Working Paper Series, N. 10, November 2012

Forecasting the distribution of aggregated time series: a bootstrap approach

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. Uncertainty due to parameter estimation 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 forecast distribution. The forecasting procedures are applied to the United States aggregate inflation.

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

Contents

1 Introduction **1**

2 The model and the competing forecasts **3**

3 The bootstrap procedure **5**

4 Comparing the predictive performances **7**

5 Simulation study **9**

6 Application **11**

7 Conclusions **14**

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

Forecasting the distribution of aggregated time series: a bootstrap approach

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. Uncertainty due to parameter estimation 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 forecast 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.

This problem has often been faced in the economic literature. For example, several aggregate and disaggregate predictors of the euro-zone inflation

are considered by Espasa *et al.* (2002) and Hubrich (2005). Similarly, Hendry & 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 *et al.*, 2007, Moser *et al.*, 2007). Besides inflation, Marcellino *et al.* (2003) also analyze real GDP, industrial production and unemployment for the euro area, while Fagan & Henry (1998) and Dedola *et al.* (2001) focus on money demand, expliciting contributions generated at a national level. Carson *et al.* (2011) analyze 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 *et al.* (2000) show that area wide models are preferable for forecasting industrial production. Many other examples exist in the literature.

Early theoretical results on aggregation versus disaggregation in forecasting can be found in Theil (1954) and Grunfeld & Griliches (1960). Other theoretical contributions include, among others, Lütkepohl (1984, 1987), Granger (1987), Giacomini & Granger (2004), Hendry & Hubrich (2011) and Sbrana (2012). See Lütkepohl (2010) for a recent survey on aggregation and forecasting.

According to the theoretical literature, except under certain conditions and when the data generating process (DGP) is known, 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.

This paper suggests the use of a bootstrap approach, in the context of vector autoregressive models (henceforth: VAR). These models are widely employed for forecasting 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 advantages of the bootstrap for univariate autoregressive (AR) models (e.g. Masarotto, 1990, Thombs & Schucany, 1990, Kabaila, 1993, Breidt *et al.*, 1995, Grigoletto, 1998, Clements & 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. Since, when using bootstrap, many replicates of the future values are generated, it is natural to examine the forecast performance using prediction intervals (Christoffersen, 1998, Kupiec,

1995), or the whole density forecast (Diebold *et al.*, 1998). This is important, since many realistic economic loss functions can't be reduced to the comparison of point forecasts, using the MSFE (Diebold & Mariano, 2002). 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 methods might be different.

The paper is organized as follows. In Section 2 the competing forecasts for the aggregate are defined, while Section 3 describes the bootstrap procedure. The criterions 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.

Model (1) can be written in backward form as

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

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 & 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}, $h = 1, \ldots, H$, are 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 $j \leq 0$.

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 $j \leq 0$.

We will focus on the following linear transformation of y_t:

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

where F is an $m \times k$ aggregation matrix of full rank m. It should be noted that, while a linearly transformed VARMA(p, q) process has a finite order VARMA representation, the same 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 & Vahid (2008). For these reasons, for our forecasting purposes we will represent x_t with equations analogous to (1) and (2). See also Lewis & 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.

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

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

2. Forecasting the disaggregated multivariate model and then aggregating the forecasts:

$$_d\hat{x}^c_t(h) = F \, \hat{y}^c_t(h) \ .$$

3. 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 $\hat{y}^c_{j,t}(h)$. The corresponding forecast of x_t will be

$$_u\hat{x}^c_t(h) = F \, (\hat{y}^c_{1,t}(h), \ldots, \hat{y}^c_{k,t}(h))' \ .$$

Notably, the univariate processes $y_{j,t}$, $j = 1, \ldots, k$, will belong to the ARMA class. 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 *et al.* (2006). 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 & 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 shows that, when the DGP is known and the MSFE is employed as evaluation criterion, ${}_d\hat{x}_t^c(h)$ is preferable to $\hat{x}_t^c(h)$ and ${}_u\hat{x}_t^c(h)$, while no unique ranking exists between $\hat{x}_t^c(h)$ and ${}_u\hat{x}_t^c(h)$. In practice, especially when the number of disaggregate variables is large, ${}_d\hat{x}_t^c(h)$ might well produce inferior results with respect to the other approaches. Also, ${}_d\hat{x}_t^c(h)$ 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 $\hat{x}_t^c(h)$, i.e. to forecast the aggregate series directly. The forecast ${}_u\hat{x}_t^c(h)$, 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. See Lütkepohl (2010) 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 ${}_d\hat{x}_t^c(h)$? 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.

