Real-time Forecasting of Inflation and Output Growth with Autoregressive Models in the Presence of Data Revisions

Michael P. Clements
Department of Economics
University of Warwick
M.P.Clements@warwick.ac.uk

Ana Beatriz Galvão*
School of Economics and Finance
Queen Mary University of London
a.ferreira@qmul.ac.uk

December 8, 2011

Abstract

We examine how the accuracy of real-time forecasts from models that include autoregressive terms can be improved by estimating the models on 'lightly-revised' data instead of using data from the latest-available vintage. The benefits to estimating autoregressive models on lightly-revised data are related to the nature of the data revision process and on the underlying process for the true values. Empirically, we find RMSFE improvements of 2-4% when forecasting output growth and inflation with univariate models, and of 8% with multivariate models. We show that multiple-vintage models, which explicitly model data revisions, require large estimation samples to deliver competitive forecasts.

Keywords: real-time data, news and noise revisions, optimal forecasts, multiple-vintage models, autoregressive models.

JEL code: C53.

*Corresponding author: Dr. Ana Beatriz Galvão. Queen Mary University of London, School of Economics and Finance, Mile End Road, E1 4NS, London, UK. Phone: ++44-20-78828825. email: a.ferreira@qmul.ac.uk.

We are grateful to Dick van Dijk and Simon van Norden for helpful comments, and to the participants at the 6th Forecasting Workshop at the ECB, Frankfurt, 2010, the 4th Time Series Conference, CIREQ, Montreal, 2010, and the 2010 Conference on Real-Time Data Analysis, Philadelphia. We also gratefully acknowledge comments from two anonymous referees of this journal.

1 Introduction

There has been much interest in the recent literature regarding the effects of different data vintages on model specification and forecast evaluation, and in the use of 'real-time data' in assessing predictability, as opposed to using 'final-revised' data, based on concerns that the use of final-revised data may exaggerate the predictive power of explanatory variables relative to what could actually have been achieved at the time using the then available data.[1] Following on from Koenig et al. (2003), we show that the accuracy of real-time forecasts from models that include autoregressive terms may be improved by estimating them on lightly-revised data, instead of using data from the latest-available vintage. We present a real-time analysis, in the sense that at each point in time the forecasting models are specified and the parameters estimated using only data for time periods up to that point in time, *and* the vintages of data used are restricted to those which would have been available at that time. Pseudo out-of-sample exercises adhere to the first aspect but use vintages of data that would not have been available at that time, whereas studies such as Clements and Galvão (2008, 2009) contain true real-time exercises.

When computing forecasts in real-time, the majority of the literature uses the 'traditional approach' to real-time forecasting. At each point in time, the values of all the observations from the latest-available vintage of data are used to estimate the forecasting model. This is known as the end-of-sample vintage approach, or EOS, following Koenig et al. (2003). To the extent that later estimates of a data point are more accurate than earlier estimates, this strategy uses the 'best' estimates of the data which are available at the time the forecast is made. However, it implies that a large part of the data used in model estimation has been revised many times, while the forecast is conditioned on data that has been just released or only revised a few times. We show that the traditional way of using real-time data for forecasting does not minimise the expected squared forecast error

[1] See, for example, Diebold and Rudebusch (1991b,a), Robertson and Tallman (1998), Orphanides (2001), Croushore and Stark (2001, 2003), Stark and Croushore (2002), Faust et al. (2003) and Orphanides and van Norden (2005). Croushore (2011a,b) provide additional background material on accounting for data revisions when forecasting, and provide a review of some of the topics covered in this paper.

in population, and, following Koenig et al. (2003), that the use of 'real-time vintage' (RTV) data can overcome the deficiencies of using EOS data.[2] Forecast accuracy is improved by reorganizing the real-time data employed in the estimation of the forecasting model in such a way that the data used in the estimation are of a similar maturity to the data on which the forecast is conditioned. Even if the target is to forecast post-revision data, the approach that uses RTV data to estimate the forecasting model reduces mean squared error in comparison with the use of EOS data.

We explicitly focus on autoregressive models. We derive analytical results that relate the properties of the forecasts generated by the use of EOS and RTV data to the properties of data revisions for an assumed data generation process that characterises data revisions as 'news' or 'noise', in the Mankiw and Shapiro (1986) sense. We use the Jacobs and van Norden (2011) statistical framework to model the 'regular' rounds of revisions which are made to the data at a level of detail that allows us to delineate between first and subsequent revisions, as the sizes of the variances of the first and subsequent revisions are found to affect the relative performance of RTV and EOS. We also allow for non-zero mean revisions as these are shown to affect the expected relative forecasting performance of RTV and EOS. It might be argued that data revisions are irregular and not amenable to modelling as a stationary process. This is likely to be true of benchmark revisions,[3] but the evidence presented by Croushore (2006) suggests that growth rates - the focus of our analysis - will be affected to a lesser degree than the levels of variables.[4] Our data generating process neglects certain characteristics of US data revisions - such as the seasonal nature of some revisions, as we explain below - but such complications would not affect the finding that RTV improves forecast accuracy relative to EOS in population.

For AR models we are able to relate the population properties of the estimators (using RTV or EOS data) back to the properties of the hypothesized data generation process,

[2] Harrison et al. (2005) consider the use of EOS data when there are measurement errors, and suggest that the most recent observations might be downweighted.

[3] Siklos (2008) identifies eight benchmark revisions in 1966, 1971, 1976, 1981, 1986, 1992, 1996 and 2001, all occurring in the data vintage of the first quarter of the year.

[4] Although in some circumstances taking growth rates may not be the best option: see, e.g., Knetsch and Reimers (2009).

including the properties of the revisions process, which yields additional insights when we can clearly categorize a series in terms of news or noise revisions. In models with explanatory variables it will be more difficult to obtain the direction of any bias of the estimator without making a range of further assumptions, including specifying the covariances between the revisions to the series being forecast and the revisions to the explanatory variables. Hence our analytical results are for the AR, as this allows for sharper predictions, although the general arguments that support the use of RTV data in the case of AR models are also applicable to ADL models.

We compare the RTV approach to forecasting with a model that uses the multiple estimates of the same observation that are typically available.[5] It is not clear that using multiple data vintages would improve forecast accuracy. In a recent review of forecasting with real-time data, Croushore (2006) concludes that the results of forecasting with state-space models that incorporate data revisions are mixed, compared to simply ignoring data revisions. In this paper, we add to this knowledge base by presenting forecasts from a vector autoregression (VAR) that models the relationships between the multiple-vintage estimates, in the spirit of recent work by Garratt et al. (2008, 2009) and Hecq and Jacobs (2009). We also consider the approach of Kishor and Koenig (2011), which specifies a model for the data revisions process which is estimated along side a vector autoregressive model for the 'post-revision' data.

A Monte Carlo exercise evaluates the relative efficacy of RTV compared to EOS under different assumptions about the model of data revisions. The simulations are based on models which are calibrated on actual US and UK data, so that we can compare the forecast performance of the different approaches for empirically-relevant models of data revisions. The Monte Carlo exercise is also designed to compare the forecast accuracy of multiple-vintage models and autoregressive models estimated with RTV data. We also check whether the gains from using RTV relative to EOS are robust to changes in the underlying statistical

[5]Examples of multiple-vintage models include Harvey et al. (1983), Howrey (1984), Patterson (1995, 2003), Jacobs and van Norden (2011), Cunningham et al. (2009), Garratt et al. (2009, 2008) and Hecq and Jacobs (2009).

framework.

Our empirical forecasting exercises compare the use of EOS and RTV data as competing approaches to real-time forecasting with simple autoregressive models of output growth and inflation, and consider the potential gains to using multiple-vintage models. We also allow for explanatory variables, and estimate a bivariate version of the Kishor and Koenig (2011) model.

In terms of related literature, our paper is closest to and is motivated by the seminal contribution of Koenig et al. (2003), who consider the use of lightly-revised data on the RHS of regression models. Our contribution is to consider explicitly the successive rounds of revisions made to national accounts data, and to allow for non-zero mean data revisions. We show how differences between estimates obtained with RTV data and EOS data depend on the nature of the process governing these rounds of revisions. Koenig et al. (2003) derive expressions for the bias of the estimator of distributed lag (DL) models for various ways of using real-time vintage data, but do not explicitly model the revisions process. They build their approach in terms of first and final estimates, and hence have a single revision value. We find that the properties of the estimators depend on factors which include the relative size of variances of the successive rounds of revision, and that knowledge of the revisions process can be used to predict the relative forecasting gains to be expected from using RTV compared to EOS data in any given instance.

A recent contribution by Nalewaik (2008) considers the impact of 'lack of signal error' (LoSE) in both the dependent and explanatory variables in regressions of output growth on asset prices. LoSE corresponds to what we term news. His focus is on the impact of LoSE on the estimates of the population relationship between the 'true' values of the variables. In terms of obtaining the most accurate forecasts in real-time as in our paper, the true relationship is not the object of interest, since there is a clear distinction between the optimal forecasting parameters and the parameters of the underlying true model, as spelt out in section 3.2.

The plan of the rest of the paper is as follows. Section 2 describes the real-time forecasting setting, and provides some intuition as to why simply using the latest-available vintage data to estimate the model may not be the best strategy. Section 3 derives the properties of RTV and EOS in a statistical framework that allows data revisions to be characterised as either news or noise. This is essentially the Jacobs and van Norden (2011) state-space model, but modified to allow for non-zero mean revisions, to reflect the nature of data revisions to macroeconomic aggregates. Section 4 presents a Monte Carlo evaluation of the improvements in forecasting accuracy from using RTV for data revision processes calibrated on US and UK data. Section 5 the empirical forecast comparison. Section 6 offers some concluding remarks. The derivations of the main results are confined to an appendix.

2 Motivation

When data are subject to revision, the forecaster at period $T+1$ has access to the vintage $T+1$ values of the observations on y up to time period T, as the first-release is published with a lag. We let y_t^{t+j} denote the vintage $t+j$ estimate of the value of the variable in period t, where $j = 1, 2, 3, \ldots$, and where $j = 1$ denotes the first-release value. Hence the forecaster has $(y_1^{T+1}, \ldots, y_{T-1}^{T+1}, y_T^{T+1})$. This is the 'latest-available vintage' data at period $T+1$, which we can write as $\left\{y_i^{T+1}\right\}_{i=1,2,\ldots,T}$. But the forecaster will also have the previous vintages, for example, the $T+1-j$ vintage, $\left\{y_i^{T+1-j}\right\}$ for $j = 1, 2, 3, \ldots$, and where $i = 1, 2, \ldots, T-j$.

The traditional approach estimates the forecasting model on the latest-available $(T+1)$ vintage, and conditions the forecasts on the $T+1$ vintage values of the forecast-origin data. So for an AR(2), the model is estimated on:

$$y_t^{T+1} = \alpha_0 + \alpha_1 y_{t-1}^{T+1} + \alpha_2 y_{t-2}^{T+1} + e_{t,EOS} \tag{1}$$

for $t = 3, \ldots, T$, and the forecast of y_{T+1} is given by:

$$\widehat{y_{T+1}}_{,EOS} = \hat{\alpha}_0 + \hat{\alpha}_1 y_T^{T+1} + \hat{\alpha}_2 y_{T-1}^{T+1}.$$

The parameter estimates $(\hat{\alpha}_0, \hat{\alpha}_1\, \hat{\alpha}_2)$ are based on data that are for the most part post-revision (or heavily-revised) data. Suppose for the sake of argument that data are revised 13 times, so that $y_t^{t+14} = \widetilde{y}_t$, where $\widetilde{y}_t$ is the true value, then a proportion $(T-13)/T$ of the T estimation observations underlying (1) are fully-revised or true data. Yet the forecasts are conditioned on a first-release observation and a once-revised observation (y_T^{T+1} and y_{T-1}^{T+1} respectively). This is the 'apples and oranges' problem of Kishor and Koenig (2011).

