statistics 31:29–44

Grunfeld Y, Griliches Z (1960) Is aggregation necessarily bad? The Review of Economics and Statistics 42:1–13

Hendry DF, Hubrich K (2011) Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. Journal of Business and Economic Statistics 29:216–227

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

Hurvich CM, Tsai CL (1993) A corrected Akaike information criterion for vector autoregressive model selection. Journal of Time Series Analysis 14:271–279

Kabaila P (1993) On bootstrap predictive inference for autoregressive processes. Journal of Time Series Analysis 14:473–484

Kilian L (1998a) Accounting for lag order uncertainty in autoregressions: The endogenous lag order bootstrap algorithm. Journal of Time Series Analysis 19:531–548

Kilian L (1998b) Small-sample confidence intervals for impulse response functions. Review of Economics and Statistics 80:218–230

Kilian L (2001) Impulse response analysis in vector autoregressions with unknown lag order. Journal of Forecasting 20:161–179

Kim JH (1997) Relationship between the forward and backward representations of the stationary VAR model. Econometric Theory 13:889–890

Kim JH (1998) The relationship between forward and backward representations of the stationary VAR models. Econometric Theory 14:691–693

Kim JH (1999) Asymptotic and bootstrap prediction regions for vector autoregression. International Journal of Forecasting 15:393–403

Kim JH (2001) Bootstrap-after-bootstrap prediction intervals for autoregressive models. Journal of Business & Economic Statistics 19:117–128

Kim JH (2002) Bootstrap prediction intervals for autoregressive models of unknown or infinite lag order. Journal of Forecasting 21:265–280

Kim JH (2004) Bias-corrected bootstrap prediction regions for vector autoregression. Journal of Forecasting 23:141–154

Kiviet JF (1986) On the rigour of some misspecification tests for modelling dynamic relationships. The Review of Economic Studies 53:241–261

Konishi S, Kitagawa G (2008) Information criteria and statistical modeling. Springer, Berlin

Kupiec PH (1995) Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 3:73–84

Lewis R, Reinsel GC (1985) Prediction of multivariate time series by autoregressive model fitting. Journal of Multivariate Analysis 16:393–411

Lütkepohl H (1984) Forecasting contemporaneously aggregated vector ARMA processes. Journal of Business & Economic Statistics 2:201–214

Lütkepohl H (1985) The joint asymptotic distribution of multistep prediction errors of estimated vector autoregressions. Economics Letters 17:103–106

Lütkepohl H (1987) Forecasting aggregated vector ARMA processes. Springer, Berlin

Lütkepohl H (2004) Univariate time series analysis. In: Lütkepohl H, Krätzig M (eds) Applied time series econometrics, Cambridge University Press, New York, pp 8–85

Lütkepohl H (2005) New introduction to multiple time series analysis. Springer, Berlin

Lütkepohl H (2010) Forecasting aggregated time series variables: A survey. Journal of Business Cycle and Measurement Analysis 2010:37–62

Lütkepohl H (2011) Forecasting nonlinear aggregates and aggregates with time-varying weights. Jahrbücher für Nationalökonomie und Statistik 231:107–133

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

Marcellino M, Stock JH, Watson MW (2006) A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics 135:499–526

Masarotto G (1990) Bootstrap prediction intervals for autoregressions. International Journal of Forecasting 6:229–239

Moser G, Rumler F, Scharler J (2007) Forecasting Austrian inflation. Economic Modelling 24:470–480

Pascual L, Romo J, Ruiz E (2004) Bootstrap predictive inference for ARIMA processes. Journal of Time Series Analysis 25:449–465

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Rossi B, Sekhposyan T (2014) Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set. International Journal of Forecasting 30:662–682

Sbrana G (2012) Forecasting aggregated moving average processes with an application to the euro area real interest rate. Journal of Forecasting 31:85–98

Sbrana G, Silvestrini A (2009) What do we know about comparing aggregate and disaggregate forecasts? CORE Discussion Paper 2009/20, Université Catholique de Louvain

Shibata R (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. The Annals of Statistics 8:147–164

Siliverstovs B, Dijk DJC (2003) Forecasting industrial production with linear, nonlinear, and structural change models. Econometric Institute Report EI 2003-16, Erasmus University Rotterdam

Stock JH, Watson MW (2007) Why has US inflation become harder to forecast? Journal of Money, Credit and Banking 39:3–33

Theil H (1954) Linear aggregation of economic relations. North-Holland, Amsterdam

Thombs LA, Schucany WR (1990) Bootstrap prediction intervals for autoregression. Journal of the American Statistical Association 85:486–492

Tierney L, Rossini AJ, Li N, Ševčíková H (2011) snow: Simple network of workstations. URL `http://CRAN.R-project.org/package=snow`

Wang J, Tsang WW, Marsaglia G (2003) Evaluating Kolmogorov's distribution. Journal of Statistical Software 8

Winder CCA (1997) On the construction of European area-wide aggregates - a review of the issues and empirical evidence. IFC Bulletin 1:15–23

Zellner A, Palm F (1974) Time series analysis and simultaneous equation econometric models. Journal of Econometrics 2:17–54

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it

Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

Department of Statistical Sciences
University of Padua
Italy