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 & 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). Here, the AIC_c criterion proposed by Hurvich & Tsai (1993) will be adopted, since it performs similarly to AIC in large samples, while having superior bias properties in small samples.

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 $\hat{A}_0, \ldots, \hat{A}_{\hat{p}}$, $\{\hat{\varepsilon}_t\}$ and by $\hat{H}_0, \ldots, \hat{H}_{\hat{p}}$, $\{\hat{\nu}_t\}$, respectively. The residuals are rescaled as suggested in Thombs & Schucany (1990).

3. B_0 pseudo-datasets are generated from

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

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

   For each pseudo-dataset, the parameters of model (2) are estimated, obtaining $\hat{H}_0^*, \ldots, \hat{H}_{\hat{p}}^*$. Denoting by $\overline{\widehat{H}_j}^*$ the sample mean of the B_0 replications of $\hat{H}_j^*$, we have $\text{bias}(\hat{H}_j) = \overline{\widehat{H}_j}^* - \hat{H}_j$, for $j = 0, \ldots, \hat{p}$. This bias is used to compute the bias corrected estimates $\{\hat{H}_j^c\}$, with the procedure introduced by Kilian (1998b). A stationarity adjustment (step 1b of the procedure) is also performed 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 $\text{bias}(\hat{A}_j) = \overline{\widehat{A}_j}^* - \hat{A}_j$, 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 $\text{bias}(\hat{A}_j)$ 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

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

7. Perform steps 1–5 for the aggregated time series x_t. When $m = 1$, the underlying processes are univariate. The aggregate forecast replicates will be indicated by $\hat{x}_T^{*c}(h)$.

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

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

4 Comparing the predictive performances

In this section we examine the predictive performance of the three competing forecasts of the aggregated time series. When a prediction for the conditional mean is desired, straightforward evaluation criteria can be computed by simply comparing the realized 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.

The bootstrap procedure described above yields many forecast replicates. In this 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.

Since the 1990's a variety of tests have been proposed to measure accuracy of prediction intervals. Christoffersen (1998) introduced a model-free approach based on the concept of violation, which occurs when the *ex-post* realization of a variable does not lie in the *ex-ante* forecast interval. The validity of prediction intervals can then be reduced to the problem of assessing whether the following hypotheses are 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. Here, the Monte Carlo versions of these tests will be used, as suggested by Christoffersen & 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. These test procedures need to be modified when $h \geq 2$, since in this case optimal forecasts at horizon h are characterized by autocorrelation of order $h-1$. Diebold *et al.* (1998) recommend using an approach based on Bonferroni bounds: the indicator variables used in the test statistics are divided in h sub-groups, which are independent under the null hypothesis. The null hypothesis 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.

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 $\{z_T\}$. While Diebold *et al.* (1998) employed mainly graphical tools, more recently (e.g. Clements *et al.*, 2003, Siliverstovs & Dijk, 2003) formal tests have been applied, such as the Kolmogorov-Smirnov test (KS). Here, this test is implemented using the technique developed in Wang *et al.* (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

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	$\psi_{11} = \psi_{22} = 0.5$; $\alpha_{12} = \alpha_{21} = 0$
DGP_4	$\psi_{11} = 0.3$; $\psi_{22} = -0.5$ $\psi_{12} = -0.66$; $\psi_{21} = -0.5$

Table 1: Definition of the DGPs used in the simulation experiment, where $A_1 = \{\alpha_{ij}\}$ and $B_1 = \{\psi_{ij}\}$.

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, an LM test is carried out, with the null hypothesis that $\gamma_i = 0$, $i = 1, \ldots, q$. When $h > 1$, we proceed as described above, partitioning the PITs in h subgroups.