Given that the forecast will be conditioned on early-release data (y_T^{T+1} and y_{T-1}^{T+1} for an AR(2)), the RTV approach estimates the parameters of the AR(2) on matching early-release data:

$$y_t^{t+1} = \beta_0 + \beta_1 y_{t-1}^t + \beta_2 y_{t-2}^t + e_{t,RTV}$$

for $t = 3, \ldots, T$, and the forecast of y_{T+1} is given by:

$$\widehat{y_{T+1}},_{RTV} = \hat{\beta}_0 + \hat{\beta}_1 y_T^{T+1} + \hat{\beta}_2 y_{T-1}^{T+1}.$$

The forecast is conditioned on exactly the same data as under the traditional approach - the latest values from the most recent vintage. Notice that the RTV approach uses first-release data for the first lag, and second-release data for the second lag, and that this generalizes for an AR(p), so that it is not the case that only first-release data is used to estimate the model. The important point is that the data maturities used in estimation match those that the forecasts are conditioned on.

In what follows we show analytically that the RTV approach yields more accurate forecasts. To do that, we need a data generating process for the different data vintages and the true values.

3 Comparing RTV and EOS using a model of data revisions

We use a model of data revisions to compare the use of RTV and EOS data when estimating and forecasting with autogressive models. Based on the statistical framework described in section 3.1, we compute the optimal parameter vector of an AR forecasting model (section

3.2), and the population values of the parameter vector obtained by EOS (section 3.3) and RTV (section 3.4).

3.1 Statistical framework

Our statistical framework for modelling data revisions relates a data vintage estimate to the true value plus an error or errors, where the errors are typically unobserved. So the period $t+s$ vintage estimate of the value of y in period t, denoted y_t^{t+s}, where $s = 1, \ldots, l$, consists of the true value $\widetilde{y}_t$, as well as (in the general case) news and noise components, v_t^{t+s} and ε_t^{t+s}, so that $y_t^{t+s} = \widetilde{y}_t + v_t^{t+s} + \varepsilon_t^{t+s}$. Data revisions are news when initially released data are optimal forecasts of later data, so news revisions are not correlated with the earlier-release data, $Cov\left(v_t^{t+s}, y_t^{t+s}\right) = 0$. Data revisions are noise when each new release of the data is equal to the true value of y_t, denoted $\widetilde{y}_t$, plus noise, so that noise revisions are not correlated with the truth, $Cov\left(\varepsilon_t^{t+s}, \widetilde{y}_t\right) = 0$. We adopt the framework of Jacobs and van Norden (2011) which stacks the l different vintage estimates of y_t, namely, $y_t^{t+1}, \ldots, y_t^{t+l}$ in the vector $\mathbf{y}_t = \left(y_t^{t+1}, \ldots, y_t^{t+l}\right)'$, and similarly $\boldsymbol{\varepsilon}_t = \left(\varepsilon_t^{t+1}, \ldots, \varepsilon_t^{t+l}\right)'$ and $\mathbf{v}_t = \left(v_t^{t+1}, \ldots, v_t^{t+l}\right)'$, so that:

$$\mathbf{y}_t = \mathbf{i}\widetilde{y}_t + \mathbf{v}_t + \boldsymbol{\varepsilon}_t \tag{2}$$

where $\mathbf{i}$ is a l-vector of ones. One way of defining a revisions process with the required characteristics is to assume a process for $\widetilde{y}_t$, for example, an AR(p) with iid disturbances $R_1\eta_{1t}$, plus a sum of l news components $v_{i,t}$:

$$\widetilde{y}_t = \rho_0 + \sum_{i=1}^{p} \rho_i \widetilde{y}_{t-i} + R_1\eta_{1t} + \sum_{i=1}^{l} v_{i,t}, \tag{3}$$

where $v_{i,t} = \sigma_{v_i}\eta_{2t,i}$ (for $i = 1, ..., l$) and both η_{1t} and $\eta_{2t,i}$ are $iid(0,1)$. We let $\rho\,(L) = \sum_{i=1}^{p} \rho_i L^i$ and assume that the roots of $(1 - \rho\,(L)) = 0$ lie outside the unit circle, so that

$\widetilde{y}_t$ is a stationary process. The news and noise components of each vintage in $\mathbf{y}_t$ are:

$$\mathbf{v}_t = \begin{bmatrix} v_t^{t+1} \\ v_t^{t+2} \\ \vdots \\ v_t^{t+l} \end{bmatrix} = - \begin{bmatrix} \sum_{i=1}^{l} v_{i,t} \\ \sum_{i=2}^{l} v_{i,t} \\ \vdots \\ v_{l,t} \end{bmatrix}, \quad \boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_t^{t+1} \\ \varepsilon_t^{t+2} \\ \vdots \\ \varepsilon_t^{t+l} \end{bmatrix} = \begin{bmatrix} \sigma_{\varepsilon_1}\eta_{3t,1} \\ \sigma_{\varepsilon_2}\eta_{3t,2} \\ \vdots \\ \sigma_{\varepsilon_l}\eta_{3t,l} \end{bmatrix}, \tag{4}$$

where $\eta_{3t,i}$ is $iid(0,1)$. The shocks are also mutually independent, that is, if $\boldsymbol{\eta}_t = [\eta_{1t}, \boldsymbol{\eta}_{2t}', \boldsymbol{\eta}_{3t}']$, then $E(\boldsymbol{\eta}_t) = 0$, with $E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t') = I$.

Therefore, the first estimate of y_t, y_t^{t+1}, is $y_t^{t+1} = \rho_0 + \sum_{i=1}^{p} \rho_i \widetilde{y}_{t-i} + R_1\eta_{1t} + \sigma_{\varepsilon_1}\eta_{3t,1}$, which does not include any news component. Later estimates may be characterised by noise, but include more news components. For example, $y_t^{t+4} = \rho_0 + \sum_{i=1}^{p} \rho_i \widetilde{y}_{t-i} + R_1\eta_{1t} + \sigma_{\varepsilon_4}\eta_{3t,4} + \sum_{i=1}^{3} v_{i,t}$ is a more accurate estimate of $\widetilde{y}_t$ than y_t^{t+1}, because it includes the news terms $(\sum_{i=1}^{3} v_{i,t})$ which are part of $\widetilde{y}_t$ (in addition to the noise component). As noted by Mankiw and Shapiro (1986), news revisions imply that $var(y_t^{t+1}) < var(y_t^{t+l})$, while noise revisions imply that $var(y_t^{t+1}) > var(y_t^{t+l})$, assuming that later estimates are less 'noisy' $(\sigma_{\varepsilon_1} > \sigma_{\varepsilon_l})$. If $\sigma_{v_l} = 0$ and $\sigma_{\varepsilon_l} = 0$ the l-vintage value is the true value, $y_t^{t+l} = \widetilde{y}_t$. The assumption that $\widetilde{y}_t$ is a stationary process ensures that $\mathbf{y}_t$ is a stationary process from (2), as both the news and noise terms are stationary.

The assumptions we have made imply that both noise and news revisions are zero mean, so that the unconditional mean of the underlying series $\{\widetilde{y}_t\}$ and the observed data $\{\mathbf{y}_t\}$ are equal at $\rho_0 (1 - \rho(1))^{-1}$. However, there is evidence that the revisions to some macroeconomic data are non-zero mean (as we discuss in section 4.1).[6] To account for this characteristic of the revisions process, we consider a modified version of the statistical model that allows data revisions to affect the mean of the $\{\widetilde{y}_t, \mathbf{y}_t\}$. We assume that each news term is instead $v_{i,t} = \mu_{v_i} + \sigma_{v_i}\eta_{2t,i}$, and the noise component is $\varepsilon_t^{t+i} = -\mu_{\varepsilon_i} + \sigma_{\varepsilon_i}\eta_{3t,i}$. The true process is:

[6] See Aruoba (2008) and Corradi et al. (2009) for recent analyses of the properties of data revisions.

$$\widetilde{y}_t = \left[\rho_0 + \sum_{i=1}^{l} \mu_{v_i}\right] + \sum_{i=1}^{p} \rho_i \widetilde{y}_{t-i} + R_1\eta_{1t} + \sum_{i=1}^{l} \sigma_{v_i}\eta_{2t,i}, \tag{5}$$

since now $\sum_{i=1}^{l} v_{it} = \sum_{i=1}^{l} \mu_{v_i} + \sum_{i=1}^{l} \sigma_{v_i}\eta_{2t,i}$. The news and noise processes of each vintage are re-specified as:

$$\mathbf{v}_t = -\begin{bmatrix} \sum_{i=1}^{l} \mu_{v_i} \\ \sum_{i=2}^{l} \mu_{v_i} \\ \vdots \\ \mu_{v_l} \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{l} \sigma_{v_i}\eta_{2t,i} \\ \sum_{i=2}^{l} \sigma_{v_i}\eta_{2t,i} \\ \vdots \\ \sigma_{v_l}\eta_{2t,l} \end{bmatrix}, \ \boldsymbol{\varepsilon}_t = -\begin{bmatrix} \mu_{\varepsilon_1} \\ \mu_{\varepsilon_2} \\ \vdots \\ \mu_{\varepsilon_l} \end{bmatrix} + \begin{bmatrix} \sigma_{\varepsilon_{1,t}}\eta_{3t,1} \\ \sigma_{\varepsilon_{2,t}}\eta_{3t,2} \\ \vdots \\ \sigma_{\varepsilon_{l,t}}\eta_{3t,l} \end{bmatrix}. \tag{6}$$

such that either news or noise revisions may be non-zero mean, when either $\mu_{v_i} \neq 0$ or $\mu_{\varepsilon_i} \neq 0$, for any i.[7]

The statistical model can be cast in state-space form using (2) as the observation equation and combining (5) and (6) to obtain the measurement equation. The parameters can be estimated by maximum likelihood using the Kalman Filter, as described by Jacobs and van Norden (2011).

3.2 Forecasting with Autoregressive Models in Real Time

Consider how forecasters normally use real-time data to compute forecasts with an autoregressive model. At time $T+1$, the $T+1$ vintage of data contains data up to period T, so that the $T+1$ vintage is used to estimate the AR(p) model, and the forecasts are obtained by conditioning on the model estimates and the lagged values of y from the latest data vintage, namely, $\mathbf{y}_T^{T+1} = \left(y_T^{T+1}, \ldots, y_{T-p+1}^{T+1}\right)'$. But does least squares estimation of the AR model using the $T+1$ data vintage minimise the expected squared forecast error? To answer this question, we first obtain the population value of the parameter vector that minimizes the real-time squared forecast error for a forecast conditioned on $\mathbf{y}_T^{T+1}$, when

[7] A possible extension suggested by a referee would be to allow for time-variation in the revision means, caused, for example, by improvements in the procedures of the statistical agency. This could be accommodated within (6), but in this paper we treat the revision means as constant. We compare the impact of non-zero mean revisions calibrated to data released from different statistical agencies in section 4.

the data are subject to revisions as described in section 3.1. For a p^{th}-order autoregression, the forecast is given by the combination $\phi_0 + \boldsymbol{\phi}'\mathbf{y}_T^{T+1}$, where ϕ_0 is the intercept and $\boldsymbol{\phi}' = (\phi_1, \ldots, \phi_p)$ contains the slope parameters. The optimal parameter values $(\phi_0^*, \boldsymbol{\phi}^*)$, in terms of minimizing the real-time squared-error loss, are the solution to:

$$(\phi_0^*, \boldsymbol{\phi}^*) = \underset{\phi_0, \boldsymbol{\phi}}{\arg\min} E\left[\left(y_{T+1}^{T+1+f} - \phi_0 - \boldsymbol{\phi}'\mathbf{y}_T^{T+1}\right)^2\right] \tag{7}$$

The vintage of data employed to compute the forecast errors is that available at $T+1+f$, so that $f = 1$ indicates first-release data is used to evaluate the forecasts. The following proposition is derived in the Appendix.

Proposition 1 *The optimal parameters $(\phi_0^*, \boldsymbol{\phi}^*)$ when the data are generated by (2), (5) and (6), for the loss function given by (7) with $f = 1$, are:*

$$\begin{aligned} \boldsymbol{\phi}^* &= \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}} + \Sigma_{\mathbf{v}} + \Sigma_{\widetilde{\mathbf{y}}\mathbf{v}} + \Sigma'_{\widetilde{\mathbf{y}}\mathbf{v}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)^{-1} \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}} + \boldsymbol{\Sigma}'_{\widetilde{\mathbf{y}}\mathbf{v}}\right) \boldsymbol{\rho} \\ \phi_0^* &= \left(1 - \boldsymbol{\phi}^{*\prime}\mathbf{i}\right) \mu_{\widetilde{y}} - \textstyle\sum_{i=1}^{l} \mu_{v_i} - \mu_{\varepsilon_1} - \boldsymbol{\phi}^{*\prime}\boldsymbol{\mu}_{\varepsilon} - \boldsymbol{\phi}^{*\prime}\boldsymbol{\mu}_{v}, \end{aligned} \tag{8}$$

where the second moment matrices $\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}}$, $\Sigma_{\mathbf{v}}$, $\Sigma_{\widetilde{\mathbf{y}}\mathbf{v}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ are defined in the Appendix, $\mathbf{i}$ is a p-vector of 1's, and $\mu_{\widetilde{y}} = (1 - \rho(1))^{-1}\left[\rho_0 + \sum_{i=1}^{l} \mu_{v_i}\right]$, $E\left(\varepsilon_t^{t+1}\right) = -\mu_{\varepsilon_1}$, $E\left(v_t^{t+1}\right) = -\sum_{i=1}^{l} \mu_{v_i}$, and $\boldsymbol{\mu}_{\varepsilon} = E(\boldsymbol{\varepsilon}_t)$ and $\boldsymbol{\mu}_{v} = E(\mathbf{v}_t)$ are $p \times 1$ vectors of the means of noise and news components.

If revisions are zero-mean, the above simplify and the resulting expressions are optimal for all $f \geq 1$, i.e., irrespective of whether the goal is to forecast the first-release or the final values. The slope parameter simplifies to $\boldsymbol{\phi}_{news}^* = \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}} + \Sigma_{\mathbf{v}} + \Sigma_{\widetilde{\mathbf{y}}\mathbf{v}} + \Sigma'_{\widetilde{\mathbf{y}}\mathbf{v}}\right)^{-1} \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}} + \boldsymbol{\Sigma}'_{\widetilde{\mathbf{y}}\mathbf{v}}\right) \boldsymbol{\rho}$, if data revisions add news, and is $\phi_{noise}^* = \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)^{-1} \boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}}\boldsymbol{\rho}$ if data revisions reduce noise. Note that the optimal slope parameter is not equal to the parameter vector of the true process ($\boldsymbol{\rho}$) in either case.

3.3 Estimating AR forecasting models using EOS data

The standard approach is to estimate the model with the latest estimates of all the past observations available at the forecast origin. So, for example, at time $T+1$, the model is estimated with observations up to T from the $T+1$ vintage, while at $T+2$, observations up to $T+1$ from the $T+2$ vintage are used for estimation. For forecasting y_{T+1}^{T+1+f} the AR forecasting model with EOS data is given by:

$$y_t^{T+1} = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^{T+1} + e_{t,EOS} \tag{9}$$

where $t = p+1, \dots, T$, assuming the latest data vintage is dated $T+1$. In matrix notation:

$$Y^{T+1} = \mathbf{i}\alpha_0 + \mathbf{Y}_{-1}\boldsymbol{\alpha} + error$$

where $\mathbf{Y}_{-1} = \left[Y_{-1}^{T+1}, \dots, Y_{-p}^{T+1}\right]$, $\mathbf{i}$ is a $T-p$ vectors of 1's, and the vectors of observations Y^{T+1} and Y_{-i}^{T+1}, $i = 1, \dots, p$, are given by $Y^{T+1} = \left[y_{p+1}^{T+1}, \dots, y_{T-1}^{T+1}, y_T^{T+1}\right]'$, and $Y_{-i}^{T+1} = \left[y_{p+1-i}^{T+1}, \dots, y_{T-i-1}^{T+1}, y_{T-i}^{T+1}\right]'$. The main result is summarized in the following proposition, which is derived in the appendix.

Proposition 2 *The population (asymptotic) value of the least-squares estimators in the autoregressive model using EOS data, when the data are generated by (2), (5) and (6), is given by:*

$$\begin{aligned} \boldsymbol{\alpha}^* &= \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}} + \boldsymbol{\Sigma}_{\underline{\mathbf{v}}} + \boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}} + \boldsymbol{\Sigma}'_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}} + \boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}}\right)^{-1} \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}} + \boldsymbol{\Sigma}'_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}}\right) \boldsymbol{\rho} \\ \alpha_0^* &= \left(1 - \boldsymbol{\alpha}^{*\prime}\mathbf{i}\right)\left(\mu_{\widetilde{y}} - \mu_{v_l} - \mu_{\varepsilon_l}\right), \end{aligned} \tag{10}$$

where $\Sigma_{\underline{\mathbf{v}}}$ and $\Sigma_{\underline{\boldsymbol{\varepsilon}}}$ are second moment matrices of the news and noise components, and $\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}}$ is the second moment matrix between the news and the underlying process, $\widetilde{y}_t$, and $\mu_{\widetilde{y}} = (1-\rho(1))^{-1}\left[\rho_0 + \sum_{i=1}^{l} \mu_{v_i}\right]$, $E\left(\varepsilon_t^{t+l}\right) = -\ \mu_{\varepsilon_l}$ and $E\left(v_t^{t+l}\right) = -\mu_{v_l}$.

Remark 1 *Forecasts of the first-released value computed using the AR model with parame-*

ter vector $(\alpha_0^*, \boldsymbol{\alpha}^*)$*, as under EOS, are biased when data revisions are described by (2), (5) and (6), with bias of* $\left(\mu_{v_l} - \sum_{i=1}^{l} \mu_{v_i}\right) + \left(\mu_{\varepsilon_l} - \mu_{\varepsilon_1}\right) - \boldsymbol{\alpha}^{*\prime}\left[\left(\boldsymbol{\mu}_v + \mathbf{i}\mu_{v_l}\right) + \left(\boldsymbol{\mu}_\varepsilon + \mathbf{i}\mu_{\varepsilon_l}\right)\right]$.

A comparison of (10) and (8) shows that the conventional use of real-time data (EOS) for estimation of the AR(p) model does not deliver the optimal population parameters when there are data revisions. That is, $\boldsymbol{\alpha}^* \neq \boldsymbol{\phi}^*$ and $\alpha_0^* \neq \phi_0^*$, so that the forecasts of $\tau + 1$ computed using $\alpha_0^* + \boldsymbol{\alpha}^{*\prime}\mathbf{y}_\tau^{\tau+1}$ (for a set of forecast origins $\tau = T, T+1, T+2, \ldots$), where $\mathbf{y}_\tau^{\tau+1} = \left(y_\tau^{\tau+1}, \ldots, y_{\tau-p+1}^{\tau+1}\right)'$, are not optimal in a squared-error loss sense. Intuitively, when the sample is large, the use of EOS data amounts to mainly using fully-revised data (i.e., data from the y_t^{t+l} vintage) whilst optimal forecasts are obtained by relating the first estimates of the LHS variable to early estimates of the RHS variables. The finding of the lack of optimality of EOS forecasts holds for news and noise revisions. The differences between the optimal and EOS population values of the slopes and intercepts, namely, $(\boldsymbol{\alpha}^* - \boldsymbol{\phi}^*)$ and $(\alpha_0^* - \phi_0^*)$ will depend on whether revisions are news or noise.

When $p = 1$, and data revisions are noise, for example, then using the expressions for $\boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ in Appendix A, we have that:

$$\alpha^* - \phi^* = \left(\frac{\sigma_{\tilde{y}}^2}{\sigma_{\tilde{y}}^2 + \sigma_{\varepsilon_l}^2} - \frac{\sigma_{\tilde{y}}^2}{\sigma_{\tilde{y}}^2 + \sigma_{\varepsilon_1}^2}\right)\rho.$$

Thus, the difference of the EOS slope relative to the optimal value is increasing in ρ, the degree of persistence of the true process (here an AR(1)) and the difference in the variances of the first and last revisions, $\sigma_{\varepsilon_1}^2$ and $\sigma_{\varepsilon_l}^2$. ($\sigma_{\tilde{y}}^2$ is the variance of the underlying process).

Summarizing, the use of EOS data to estimate AR models for forecasting in real-time when data are subject to news and noise revisions delivers predictions that are not optimal, that is, the resulting forecasts do not minimise expected quadratic forecast loss. The forecasts are also biased (Remark 1) unless the revisions are zero mean.

3.4 Estimating AR forecasting models using RTV data

In this section we consider a way of using real-time data, motivated by the approach suggested by Koenig et al. (2003) in the context of distributed lag models, that delivers optimal estimators of the forecasting model in population. We apply their approach to the estimation of AR models by regressing the period $t+1$ vintage value of y_t on the t-vintage data values of the lags, y_{t-i}, $i=1,\ldots,p$, for $t=p+1,\ldots,T$. The reason for using the t-vintage value of the lags (rather than the $t+1$-vintage value) is that when the model is used for forecasting, say, the $T+2$ vintage value of y_{T+1}, the forecast will be conditioned on the previous $T+1$-vintage values of the explanatory variables. The autoregressive model with RTV data is:

$$y_t^{t+1} = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i}^t + e_{t,RTV} \tag{11}$$

where $t=p+1,\ldots,T$. In matrix notation, $Y^t = \mathbf{i}\beta_0 + \mathbf{Y}_{-1}^t\boldsymbol{\beta} + \mathbf{e}$, where Y^t and $\mathbf{Y}_{-1}^t = \left[Y_{-1}^t,\ldots,Y_{-p}^t\right]$ are given by $Y^t = \left[y_{p+1}^{p+2},\ldots,y_{T-1}^{T},y_T^{T+1}\right]'$, and $Y_{-i}^t = \left[y_{p+1-i}^{p+1},\ldots,y_{T-i-1}^{T-1},y_{T-i}^{T}\right]'$ for $i=1,\ldots,p$. Note that Y^t and Y_{-i}^t contain early vintages of data relative to the period $T+1$-vintage: these observations are not replaced with the latest available ($T+1$-vintage) values for these observations.