5 Simulation study

The simulation framework is designed to compare the performances of $\hat{x}_t^c(h)$, ${}_d\hat{x}_t^c(h)$ and ${}_u\hat{x}_t^c(h)$ in small samples and when the coefficients and the order of the DGPs are unknown. The experiments suggested by Lütkepohl (1984) 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 contribution, 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 an out-of-sample period can be used to choose between competing forecasting procedures.

We are going to consider two-dimensional DGPs. If we indicate by α_{ij} the generic element of A_1 in equation (1), the 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: $y_t = B_0 + \varepsilon_t + B_1 \varepsilon_{t-1}$. Table 1 defines the VMA(1) DGPs adopted, with $\{\psi_{ij}\}$ denoting the generic element of B_1. Also in this case, all the elements of the intercept vector B_0 are set equal to 1. Innovations are Gaussian i.i.d. with $\mathrm{E}(\varepsilon_t) = 0$ and $\Sigma_\varepsilon = I$.

General results (Lütkepohl, 1987) 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, in this context, there is no uncertainty originated e.g. from parameters and order estimation, 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, $i = 2, 4$, the MSFE for ${}_d x_t(h)$ will be lower than that for the other two predictors. 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).

The aggregation matrix in Equation (3) is $F = (1, \ldots, 1)$, so that x_t is a simple sum of the components in y_t. Results are obtained for samples with an initial size $T = 100$. Samples are recursively expanded to include an out-of-sample period having length $S = 100$.

Prediction intervals and density forecasts are computed with $B = 2000$ bootstrap replicates; bias correction is based on $B_0 = 1000$ replicates.

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

We will consider prediction horizons $h = 1, 2, 3$. The out-of sample period is used to compute the root MSFE (RMSFE), with respect to point forecasts given by the mean of the bootstrap forecasting distributions. The tests described in Section 4 will also be performed. RMSFEs and test p-values are then averaged over 200 Monte Carlo repetitions.

It should be mentioned that while the necessary computations are rather CPU-intensive, they are easy to parallelize. In fact, the results presented here have been obtained on a computer cluster.

Table 2 shows the RMSFEs for the different processes, when estimation uncertainty comes into play. When the component univariate processes are independent (DGP_1 and DGP_3), the three methods yield similar RMSFEs. On the contrary, when the component processes are dependent (DGP_2 and DGP_4), at the prediction horizon $h = 1$, ${}_d\hat{x}_t^c(h)$ produces smaller RMSFEs, while $\hat{x}_t^c(h)$ and ${}_u\hat{x}_t^c(h)$ perform similarly. As h increases, differences become less marked. These results are of course consistent with those found in previous literature, as e.g. Lütkepohl (1984) and Hendry & Hubrich (2011).

Let us now turn to the more general evaluation of the performance of the predictive distribution. Table 3 displays the 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.

The KS tests show similar performances for the three methods when the component processes are independent (DGP_1 and DGP_3), while ${}_u\hat{x}_t^c(h)$ is preferable (or at least comparable) when the component processes are interrelated (DGP_2 and DGP_4). It should be noted that in this case the LM test detects (at $h = 1$) the inability of ${}_u\hat{x}_t^c(h)$ to fully represent the linear dependence structure.

While the KS test concerns the whole forecasting distributions, when prediction intervals are of interest results for the CC test are particularly relevant.

	DGP$_1$			DGP$_2$		
	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	1.426	1.586	1.617	1.422	1.717	1.805
${}_u\hat{x}_t^c(h)$	1.420	1.580	1.613	1.708	1.738	1.869
$\hat{x}_t^c(h)$	1.418	1.579	1.611	1.687	1.750	1.886
	DGP$_3$			DGP$_4$		
	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	1.454	1.618	1.602	1.511	1.913	1.899
${}_u\hat{x}_t^c(h)$	1.445	1.607	1.593	1.710	1.860	1.855
$\hat{x}_t^c(h)$	1.442	1.606	1.593	1.693	1.879	1.867

Table 2: RMSFE in $S = 100$ out-of-sample steps (average over 200 Monte Carlo runs). The sample size is $T = 200$.