Consider estimating equation (11) by OLS. A typical observation on the LHS and RHS variables is $\left\{y_t^{t+1}, \mathbf{y}_{t-1}^t = \left(y_{t-1}^t \cdots y_{t-p}^t\right)'\right\}$. This is a covariance stationary process, so we can calculate the population values of the OLS estimators as the values that satisfy:

$$(\beta_0^*, \boldsymbol{\beta}^*) = \underset{\beta_0,\boldsymbol{\beta}}{\arg\min} E\left(y_t^{t+1} - \beta_0 - \boldsymbol{\beta}\mathbf{y}_{t-1}^t\right)^2.$$

This estimation loss function is identical to the real-time forecast loss function (7), when $f=1$, so that the solutions to the two coincide. Hence the use of RTV data to estimate the AR model delivers optimal forecasts, that is, $\beta_0^* = \boldsymbol{\phi}_\mathbf{0}^*$ and $\boldsymbol{\beta}^* = \boldsymbol{\phi}^*$. The unbiasedness of using RTV data to compute forecasts follows directly because $(\beta_0^*, \boldsymbol{\beta}^*)$ satisfy $E\left(y_t^{t+1} - \beta_0 - \boldsymbol{\beta}\mathbf{y}_{t-1}^t\right) = 0$ (which is one of the FOCs), and so by stationarity $E\left(y_{T+1}^{T+2} - \beta_0^* - \boldsymbol{\beta}^*\mathbf{y}_T^{T+1}\right) = 0$. The unbiasedness of the forecasts holds irrespective of whether or not the revisions are zero mean, provided that $f = 1$ when revisions are non-zero mean. Suppose the goal were instead to forecast y_{T+1}^{T+1+f}, when $f > 1$, then the estimation and out-of-sample loss criteria no longer match, and RTV estimation would yield a systematic forecast error if $E\left(y_{T+1}^{T+1+f} - y_{T+1}^{T+2}\right) \neq 0$.[8] However, a simple solution suggests itself - the forecast should be corrected by an estimate of the difference between the vintage we wish to forecast and the first release, e.g., the sample mean of $y_t^{t+f} - y_t^{t+1}$, $t = 1, \ldots, T - f$.

An alternative solution is to use an efficient estimate of $\tilde{y}_t$ as the LHS variable in the RTV regression, that is, an estimate y_t^{t+s} such that $E[y_t^{t+s}] = \tilde{y}_t$, where $s > 1$. This is a suggestion of Koenig et al. (2003), although they do not explicitly consider this in the context of non-zero mean revisions. This would also provide unbiased forecasts in population, but because the LHS variable is now y_t^{t+f}, instead of y_t^{t+1}, the model is estimated with $T - f + 1$ observations instead of T, so that there are small-sample considerations. The relative forecasting performance of these two approaches in finite samples is compared in section 4.

Summarizing, the use of RTV data to estimate and forecast with AR(p) models in real time delivers forecasts that minimise the real-time expected loss (optimal forecasts) and unbiased forecasts of y_{T+1}^{T+2}. Unbiasedness also holds for $f > 1$ if data revisions are zero mean, but in the event of non-zero mean revisions a simple correction can be applied to the intercept.

4 Monte Carlo Evaluation of the use of EOS and RTV data

We perform a number of simulation experiments to assess the potential size of the improvements in forecast accuracy from using RTV instead of EOS for AR models.

[8] The expected forecast error is $E\left(y_{T+1}^{T+f+1} - y_{T+1}^{T+2}\right) = E\left(v_{T+1}^{T+f+1} - v_{T+1}^{T+2} + \varepsilon_{T+1}^{T+f+1} - \varepsilon_{T+1}^{T+2}\right) = -\sum_{i=f}^{l} \mu_{v_i} + \sum_{i=1}^{l} \mu_{v_i} + \left(\mu_{\varepsilon_1} - \mu_{\varepsilon_f}\right) = -\sum_{i=1}^{f-1} \mu_{v_i} + \left(\mu_{\varepsilon_1} - \mu_{\varepsilon_f}\right)$, where $E\left(\varepsilon_t^{t+i}\right) = -\mu_{\varepsilon_i}$, and $E\left(v_t^{t+1}\right) = -\sum_{i=1}^{l} \mu_{v_i}$ and $E\left(v_t^{t+l}\right) = -\mu_{v_l}$, and so would be non-zero if either $\mu_{\varepsilon_f} \neq \mu_{\varepsilon_1}$ under noise revisions, or if $\sum_{i=1}^{f-1} \mu_{v_i} \neq 0$ when revisions are news.

4.1 Calibrating the key parameters

Figure 1 displays some of the key characteristics of the revisions processes for US real output growth and two measures of inflation (GDP deflator, PCE deflator) and UK real output growth.[9] We will use these as a guide when we calibrate our statistical model to ensure that we are calculating the potential size of the forecast losses from using EOS data for typical macroeconomic data subject to revisions. All four series are expressed in quarterly percentage differences. In the first panel, we plot the mean of each revision as a proportion of the mean of the first-released data. The figure plots the sample averages of revisions defined as $r_t^{(i)} = y_t^{t+1+i} - y_t^{t+i}$, for $i = 1, ..., 14$, calculated for the full period (1965Q3 onwards with US data and 1975Q4 onwards with UK data)[10]. It is apparent that the first revision tends to increase the mean of the first-released data by around 6% for US output growth and 3% for US GDP inflation. After the first revision, subsequent revisions tend to be smaller in terms of first-moment effects. The mean of the first revision to UK output growth (right axis) is twice as large as for the US, and there are spikes at later revisions.

In the top right panel, we present the standard deviation of each revision as a proportion of the standard deviation of the first-released data. For US output growth and GDP inflation, the first revision has a variance that is almost 25% of the first-release data. A salient feature is that the standard error of revisions tends to decrease in i. The standard deviation of revisions to UK output growth peaks at $i = 5$, and is of the same order of magnitude as the standard deviation of the first-release data (right axis). In contrast with US data, the variance of the revision to the fifth UK growth estimate (i.e., the $i = 5$ revision) is larger than the revision to the first estimate.

The bottom panel is the same as the top right, only restricted to the sub-period of the Great Moderation (McConnell and Perez-Quiros, 2000). It is evident that the relative importance of US revisions increases after 1985, as the lower variability of the measured

[9] All US real-time data employed in this paper are from the Real Time Data Set of the Philadelphia Fed available at the Philadelphia Fed webpage, http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/. See Croushore and Stark (2001). Real-time data on UK output growth are from the Bank of England dataset available at http://www.bankofengland.co.uk/statistics/gdpdatabase/ (long-run spreadsheet). We use vintages from both datasets up to 2009Q1.

[10] We plot $r_t^{(i)}$ for $i = 1$ to 14, as $r_t^{(14)}$ will account for all the 'regular' revisions.

first-release series is not matched one-for-one by a reduction in the variability of revisions. The first-revision standard deviation is now 40% of the standard deviation of the first-release data. In contrast the data revisions to UK output growth are relatively less important in the post-85 period. The relative sizes are now similar to US values, but the pattern over i indicates peaks at $i = 3$ and 6 (i.e., the revisions to the 3rd and 6th estimates, $y_t^{t+4} - y_t^{t+3}$ and $y_t^{t+7} - y_t^{t+6}$).

Based on these estimates of the typical sizes of revisions, we construct four sets of parameters for the statistical model in section 3.1, assuming throughout that $p = 2$ and $l = 14$. Firstly, consider the AR coefficients of the model for $\widetilde{y}_t$. The first two sets of parameters in Table 1 have autoregressive coefficients that sum up to 0.4, while those in the last two sum to 0.8. The first block is more typical of a process such as output growth which exhibits moderate persistence, while the second is typical of a more persistent process such as inflation. Comparisons between these two blocks will therefore be informative about whether the relative forecast accuracy of EOS and RTV estimation depends upon the persistence of the underlying process.

For all set of parameters, we assume that the means of the first and the fifth revisions are non-zero. The mean values of DGPs 1 and 3 were chosen in conjunction with the values of the AR coefficients to give ratios of revision means to first-released data of 4% and 2%, as suggested by US data in Figure 1. The mean values of DGPs 2 and 4 are twice as large as suggested by UK data values. Calibrated to US data, DGPs 1 and 3 have a relatively large first revision variance ($\sigma_{r_1}/\sigma_{y_t^{t+1}} = .4$), followed by equal-sized revisions of smaller variance ($\sigma_{r_i}/\sigma_{y_t^{t+1}} = .2$, for $i = 2, ...13$), with a small final revision ($\sigma_{r_{14}}/\sigma_{y_t^{t+1}} = .1$). [11] DGPs 2 and 4 have proportionately larger revision standard deviations after the first release, motivated by considering the UK data revisions after 1985.

The parameter values set out in Table 1 will be applied under the assumption that revisions are news, and under the assumption of noise. This will allow us to determine whether the news versus noise issue is relevant to the relative forecast accuracy of EOS and

[11] Specifically, for the purpose of computing the DGP parameters, we use $\sigma^2_{y_t^{t+1}} = \frac{R_1^2(1-\rho_2)}{(1+\rho_2)[(1-\rho_2)^2-\rho_1^2]}$.

RTV estimation of AR models for forecasting. We have calibrated our design parameters so that our simulations should be informative about empirical outcomes.

4.2 Monte Carlo findings

Table 2 shows the relative forecasting performance of RTV and EOS for small-estimation samples, and also 'large-sample' results based on numerical evaluation of the population value of the squared-error loss function when the parameter vectors of the AR models estimated by RTV and EOS take on their population values. We allow $f = 1$ and $f = 14$. When data revisions are non-zero mean, forecasts are computed by either bias-correcting the forecasts of the initial release, as proposed in section 3.4, or by using post-revision values as the LHS variable (Koenig et al., 2003). These two approaches are asymptotically equivalent, but not so in small samples. The DGP parameters are described in Table 1. The RMSFE ratios are computed with samples of size $T = 50, 100, 200$ and 500, and employ $10,000$ replications. The middle (vertical) panel records the bias-corrected forecasts, and the last panel the results for post-revision values as the LHS variable (effective sample size of $T - 13$).

Our analytical formulae generally provide a good match to the simulation results when the sample is large ($T = 500$). The large-sample results show the impact of data revisions is larger for more persistent data (compare DGPs 3 and 4 versus 1 and 2). The loss in terms of MSFE from EOS estimation in comparison with optimal forecasting is in the 2-15% range for the more persistent data (DGPs 3 and 4) but it is smaller for the less persistent data (.5 up to 7%). The MSFE reduction from using RTV instead of EOS data when forecasting first releases are generally larger than when forecasting post-revision data ($f = 14$). The accuracy improvements from RTV are larger when data revisions are news in contrast with noise revisions using the same set of parameters.

An important finding for practical forecasting is that the losses to using EOS instead of RTV data are markedly larger for small samples when revisions are news, so that the analytical results for forecasting first-released data downplay the likely empirical relevance of RTV estimation.

Even though gains from using RTV data are generally smaller when forecasting post-revision data, the use of bias-corrected forecasts is generally preferred to the use of post-revision data on the LHS: the latter approach fares relatively poorly when the sample is short, when the loss of observations has more of an impact.

In summary, we would expect that the forecast loss of using EOS data instead of RTV data would be higher when i) the estimation sample is relatively small, ii) the process is reasonably persistent, iii) revisions primarily add news, and iv) the forecasting target is first-release data.

4.3 Using a VAR of multiple vintages and the Kishor and Koenig (2011) model

The model of data revisions described in section 3 does not allow that data revisions are serially correlated. In this Monte Carlo exercise, we consider two additional approaches to modelling data revisions that relax this assumption. These alternative approaches are also candidate forecasting models, and estimated models of the DGPs are included in the forecast comparisons along with the autoregressive models estimated with EOS and RTV data.

The first alternative is a version of the vintage-balanced VAR (VB-VAR) of Hecq and Jacobs (2009). The VB-VAR is closely related to the VAR models used by Garratt et al. (2008, 2009) to model real-time data. The VB-VAR is given by:

$$\mathbf{z}^{t+1} = \mathbf{c}_0 + \boldsymbol{\Gamma}\mathbf{z}^t + \boldsymbol{\varepsilon}^{t+1} \tag{12}$$

where $\mathbf{z}^{t+1-i} = \left[y_{t-i}^{t+1-i}, y_{t-1-i}^{t+1-i}, \ldots, y_{t-q+1-i}^{t+1-i}\right]'$, $i = 0, 1$, $\mathbf{c}_0$ is $q \times 1$, $\boldsymbol{\Gamma}$ is $q \times q$, and $\boldsymbol{\varepsilon}_t$ is $q \times 1$. The vector $\mathbf{z}^{t+1}$ (with $q = l$) differs from the vector $\mathbf{y}_t$ from the statistical framework described in section 3.1, as it consists of the last q observations of vintage $t+1$, instead of different estimates of an observation at time t. This way of employing the panel of data vintages means that the forecasts will be conditioned on the latest-available estimates of the current and lagged observations. For example, for $p = 1$, the forecast of y_{T+1}^{T+2} will be the first element of the 1-step ahead vector of forecasts, $\hat{\mathbf{z}}^{T+2|T+1} = \widehat{\mathbf{c}}_0 + \widehat{\mathbf{\Gamma}}_1 \mathbf{z}^{T+1}$, while the forecast of y_{T+1}^{T+15} is the last element of the 14-step ahead vector of forecast values, $\hat{\mathbf{z}}^{T+15|T+1}$ (forecasts are computed by iteration). The VAR model captures the dynamics between different vintage estimates of the observations, but does not clearly disentangle news and noise revisions.

The second alternative is the Kishor and Koenig (2011) approach (henceforth KK) that builds on Howrey (1978, 1984) and Sargent (1989), and aims to forecast post-revision data. The Kishor and Koenig (2011) model includes a vector autoregressive process for the 'true' data, as well as a model of data revisions. The Kalman filter is used to compute 'true' values of recent observations by employing the dynamic structure of the model of data revisions. Then the forecasting model with 'true' data is employed for forecasting post-revision data.

As before, $\mathbf{z}^{t+1} = [y_t^{t+1}, ..., y_{t-q+1}^{t+1}]'$ contains the vintage $t+1$ data, while $\mathbf{z}_t = [y_t, ..., y_{t-q+1}]'$ is a vector of the same dimension containing the 'true values'. Including q estimates in the vector $\mathbf{z}^{t+1}$ implicitly assumes that the q^{th} estimate, y_{t-q+1}^{t+1}, is an efficient estimate of the true value, y_{t-q+1}. The model is written in state-space form with measurement and state equations:[12]

$$\mathbf{z}^{t+1} = \begin{bmatrix} I_q & I_q \end{bmatrix} \begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}^{t+1} - \mathbf{z}_t \end{bmatrix}, \tag{13}$$

and:

$$\begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}^{t+1} - \mathbf{z}_t \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{F} & 0_{q\times q} \\ 0_{q\times q} & \mathbf{K} \end{bmatrix} \begin{bmatrix} \mathbf{z}_{t-1} \\ \mathbf{z}^{t} - \mathbf{z}_{t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{v}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix}. \tag{14}$$

The disturbances vectors are $\mathbf{v}_t = (v_{1t}, 0, ..., 0)'$ and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{q-1t}, 0)'$. The model allows that the errors in the data revision equations $\boldsymbol{\varepsilon}_t$ are contemporaneously correlated, and may also be correlated with disturbances to the true values v_{1t}. If we define $\boldsymbol{\upsilon}_t = (\mathbf{v}_t' \ \boldsymbol{\varepsilon}_t')'$, then $E(\boldsymbol{\upsilon}_t \boldsymbol{\upsilon}_t') = \mathbf{Q}$ and $\mathbf{Q}$ is non-diagonal. The $q \times 1$ vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ are $\mathbf{c}_1 = (c_1, 0, ..., 0)'$ and $\mathbf{c}_2 = (c_{21}, ..., c_{2q-1}, 0)'$.

[12] These are equivalent, respectively, to equations (36) and (35) in Kishor and Koenig (2011).

The first block of (14) defines the process for the true values as being autoregressive:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f} & 0_{1\times(q-p)} \\ I_{q-1} & 0_{(q-1)\times 1} \end{bmatrix}$$

where $\mathbf{f} = (f_1, \ldots, f_p)$. The matrix $\mathbf{K}$ describes the dynamics of $q-1$ data revisions $\mathbf{z}^{t+1} - \mathbf{z}_t$:

$$\mathbf{K} = \begin{bmatrix} k_{1,1} & \ldots & k_{1,q-1} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ k_{q-1,1} & \ldots & k_{q-1,q-1} & 0 \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

Kishor and Koenig (2011) suggest estimating the coefficients $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{F}$, $\mathbf{K}$ and $\mathbf{Q}$ by SURE, assuming that $y_{t-q+1} = y_{t-q+1}^{t+1}$. The Kalman Filter is then applied to obtain filtered estimates of $\mathbf{y}_t$ using data vintages up to $T+1$. Forecasts of post-revision future observations $(y_{T+1}, ..., y_{T+h})$ are obtained by iteration of the estimated state equation. Forecasts of first release values y_{T+1}^{T+2},..., y_{T+h}^{T+h+1} are computed using forecasts of data revisions, $(y_{T+h}^{T+h+1} - y_{T+h})$, and the forecasts of the post-revision values.

The statistical model described in section 3.1 is a multiple-vintage model, and maximum likelihood estimation coupled with the Kalman filter can be employed to obtain estimates of the parameters and unobserved components as in Jacobs and van Norden (2011). However, we do not consider this model for real-time forecasting. The first reason is that we need $l-1$ revisions of each observation in order to estimate the model. Thus, forecasting at the time when we have the $T+1$ data vintage, say, the estimation of the model will not draw on the observations from $T-l+1$ up to T. This may be a serious shortcoming in short samples. Secondly, the successive re-estimation of the large number of parameters associated with the unobserved components may engender parameter instabilities and lead to convergence issues. In contrast, both the VB-VAR and KK models are easy to estimate (by OLS or SURE), and use observations up to T either directly in the model estimation (VB-VAR) or within the application of the Kalman filter (KK).

Table 3 reports the relative performance of AR models estimated with RTV and EOS data when data are generated from estimated VB-VAR and KK models. The models were estimated on US output growth, GDP deflator inflation, PCE deflator inflation, and UK output growth. In each case, the VB-VAR and the KK models were specified with $q = 14$. The KK data generating processes were estimated with $p = 1$ for output growth and $p = 4$ for all other variables, and with a diagonal $\mathbf{K}$ matrix. The autoregressive order of the AR models are set to $p = 1$ for DGPs based on output growth and $p = 4$ for all other variables. We consider forecasts of both first release and post-revision data. When forecasting post-revision data we use both approaches described in the previous section: RTV implies that first release forecasts are bias corrected based on past data, and RTV_KDP means that parameters are estimated using post-revision data as the LHS variable. Both VB-VAR and KK forecasting models are included in the comparisons, where the models are estimated on each replication with the same specification as the DGP.

The results in Table 3 suggest that the model of data revisions corresponding to the DGP generally delivers the most accurate forecasts when the sample is large ($T = 200, 500$), but for shorter samples loses out to the simple AR models. Both the KK and the VB-VAR model require the estimation of a large number of parameters in contrast with the autoregressive model, which adversely affects their forecast performance unless the estimation sample is large. For the shorter estimation samples ($T = 50, 100$), RTV has the edge over EOS for forecasting first-release data generated by the VB-VAR, and also for the KK-simulated data for the two output growth calibrations. For the longer estimation samples, RTV is generally at least as good as EOS, although differences between the two tend to be small. For the shorter samples, RTV_KDP is less accurate than RTV.

In summary, we expect larger gains from using RTV instead of EOS data when forecasting first-release values with a relatively short estimation sample. Approaches that attempt to model data revisions improve accuracy only if the sample size is large enough.

5 Forecasting US output growth and inflation

Our first empirical forecasting exercise compares the use of RTV and EOS data for simple autoregressive models for predicting US quarterly output growth and inflation. In addition, we include the VB-VAR and the KK as the multiple-vintage models to see whether the use of models of data revisions improves forecast accuracy.

We then present a forecasting exercise for output growth and inflation that allows for explanatory variables in addition to the autoregressive terms. This parallels the pseudo out-of-sample forecasting exercises of Stock and Watson (2003, 2008) for these two variables, albeit using only a handful of candidate explanatory variables. Nevertheless, we are able to assess whether the theoretical advantages to the use of RTV over EOS in real-time forecasting are realized in practice. We will also consider very short horizon forecasts (or 'nowcasts') as Koenig et al. (2003) found marked gains to RTV-estimation of models for quarterly output growth using monthly indicators at such horizons. We also include a bivariate version of the KK model (Kishor and Koenig, 2011) in the comparisons.

5.1 Univariate models of output growth and inflation

In this subsection we assess the empirical relevance of the analytical and Monte Carlo results for forecasting with autoregressive models of US output growth and the two measures of inflation. The autoregressive model estimated with EOS data is the benchmark model, against which we compare three alternative ways of handling data revisions:

(i) The use of RTV data (combined with bias-correction when predicting post-revision values).

(ii) The VB-VAR model, which is a vector autoregression of q estimates published at vintage $t + 1$.

(iii) The KK model, which aims to forecast post-revision values, while modelling the dynamics of data revisions.

Section 4.3 outlines the mechanics of forecasting with (ii) and (iii). Because the Monte Carlo results suggest that relatively large estimation samples (of size 200) are required for (ii) and (iii) to provide competitive forecasts relative to the simpler AR models, and because we need sufficient out-of-sample observations to have power to detect differences in forecast accuracy, we use only the US data. We have 174 vintages of US output growth and inflation data (1965Q4-2009Q1), compared to 134 vintages (1976Q1-2009Q1) of UK real GDP growth. We use the US vintages from 1985Q4-2009Q1 for the out-of-sample forecasting exercise. This means that the number of observations available for estimation range from 80 up to 170. Based on the results in section 4.3, the AR with RTV data might be expected to be more accurate than the KK and VB-VAR model forecasts, with larger gains from forecasting first-release values.