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$_4$, as for this process the CC test p-values are rather larger 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 $h = 2, 3$.

In summary, when computing confidence intervals, $\hat{x}_t^c(h)$ appears to be the most desirable choice, at least at this sample size and for the DGPs considered. It should be noted that this result partially contradicts the suggestions based on the RMSFE criterion.

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), analyzing euro area inflation, finds that neither approach works necessarily better, while results in Bermingham & D'Agostino (2011) are more in favor of aggregating disaggregate forecasts. See Faust & Wright (2012) for a recent discussion on inflation forecasting.

As in Hendry & 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. The data set can

		DGP_1			DGP_2		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	KS	0.541	0.381	0.276	0.533	0.398	0.244
	LM	0.526	0.321	0.264	0.511	0.307	0.262
	CC	0.454	0.298	0.249	0.435	0.311	0.241
${}_u\hat{x}_t^c(h)$	KS	0.522	0.363	0.272	0.670	0.417	0.324
	LM	0.501	0.313	0.255	0.004	0.233	0.150
	CC	0.480	0.317	0.262	0.383	0.286	0.236
$\hat{x}_t^c(h)$	KS	0.540	0.366	0.287	0.419	0.342	0.219
	LM	0.522	0.319	0.260	0.462	0.302	0.234
	CC	0.473	0.321	0.249	0.425	0.290	0.229
		DGP_3			DGP_4		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	KS	0.493	0.350	0.255	0.589	0.362	0.251
	LM	0.498	0.314	0.248	0.499	0.331	0.224
	CC	0.391	0.274	0.242	0.380	0.244	0.198
${}_u\hat{x}_t^c(h)$	KS	0.492	0.356	0.253	0.694	0.372	0.249
	LM	0.476	0.292	0.256	0.009	0.322	0.235
	CC	0.410	0.283	0.245	0.373	0.264	0.207
$\hat{x}_t^c(h)$	KS	0.493	0.348	0.254	0.571	0.389	0.270
	LM	0.485	0.299	0.242	0.514	0.315	0.230
	CC	0.425	0.292	0.246	0.445	0.306	0.248

Table 3: p-values for the Kolmogorov-Smirnov, LM (no serial correlation of order 1) and Christoffersen (conditional coverage, 95%) tests on the predictive distributions. Results are obtained for 100 out-of-sample steps and are averaged over 200 Monte Carlo runs. The sample size is $T = 200$.

be retrieved from the U.S. Bureau of Labor Statistics (BLS)[1]. 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 behavior.

We present results for models estimated using monthly changes in year-on-year inflation. In fact, we found that modeling 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.

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

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

	1970(1)–1983(12)			1984(1)–2004(12)			1984(1)–2012(8)		
	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	0.449	0.725	1.072	0.280	0.484	0.615	0.398	0.699	0.903
${}_u\hat{x}_t^c(h)$	0.424	0.709	1.028	0.256	0.442	0.573	0.384	0.680	0.896
$\hat{x}_t^c(h)$	0.387	0.634	0.919	0.259	0.442	0.563	0.363	0.654	0.860

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

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 & Watson (2007), Hendry & Hubrich (2011) and Groen *et al.* (2012).

More precisely, the out-of-sample evaluation is based on periods 1970(1)–1983(12), 1984(1)–2004(12) and 1984(1)–2012(8). 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.

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 show changing volatility 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.

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.

Tables 4 and 5 describe the root MSFE (RMSFE) for the year-on-year inflation (percentage points) and the 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.

The period 1970(1)–1983(12) was characterized 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 during this period. On the other hand, ${}_u\hat{x}_t^c(h)$ appears to be preferable if correct coverage for the prediction intervals is desired, since it passes the CC test, 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 volatil-