The Monte Carlo results also indicate larger gains from RTV than EOS when revisions are news, and the more persistent the underlying autoregressive processes. In view of this expected dependence of the relative performance of the two approaches on the characteristics of the data revision process, we present some relevant summary statistics in Table 4. Descriptive statistics are computed for two sub-samples, where the second covers the out-of-sample period. The variables are defined as (one hundred times) the quarterly difference of the log of the level. We consider first-released data y_t^{t+1}, data available three and a half years later y_t^{t+14}, as well as latest-available, which in our case is from the 2009Q1 vintage dataset, denoted y_t^{09Q1}. Table 4 presents means, standard deviations and first-order autocorrelations for the three data series, as well as p-values of tests for whether revisions ($y_t^{t+14} - y_t^{t+1}$ and $y_t^{09:1} - y_t^{t+1}$) are noise, or add news, and whether they are zero-mean. Recall that revisions are defined as noise if the initial estimate is an observation on the final series but measured with error, so that the revisions are uncorrelated with final value, but are correlated with data available when the initial estimate was made. Hence noisy revisions are predictable. Alternatively, revisions are news if the initial estimate is an efficient forecast of the final value, such that the revision is unpredictable from information available at the time the initial estimate was made. We test for news and noise revisions using, respectively, the following auxiliary regressions:

$$\begin{aligned} y_t^{t+l} - y_t^{t+1} &= \alpha + \beta y_t^{t+1} + \omega_t \\ y_t^{t+l} - y_t^{t+1} &= \alpha + \beta y_t^{t+l} + \omega_t \end{aligned}$$

where the null hypothesis is that $\alpha = \beta = 0$ in both cases. In place of y_t^{t+l}, we use both y_t^{t+14} and y_t^{09Q1}. We also test separately whether revisions are zero mean (H_0: $\alpha = 0$ in $y_t^{t+l} - y_t^{t+1} = \alpha + \omega_t$).

Data revisions to output growth and inflation are seen to have different characteristics, and show some variation across the forecast and estimation periods. For output growth we can reject the noise hypothesis for both periods using the 2009Q1 data vintage, and there is no evidence against the news hypothesis using the $t + 14$ data vintage. For both inflation measures the revisions relative to the 2009Q1 vintage can be assumed to be noise over the forecast period, and news over the estimation period, while the revisions relative to y_t^{t+14} cannot be so easily categorised. There is also evidence that the 2009Q1 revisions to output growth have been significantly upward, as have the $t + 14$ revisions to both inflation rates over the forecast period. At least for the period after 1985 (out-of-sample period), data revisions to output growth appear to be mainly news, and those to inflation mainly noise, while the first-order autocorrelation of inflation is twice that of output growth. When this information is combined with the Monte Carlo results, one might expect small gains from using RTV instead of EOS for both series: for inflation, because of the persistence of the series, and for output growth, because revisions are predominantly news.

We compare forecasts computed with RTV and EOS data for the three series, using autoregressive models with fixed specifications - the lag length for output growth is 1, and for the inflation series, 4. These autoregressive orders were chosen based on the performance of these models when forecasting with final data. We evaluate forecasts of first-released data (y_{T+1}^{T+2}), data after three and a half years of revision (y_{T+1}^{T+15}) and data from the last available vintage (y_{T+1}^{09Q1}). We use past data ($y_{t-13}^{t+1} - y_{t-13}^{t-12}$ for $t = .., T$) to compute the bias correction to predict post-revision values (y_{T+1}^{T+15} and y_{T+1}^{09Q1}) at each forecast origin ($T+1,...,T+N$). Table 5 presents ratios of RMSFEs using RTV and EOS data, such that values smaller than one favour RTV. We give results for 1-step ahead forecasts, and 4-step ahead forecasts (computed by iteration). There are gains for forecasting first release values from using RTV data for all three series, some of which are around 4%.

Can we do better still with the models that draw on multiple data vintages or model the data revisions process - namely, the VB-VAR and KK models? Table 5 reports the ratios of the RMSFEs of the VB-VAR model forecasts to the AR (estimated with EOS data), for a lag order of 1, using two values of q (the number of vintages). $q = 5$ implies that the final values are well approximated by the data published one year after the first release and $q = 14$ suggests that good approximations are obtained after three and a half years of revisions. In general, gains from using VB-VAR with respect to the AR estimated with EOS data are similar in size to the gains from using RTV data for AR forecasting, and in many cases the VB-VAR performance is worse. The KK model is also specified with similar values of q, with an AR order for y_t of four for inflation, and one for output growth. The KK model only improves forecasts of next year's output growth.

These results generally confirm the Monte Carlo. The multiple-vintage models do not dominate the AR model estimated with RTV data on samples of the size we have in the empirical exercise, and that RTV gains over EOS are generally small and mainly for forecasting first-release values.

5.2 ADL models of output growth and inflation

Our analytical results in section 3 show that the use of RTV data to estimate autoregressive models delivers optimal forecasts while EOS data does not because forecasts in real-time are computed conditioned on lightly-revised data, while EOS-estimation uses mainly heavily-revised data. The same argument applies more generally, to multivariate models which include explanatory variables which are subject to revision. In this section, we exploit RTV data to estimate ADL models to forecast output growth and inflation. Of the multiple-vintages approaches considered in the previous section, we estimate a bivariate version of the KK model to compute forecasts, as suggested by Kishor and Koenig (2010). The usefulness of the VB-VAR in a multivariate settings is curtailed by the rapid increase in dimensionality as additional variables are permitted.[13]

We consider two indicators: industrial production and employment. The choice of these indicators is supported by their popularity in forecasting exercises that aim at assessing the predictive power of economic activity variables, and is in part due to their timeliness. Initial estimates of these variables are released prior to the first estimates of output growth and inflation.[14] We consider forecasts of quarterly growth (at an annual rate) 1 and 4 quarters ahead, as before, and in addition we generate 'nowcasts': in this case the horizon is $h = 0$. The ADL models specified for nowcasting correspond closely to the distributed lag models of Koenig et al. (2003).[15] We also consider two in-sample periods. The first begins in 1965Q3, as in the previous sub-section; the second is the shorter estimation period beginning in 1979Q1, to be comparable with Koenig et al. (2003).

The implementation of the RTV approach for ADL(p_y, p_x) models at each forecast horizon h ($h \geq 1$) follows:

$$y_t^{t+1} = \beta_0 + \sum_{i=0}^{p_y-1} \beta_{(1+i)} y_{t-h-i}^{t+1-h} + \sum_{i=0}^{p_x-1} \gamma_i x_{t-h-i}^{t+1-h} + e_{RTV,t} \quad (15)$$

for $t = \max(p_y + h, p_x + h) + 1, ..., T$, where p_y is the autoregressive order and p_x is the number of lags of the indicator. In the case of nowcasting ($h = 0$), this becomes:

$$y_t^{t+1} = \beta_0 + \sum_{i=1}^{p_y} \beta_i y_{t-i}^{t} + \sum_{i=0}^{p_x-1} \gamma_i x_{t-i}^{t+1} + e_{RTV,t}.$$

[13] For example, there will be a doubling of the number of coefficients to be estimated in each equation of the system in an unrestricted bivariate VB-VAR relative to the single-variable models we consider.

[14] Data are obtained from the real-time dataset of the Philadelphia Fed. Industrial production (total) and employment (non-farm payroll) data are published monthly. We construct quarterly vintages by using the vintage released in the first month of each quarter. So when the observation t refers, say, to a first quarter of the year, the vintage $t + 1$ refers to the April release, such that x_t^{t+1} is an initial release.

[15] Koenig et al. (2003) regressed quarterly output growth on the contemporaneous and 4 lags of the indicator sampled monthly. Preliminary investigation suggested nothing is lost by using the contemporaneous and one lag of the indicator sampled quarterly instead.

This implies that when computing nowcasts of y_{T+1}, we use information on the indicator up to x_{T+1}^{T+2}, while one-step-ahead forecasts of y_{T+1} employ information up to x_T^{T+1}. The nowcasts are feasible in real-time given the earlier availability of the indicator data. Forecasts of post-revision data (y_{T+h}^{T+h+14} and y_{T+h}) use forecasts of first release data plus the bias-correction.

Appendix B describes our implementation of Kishor and Koenig (2011), including our adaptation of the model for nowcasting. Because the number of data estimates (q) is assumed to be the same for y and x, we set $q = 5$, since the data revisions on the indicators generally occur up to one year after the first release. The RMSFE of forecasts generated from the KK model is compared to the RMSFE values of an ADL model estimated with EOS data, so that the benchmark is a model that ignores the fact that data are revised over time.

Table 6 shows improvements in forecasting accuracy from using RTV instead of EOS data to estimate ADL models of around 4 up to 10% for forecasting next quarter's growth and GDP deflator inflation (both $h = 0$ and $h = 1$). These gains exceed those obtained with univariate models. However, the gains observed for output growth depend on the estimation period: gains for both first-released and post-revision values arise only when the estimation sample starts in 1979Q1, as in Koenig et al. (2003). This primarily reflects a worsening in the EOS-based forecasts when the sample is shortened, with the RTV-generated forecasts being less affected by sample size. Only 2 to 3% reductions in RMSFE are observed when forecasting PCE inflation.

Forecasts using the KK model are generally not more accurate than ADL forecasts computed with RTV data, especially for the shorter in-sample period. However, when using industrial production to predict next year's post-revision values of output growth and both measures of inflation, the KK is more accurate than the ADL model. This suggests that KK may have a relative advantage at longer horizons ($h = 4$), while ADL models with RTV data are preferred at shorter horizons ($h = 0, 1$).

In general, we find that the inclusion of explanatory variables results in only modest reductions in RMSFEs in comparison with an autoregressive model, with the exception of nowcasting output growth. Thus the results of our real-time forecasting exercise are in tune with those of the pseudo out-of-sample exercise of Stock and Watson (2008), who report that indicator variables only episodically enhance inflation forecasts over the period from 1985 onwards.

In summary, the estimation of multivariate models with RTV data as a method of handling the effect of data revisions improves forecasts in comparison with the traditional real-time estimation approach. The reductions in RMSFE with bivariate models are larger than with univariate models. The use of a bivariate KK model does not generally improve forecasts in comparison with ADL models estimated with RTV data.

6 Conclusions

In recent times there has been a growing appreciation of the potential importance of data revisions for forecasting (see the review by Croushore (2011b)). We have tackled a number of aspects of forecasting when there are data revisions. Firstly, can we improve real-time forecasts of autoregressive models by exploiting better the real-time data currently available? Our analytical results show that the use of the latest-available vintage at each point in time (EOS) will lead to AR model forecasts which do not minimise the real-time forecasting loss function, matching the findings of Koenig et al. (2003) for distributed lag models. The use of lightly-revised data in estimation (RTV) improves forecast accuracy in population even if data revisions are non-zero mean. The forecast accuracy spread between RTV and EOS for autoregressive models is larger for news revisions and for highly persistent autoregressive processes. Based on parameters calibrated to US and UK data, the implication of our analytical results is that the RMSFE reductions are small (2%) when forecasting first-release data and even smaller when predicting post-revision data. The empirical forecast performance of autoregressive models in predicting US output growth and inflation support these analytical results. However, when autoregressive-distributed lag models are employed, the observed RMSFE reductions are in the 3-10% range.