		1970(1)–1983(12)			1984(1)–2004(12)			1984(1)–2012(8)		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
${}_d\hat{x}_t^c(h)$	KS	0.058	0.126	0.179	0.292	0.152	0.335	0.196	0.447	0.556
	LM	0.163	0.027	0.133	0.420	0.063	2e-5	0.784	0.081	6e-5
	CC	5e-4	5e-4	0.007	0.095	0.585	0.227	5e-4	0.001	0.135
${}_u\hat{x}_t^c(h)$	KS	0.048	0.241	0.146	0.099	0.300	0.268	0.057	0.745	0.586
	LM	0.009	0.003	0.054	0.259	0.341	1.5e-4	0.680	0.643	0.002
	CC	0.031	0.015	0.033	0.504	0.588	0.123	0.034	0.002	0.063
$\hat{x}_t^c(h)$	KS	0.065	0.057	0.073	0.974	0.582	0.128	0.857	0.830	0.498
	LM	0.018	0.087	0.302	0.700	0.004	0.001	0.959	0.049	0.010
	CC	0.009	0.002	0.016	0.492	0.440	0.245	0.190	0.089	0.230

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

ity have occurred, and the forecasting methods perform generally better. In agreement with the results in Hendry & Hubrich (2011), the RMSFEs are smaller than those for the 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 rejection of the null hypothesis for the LM test when $h = 3$ suggests the presence, at this forecast horizon, of linear dependence that is not accounted for by the models.

The inflation volatility for 1984(1)–2012(8) 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 often preferable also when using the KS, LM and CC tests as ranking criterions.

7 Conclusions

Since the choice between combining disaggregate forecasts and forecasting the aggregate depends on the properties of the the process underlying observations, which in practice is unknown, the selection of the best forecast of the aggregate of interest is essentially an empirical problem.

The techniques described here show how estimation uncertainty can be considered when choosing an appropriate procedure to forecast the aggregate, and how the ranking between different approaches can be based on the evaluation of the performance of the whole predictive distribution, rather than on

a single point forecast.

Of course, the level of estimation uncertainty can greatly affect conclusions. Also, rankings based on the performance of point forecasts do not always agree with those derived from the more general evaluation of the forecasting distribution. This is confirmed by simulations and a real data example concerning the U.S. CPI inflation. The techniques analyzed appear then to be a useful instrument for forecasters desiring to discriminate among different prediction methods in an applied context.

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 e.g. from allowing for breaks in the error variance and/or regression parameters. This appears to be a promising direction for further research.

References

Athanasopoulos, G., & Vahid, F. 2008. VARMA versus VAR for macroeconomic forecasting. *Journal of Business and Economic Statistics*, **26**(2), 237–252.

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

Bodo, G., Golinelli, R., & Parigi, G. 2000. Forecasting industrial production in the euro area. *Empirical economics*, **25**(4), 541–561.

Breidt, F.J., Davis, R.A., & Dunsmuir, W. 1995. Improved bootstrap prediction intervals for autoregressions. *Journal of Time Series Analysis*, **16**(2), 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), 1–22.

Carson, R.T., Cenesizoglu, T., & Parker, R. 2011. Forecasting (aggregate) demand for US commercial air travel. *International Journal of Forecasting*, **27**(3), 923–941.

Christoffersen, P., & Pelletier, D. 2004. Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, **2**(1), 84–108.

Christoffersen, P.F. 1998. Evaluating interval forecasts. *International economic review*, 841–862.

Clements, M.P., & Taylor, N. 2001. Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting*, **17**(2), 247–267.

Clements, M.P., Franses, P.H., Smith, J., & Van Dijk, D. 2003. On SETAR non-linearity and forecasting. *Journal of Forecasting*, **22**(5), 359–375.

Dedola, L., Gaiotti, E., & Silipo, L. 2001. Money demand in the euro area: do national differences matter? *Economic Working Papers, Bank of Italy, Economic Research Department*.

Diebold, F.X., & Mariano, R.S. 2002. Comparing predictive accuracy. *Journal of business and economic statistics*, **20**(1), 134–144.

Diebold, F.X., Gunther, T.A., & Tay, A.S. 1998. Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, **39**(4), 863–883.

Dufour, J.M. 2006. Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, **133**(2), 443–477.

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**(4), 402–421.

Fagan, G., & Henry, J. 1998. Long run money demand in the EU: Evidence for area-wide aggregates. *Empirical Economics*, **23**(3), 483–506.

Faust, Jon, & Wright, Jonathan H. 2012. Inflation forecasting. *Handbook of Forecasting, forthcoming*.

Giacomini, R., & Granger, C.W.J. 2004. Aggregation of space-time processes. *Journal of econometrics*, **118**(1-2), 7–26.

Granger, C.W.J. 1987. Implications of aggregation with common factors. *Econometric Theory*, **3**(2).

Grigoletto, M. 1998. Bootstrap prediction intervals for autoregressions: some alternatives. *International Journal of Forecasting*, **14**(4), 447–456.

Grigoletto, M. 2005. Bootstrap prediction regions for multivariate autoregressive processes. *Statistical Methods & Applications*, **14**(2), 179–207.

Groen, J.J.J., Paap, R., & Ravazzolo, F. 2012. Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics, forthcoming*.

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

Hendry, D.F., & Hubrich, K. 2011. Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business and Economic Statistics*, **29**(2), 216–227.

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

Hurvich, C.M., & Tsai, C.L. 1993. A corrected Akaike information criterion for vector autoregressive model selection. *Journal of time series analysis*, **14**(3), 271–279.

Kabaila, P. 1993. On bootstrap predictive inference for autoregressive processes. *Journal of Time Series Analysis*, **14**(5), 473–484.

Kilian, L. 1998a. Accounting for lag order uncertainty in autoregressions: the endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis*, **19**(5), 531–548.

Kilian, L. 1998b. Small-sample confidence intervals for impulse response functions. *Review of economics and statistics*, **80**(2), 218–230.

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

Kim, J.H. 1999. Asymptotic and bootstrap prediction regions for vector autoregression. *International Journal of Forecasting*, **15**(4), 393–403.

Kim, J.H. 2002. Bootstrap prediction intervals for autoregressive models of unknown or infinite lag order. *Journal of Forecasting*, **21**(4), 265–280.

Kim, J.H. 2004. Bias-corrected bootstrap prediction regions for vector autoregression. *Journal of Forecasting*, **23**(2), 141–154.

Konishi, S., & Kitagawa, G. 2008. *Information criteria and statistical modeling*. Springer Verlag.

Kupiec, P.H. 1995. Techniques for Verifying the Accuracy of Risk Measurment Models. *Journal of Derivatives*, **3**(2), 73–84.

Lewis, R., & Reinsel, G.C. 1985. Prediction of multivariate time series by autoregressive model fitting. *Journal of multivariate analysis*, **16**(3), 393–411.

Lütkepohl, H. 1984. Forecasting contemporaneously aggregated vector ARMA processes. *Journal of Business & Economic Statistics*, 201–214.

Lütkepohl, H. 1985. The joint asymptotic distribution of multistep prediction errors of estimated vector autoregressions. *Economics Letters*, **17**(1-2), 103–106.

Lütkepohl, H. 1987. *Forecasting aggregated vector ARMA processes*. Vol. 284. Berlin: Springer.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Berlin: Springer.

Lütkepohl, H. 2010. Forecasting aggregated Time Series Variables: A Survey. *Journal of Business Cycle and Measurement Analysis*, **2010**(2), 37–62.

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

Masarotto, G. 1990. Bootstrap prediction intervals for autoregressions. *International Journal of Forecasting*, **6**(2), 229–239.

Moser, G., Rumler, F., & Scharler, J. 2007. Forecasting Austrian inflation. *Economic Modelling*, **24**(3), 470–480.

Pascual, L., Romo, J., & Ruiz, E. 2006. Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics & Data Analysis*, **50**(9), 2293–2312.

Sbrana, G. 2012. Forecasting Aggregated Moving Average Processes with an Application to the Euro Area Real Interest Rate. *Journal of Forecasting*, **31**(1), 85–98.

Siliverstovs, B., & Dijk, D.J.C. 2003. *Forecasting industrial production with linear, nonlinear, and structural change models*. Tech. rept. Econometric Institute Report EI 2003-16.

Stock, J.H., & Watson, M.W. 2007. Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, **39**(s1), 3–33.

Theil, H. 1954. *Linear aggregation of economic relations*. Amsterdam: North-Holland.

Thombs, L.A., & Schucany, W.R. 1990. Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, 486–492.

Wang, J., Tsang, W.W., & Marsaglia, G. 2003. Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, **8**(i18).

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