Secondly, do more elaborate, complicated models that simultaneously model the true process and the revisions process yield more accurate forecasts? The findings of the recent empirical forecast comparison literature suggests that 'simple' models may often generate forecasts that are competitive with those of more sophisticated models,[16] so this is an interesting issue to explore. We use two multiple-vintage models to help us to answer this question: the vintage-based VAR, adapted from Hecq and Jacobs (2009), and the Kishor and Koenig (2011) model. Our Monte Carlo results show that even when these models are the data generating process, their forecast performance may be inferior to autoregressive models estimated with RTV data in sample sizes smaller than 200 observations. The empirical exercise supports this claim when forecasting US output growth and inflation. We also compare forecasting models that are able to exploit the content of indicators, which are also subject to revisions, to predict output growth and inflation. We find that the autoregressive distributed lag model estimated with RTV data is more accurate than the Kishor and Koenig (2011) multiple-vintage model at short horizons (nowcasting and next quarter), while the multiple-vintage model appears to be more accurate at longer horizons (next year).

The reorganisation of past real-time data in real-time data vintages (RTV), as suggested by Koenig et al. (2003), provides a simple way of handling the impact of data revisions in forecasting with models. We show that the method delivers forecasts that are generally more accurate than multiple-vintage models for the sample sizes of national accounts data typically available, and especially when the target is first-release instead of post-revision data.

References

Aruoba SB. 2008. Data revisions are not well-behaved. *Journal of Money, Credit and Banking* **40**: 319–340.

Clements MP, Galvão AB. 2008. Macroeconomic forecasting with mixed-frequency data:

[16] See, for example, Fildes and Makridakis (1995), Clements and Hendry (1999) and Makridakis and Hibon (2000).

Forecasting output growth in the United States. *Journal of Business and Economic Statistics* **26**: 546–554. No. 4.

Clements MP, Galvão AB. 2009. Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics* **24**: 1187–1206.

Clements MP, Hendry DF. 1999. *Forecasting Non-stationary Economic Time Series*. Cambridge, Mass.: MIT Press.

Corradi V, Fernandez A, Swanson NR. 2009. Information in the revision process of real-time datasets. *Journal of Business and Economic Statistics* **27**: 455–467.

Croushore D. 2006. Forecasting with real-time macroeconomic data. In Elliott G, Granger C, Timmermann A (eds.) *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*. Elsevier, Horth-Holland, 961–982.

Croushore D. 2011a. Forecasting with real-time data vintages, chapter 9. In Clements MP, Hendry DF (eds.) *The Oxford Handbook of Economic Forecasting*. Oxford University Press, 247–267.

Croushore D. 2011b. Frontiers of real-time data analysis. *Journal of Economic Literature* **49**: 72–100.

Croushore D, Stark T. 2001. A real-time data set for macroeconomists. *Journal of Econometrics* **105**: 111–130.

Croushore D, Stark T. 2003. A real-time data set for macroeconomists: Does the data vintage matter? *The Review of Economics and Statistics* **85**: 605–617.

Cunningham A, Eklund J, Jeffery C, Kapetanios G, Labhard V. 2009. A state space approach to extracting the signal from uncertain data. *Journal of Business & Economic Statistics* Ahead of print. doi:10.1198/jbes.2009.08171.

Diebold FX, Rudebusch GD. 1991a. Forecasting output with the composite leading index: A real-time analysis. *Journal of the American Statistical Association* **86**: 603–610.

Diebold FX, Rudebusch GD. 1991b. Turning point prediction with the composite leading index: An ex ante analysis. In Lahiri K, Moore GH (eds.) *Leading economic indicators: New approaches and forecasting records*. Cambridge University Press, Cambridge, 231–256.

Faust J, Rogers JH, Wright JH. 2003. Exchange rate forecasting: The errors we've really made. *Journal of International Economic Review* **60**: 35–39.

Fildes RA, Makridakis S. 1995. The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review* **63**: 289–308.

Garratt A, Lee K, Mise E, Shields K. 2008. Real time representations of the output gap. *Review of Economics and Statistics* **90**: 792–804.

Garratt A, Lee K, Mise E, Shields K. 2009. Real time representations of the UK output gap in the presence of model uncertainty. *International Journal of Forecasting* **25**: 81–102.

Harrison R, Kapetanios G, Yates T. 2005. Forecasting with measurement errors in dynamic models. *International Journal of Forecasting* **21**: 595–607.

Harvey AC, McKenzie CR, Blake DPC, Desai MJ. 1983. Irregular data revisions. In Zellner A (ed.) *Applied Time Series Analysis of Economic Data*. US Department of Commerce, Washington D.C., Economic Research Report ER-5, 329–347.

Hecq A, Jacobs JPAM. 2009. On the VAR-VECM representation of real time data. Technical report, mimeo, University of Maastricht, Department of Quantitative Economics.

Howrey EP. 1978. The use of preliminary data in economic forecasting. *The Review of Economics and Statistics* **60**: 193–201.

Howrey EP. 1984. Data revisions, reconstruction and prediction: an application to inventory investment. *The Review of Economics and Statistics* **66**: 386–393.

Jacobs JPAM, van Norden S. 2011. Modeling data revisions: Measurement error and dynamics of 'true' values. *Journal of Econometrics* **161**: 101–109.

Kishor NK, Koenig EF. 2010. Yield-spreads as predictors of economic activity: A real-time VAR analysis Federal Reserve Bank of Dallas.

Kishor NK, Koenig EF. 2011. VAR estimation and forecasting when data are subject to revision. *Journal of Business and Economic Statistics* Forthcoming.

Knetsch TA, Reimers HE. 2009. Dealing with benchmark revisions in real-time data: The case of german production and orders statistics. *Oxford Bulletin of Economics and Statistics* **71**: 209–235.

Koenig EF, Dolmas S, Piger J. 2003. The use and abuse of real-time data in economic forecasting. *The Review of Economics and Statistics* **85(3)**: 618–628.

Makridakis S, Hibon M. 2000. The M3 Competition: results, conclusions and implications. *International Journal of Forecasting* **16**: 451–476.

Mankiw NG, Shapiro MD. 1986. News or noise: An analysis of GNP revisions. *Survey of Current Business (May 1986), US Department of Commerce, Bureau of Economic Analysis* : 20–25.

McConnell MM, Perez-Quiros GP. 2000. Output fluctuations in the United States: What has changed since the early 1980s? *American Economic Review* **90**: 1464–1476.

Nalewaik JJ. 2008. Lack of signal error (LoSE) and implications for ols regression: Measurement error for macro data. Finance and economics discussion series, 2008-15, Federal Reserve Board, Washington, D.C.

Orphanides A. 2001. Monetary policy rules based on real-time data. *American Economic Review* **91(4)**: 964–985.

Orphanides A, van Norden S. 2005. The reliability of inflation forecasts based on output gaps in real time. *Journal of Money, Credit and Banking* **37**: 583–601.

Patterson KD. 1995. An integrated model of the data measurement and data generation processes with an application to consumers' expenditure. *Economic Journal* **105**: 54–76.

Patterson KD. 2003. Exploiting information in vintages of time-series data. *International Journal of Forecasting* **19**: 177–197.

Robertson JC, Tallman EW. 1998. Data vintages and measuring forecast model performance. *Federal Reserve Bank of Atlanta Economic Review, Fourth Quarter* : 4–20.

Sargent TJ. 1989. Two models of measurements and the investment accelerator. *Journal of Political Economy* **97**: 251–287.

Siklos PL. 2008. What can we learn from comprehensive data revisions for forecasting inflation: Some US evidence. In Rapach DE, Wohar ME (eds.) *Forecasting in the Presence of Structural Breaks and Model Uncertainty. Frontiers of Economics and Globalization. Volume 3*. Emerald, 271–299.

Stark T, Croushore D. 2002. Forecasting with a real-time data set for macroeconomists. *Journal of Macroeconomics* **24**: 507–531.

Stock JH, Watson MW. 2003. How Did Leading Indicator Forecasts Perform During the 2001 Recession. *Federal Reserve Bank of Richmond, Economic Quarterly* **89/3**: 71–90.

Stock JH, Watson MW. 2008. Phillips Curve Inflation Forecasts. Working paper 14322, NBER, Cambridge, MA.

A Proofs

Proof. Proposition 1: Optimal AR population parameter.

Firstly, suppose revisions are zero-mean. The expected squared forecast error is given by:

$$
\begin{aligned}
& E\left(y_{T+1}^{T+1+f}-\boldsymbol{\phi}'\mathbf{y}_T^{T+1}\right)^2 \\
= \quad & E\left(\rho_0+\boldsymbol{\rho}'\widetilde{\mathbf{y}}_T+R_1\eta_{1T+1}-v_{T+1}^{T+2}+v_{T+1}^{T+1+f}+\varepsilon_{T+1}^{T+1+f}-\phi_0-\boldsymbol{\phi}'\widetilde{\mathbf{y}}_T-\boldsymbol{\phi}'\mathbf{v}_T^{T+1}-\boldsymbol{\phi}'\boldsymbol{\varepsilon}_T^{T+1}\right)^2
\end{aligned}
$$

where $\mathbf{y}_T^{T+1}=\widetilde{\mathbf{y}}_T+\mathbf{v}_T^{T+1}+\boldsymbol{\varepsilon}_T^{T+1}$, with $\mathbf{y}_T^{T+1\prime}=\left(y_T^{T+1},\ldots,y_{T-p+1}^{T+1}\right)$, $\widetilde{\mathbf{y}}_T'=(\widetilde{y}_T,\ldots,\widetilde{y}_{T-p+1})$, $\mathbf{v}_T^{T+1\prime}=\left(v_T^{T+1},\ldots,v_{T-p+1}^{T+1}\right)$, with typical element $v_{T-j}^{T+1}=-\sum_{i=j+1}^{l}\sigma_{v_i}\eta_{2,T-j,i}$, for $j<l$, and $v_{T-j}^{T+1}=\sigma_{v_i}\eta_{2,T-j,l}$, for $j\geq l$, and $\boldsymbol{\varepsilon}_T^{T+1\prime}=\left(\varepsilon_T^{T+1},\ldots,\varepsilon_{T-p+1}^{T+1}\right)$. Note that if $f=1$, $\mathbf{y}_{T+1}^{T+2}=\widetilde{\mathbf{y}}_{T+1}+\mathbf{v}_{T+1}^{T+2}+\boldsymbol{\varepsilon}_{T+1}^{T+2}=\rho_0+\boldsymbol{\rho}'\widetilde{\mathbf{y}}_T+R_1\eta_{1T+1}+\varepsilon_{T+1}^{T+2}$. We let $E\left(\widetilde{\mathbf{y}}_T\widetilde{\mathbf{y}}_T'\right)=\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}}+\boldsymbol{\mu}_{\widetilde{y}}\boldsymbol{\mu}_{\widetilde{y}}'$, where $\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}}=Var\left(\widetilde{\mathbf{y}}_T\right)$, $\boldsymbol{\mu}_{\widetilde{y}}=\mathbf{i}\mu_{\widetilde{y}}=E\left(\widetilde{\mathbf{y}}_T\right)$, $\mathbf{i}$ a p-dimensional vector of 1's; $\boldsymbol{\Sigma}_{\mathbf{v}}\equiv Var\left(\mathbf{v}_T^{T+1}\mathbf{v}_T^{T+1\prime}\right)=E\left(\mathbf{v}_T^{T+1}\mathbf{v}_T^{T+1\prime}\right)=diag\left(\sum_{i=1}^{l}\sigma_{v_i}^2,,\ldots,\sum_{i=p}^{l}\sigma_{v_i}^2\right)$ for $p\leq l$, with terms of $\sigma_{v_l}^2$ on the diagonal for $p>l$; $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\equiv Var\left(\boldsymbol{\varepsilon}_T^{T+1}\boldsymbol{\varepsilon}_T^{T+1\prime}\right)=E\left(\boldsymbol{\varepsilon}_T^{T+1}\boldsymbol{\varepsilon}_T^{T+1\prime}\right)=diag\left(\sigma_{\varepsilon_1}^2,\sigma_{\varepsilon_2}^2,\ldots\sigma_{\varepsilon_p}^2\right)$, with $\sigma_{\varepsilon_s}^2=\sigma_{\varepsilon_l}^2$ for $s>l$, and $\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}\mathbf{v}}\equiv Var\left(\widetilde{\mathbf{y}}_T\mathbf{v}_T^{T+1\prime}\right)=E\left(\widetilde{\mathbf{y}}_T\mathbf{v}_T^{T+1\prime}\right)$. Solving the first-order conditions $\partial E\left(y_{T+1}^{T+f}-\boldsymbol{\phi}'\mathbf{y}_T^{T+1}\right)^2/\partial\boldsymbol{\phi}=\mathbf{0}$ and $\partial E\left(y_{T+1}^{T+f}-\boldsymbol{\phi}'\mathbf{y}_T^{T+1}\right)^2/\partial\phi_0=\mathbf{0}$ gives:

$$
\begin{aligned}
\boldsymbol{\phi}^* &= \left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}}+\Sigma_{\mathbf{v}}+\Sigma_{\widetilde{\mathbf{y}}\mathbf{v}}+\Sigma_{\widetilde{\mathbf{y}}\mathbf{v}}'+\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)^{-1}\left(\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}}+\boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}\mathbf{v}}'\right)\boldsymbol{\rho} \\
\phi_0^* &= \left(1-\boldsymbol{\phi}^{*\prime}\mathbf{i}\right)\mu_{\widetilde{y}}.
\end{aligned}
$$

These are optimal for $f\geq 1$ if revisions are zero mean. Suppose $f=1$ and now allow for non-zero mean revisions. The optimal slope parameter $\boldsymbol{\phi}^*$ only depends on (centred) second moment matrices, so is unaffected whether the DGP is given by (2)-(5)-(6) or (2)-(4). The second of the two FOC's implies $E\left(y_{T+1}^{T+2}\right)=\phi_0+E\left(\boldsymbol{\phi}'\mathbf{y}_T^{T+1}\right)$. Assuming (2)-(5)-(6) hold, we have $E\left(y_{T+1}^{T+2}\right)=E\left(\widetilde{y}_t+v_t^{t+1}+\varepsilon_t^{t+1}\right)=\mu_{\widetilde{y}}-\sum_{i=1}^{l}\mu_{v_i}-\mu_{\varepsilon_1}$, and $E\left(\boldsymbol{\phi}'\mathbf{y}_T^{T+1}\right)=\boldsymbol{\phi}'\left(\mathbf{i}\mu_{\widetilde{y}}+\boldsymbol{\mu}_\varepsilon+\boldsymbol{\mu}_v\right)$, with $\mathbf{i}\mu_{\widetilde{y}}=E\left([\widetilde{\mathbf{y}}_{t-1}]\right)=E\left([\widetilde{y}_{t-1},\ldots,\widetilde{y}_{t-p}]'\right)$, $\mathbf{i}$ a p-dimensional vector of 1's, and $\boldsymbol{\mu}_\varepsilon$ and $\boldsymbol{\mu}_v$ are the means of the noise and news revisions, $\boldsymbol{\mu}_\varepsilon\equiv E\left([\varepsilon_{t-1}^t,\ldots,\varepsilon_{t-p}^t]\right)=\left[-\mu_{\varepsilon_1},...,-\mu_{\varepsilon_p}\right]$, $\boldsymbol{\mu}_v\equiv E\left([v_{t-1}^t,\ldots,v_{t-p}^t]\right)=\left[-\sum_{i=1}^{l}\mu_{v_i},...,-\sum_{i=p}^{l}\mu_{v_i}\right]$. Finally, $\mu_{\widetilde{y}}=(1-\rho(1))^{-1}\left[\rho_0+\sum_{i=1}^{l}\mu_{v_i}\right]$. Hence it follows that:

$$
\phi_0^*=\left(1-\boldsymbol{\phi}^{*\prime}\mathbf{i}\right)\mu_{\widetilde{y}}-\sum_{i=1}^{l}\mu_{v_i}-\mu_{\varepsilon_1}-\boldsymbol{\phi}^{*\prime}\boldsymbol{\mu}_\varepsilon-\boldsymbol{\phi}^{*\prime}\boldsymbol{\mu}_v.
$$

■

Proof. Proposition 2: Estimation of AR(p) with EOS data.

If we allow $T\to\infty$ for a fixed l, then the OLS estimator using EOS data is asymptotically equivalent to the estimator obtained from a regression using only fully-revised data. Denote this data by $\{y_t, y_{t-1} \dots y_{t-p}\}$, where $y_t = \widetilde{y}_t + v_t^{t+l} + \varepsilon_t^{t+l}$, and $y_t = \widetilde{y}_t$ if the true data are eventually revealed. As the data $\left\{y_t, \mathbf{y}_{t-1} = (y_{t-1} \dots y_{t-p})'\right\}$ are covariance stationary we can use the population moments (assuming that $T \to \infty$) to compute the OLS estimators as values that satisfy

$$(\alpha_0^*, \boldsymbol{\alpha}^*) = \underset{\alpha_0, \boldsymbol{\alpha}}{\arg\min} E\left(y_t - \alpha_0 - \boldsymbol{\alpha}\mathbf{y}_{t-1}\right)^2,$$

which are:

$$\alpha_0^* = E(y_t) - \boldsymbol{\alpha}E(\mathbf{y}_{t-1}) \tag{16}$$

and

$$E\left(\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right)\boldsymbol{\alpha}^* + \alpha_0^* E(\mathbf{y}_{t-1}) - E(y_t\mathbf{y}_{t-1}) = 0. \tag{17}$$

Combining (16) and (17) gives the standard FOC for $\boldsymbol{\alpha}^*$,

$$Cov\left(\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right)\boldsymbol{\alpha} = Cov(y_t\mathbf{y}_{t-1}) \tag{18}$$

where $Cov\left(\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right) = E\left(\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right) - E(\mathbf{y}_{t-1})E\left(\mathbf{y}_{t-1}'\right)$, $Cov(y_t\mathbf{y}_{t-1}) = E(y_t\mathbf{y}_{t-1}) - E(y_t)E(\mathbf{y}_{t-1})$.

The moments in (16) are obtained from $E(y_t) = E\left(\widetilde{y}_t + v_t^{t+l} + \varepsilon_t^{t+l}\right) = \mu_{\widetilde{y}} - \mu_{v_l} - \mu_{\varepsilon_l}$, and $E(\mathbf{y}_{t-1}) = \mathbf{i}\mu_{\widetilde{y}} - \mathbf{i}\mu_{v_l} - \mathbf{i}\mu_{\varepsilon_1}$, since $\mathbf{y}_{t-1} = \widetilde{\mathbf{y}}_{t-1} + \mathbf{v}_{t-1}^{t+l} + \boldsymbol{\varepsilon}_{t-1}^{t+l}$, with $\widetilde{\mathbf{y}}_{t-1} = [\widetilde{y}_{t-1}, \dots, \widetilde{y}_{t-p}]'$, $\mathbf{v}_{t-1}^{t+l} = \left[v_{t-1}^{t+l}, \dots, v_{t-p}^{t+l}\right]'$, $\boldsymbol{\varepsilon}_{t-1}^{t+l} = \left[\varepsilon_{t-1}^{t+l}, \dots, \varepsilon_{t-p}^{t+l}\right]'$, so that:

$$\alpha_0^* = \left(1 - \boldsymbol{\alpha}^{*\prime}\mathbf{i}\right)\left(\mu_{\widetilde{y}} - \mu_{v_l} - \mu_{\varepsilon_l}\right). \tag{19}$$

For (18) we obtain:

$$Cov\left(\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\right) = \Sigma_{\widetilde{y}} + \boldsymbol{\Sigma}_{\underline{\mathbf{v}}} + \boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}} + \Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}} + \Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}}' \tag{20}$$

where $\Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}} \equiv E\left[(\widetilde{\mathbf{y}}_{t-1} - E(\widetilde{\mathbf{y}}_{t-1}))\left(\mathbf{v}_{t-1}^{t+l} - E\left(\mathbf{v}_{t-1}^{t+l}\right)\right)'\right]$, $\boldsymbol{\Sigma}_{\underline{\mathbf{v}}} = E\left[\mathbf{v}_{t-1}^{t+l} - E\left(\mathbf{v}_{t-1}^{t+l}\right)\left(\mathbf{v}_{t-1}^{t+l} - E\left(\mathbf{v}_{t-1}^{t+l}\right)\right)'\right] = \sigma_{v_l}^2 I_p$, $\boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}} = E\left[\boldsymbol{\varepsilon}_{t-1}^{t+l} - E\left(\boldsymbol{\varepsilon}_{t-1}^{t+l}\right)\left(\boldsymbol{\varepsilon}_{t-1}^{t+l} - E\left(\boldsymbol{\varepsilon}_{t-1}^{t+l}\right)\right)'\right] = \sigma_{\varepsilon_l}^2 I_p$. Note that $\Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}}$ is upper diagonal, and its diagonal is (minus) the diagonal of $\boldsymbol{\Sigma}_{\underline{\mathbf{v}}}$. Also:

$$Cov(y_t\mathbf{y}_{t-1}) = V\left(\widetilde{\mathbf{y}}_{t-1}\widetilde{\mathbf{y}}_{t-1}'\right)\rho + V\left(\mathbf{v}_{t-1}^{t+l}\widetilde{\mathbf{y}}_{t-1}'\right)\rho = \Sigma_{\widetilde{\mathbf{y}}}\rho + \Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}}'\rho \tag{21}$$

Substituting (20) and (21) into (18) gives:

$$\boldsymbol{\alpha}^* = \left(\Sigma_{\widetilde{\mathbf{y}}} + \boldsymbol{\Sigma}_{\underline{\mathbf{v}}} + \boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}} + \Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}} + \Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}}'\right)^{-1}\left(\Sigma_{\widetilde{\mathbf{y}}} + \Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}}'\right)\rho. \tag{22}$$

Note that $\boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}} \neq \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = diag\left\{\sigma_{\varepsilon_1}^2, \dots \sigma_{\varepsilon_p}^2\right\}$, $\Sigma_{\widetilde{\mathbf{y}}\underline{\mathbf{v}}} \neq \boldsymbol{\Sigma}_{\widetilde{\mathbf{y}}\mathbf{v}}$ and $\boldsymbol{\Sigma}_{\underline{\mathbf{v}}} \neq \boldsymbol{\Sigma}_{\mathbf{v}}$ from the assumption that in large samples the use of EOS data is approximated by the use of data from the $t+l$ vintage. Hence when $\sigma_{v_l} = \sigma_{\varepsilon_l} = 0$, the use of EOS gives the same large-sample estimates of the model parameters as using the true data $\{\widetilde{y}_t\}$. ■

Proof. **Remark 1:** Forecasts computed using EOS data to estimate AR(p) model are biased when data revisions are described by (2)-(5)-(6).

Consider the first moment properties of the forecast errors when an AR(p) is estimated by EOS in the presence of non-zero mean data revisions. Suppose the aim is to forecast the first vintage estimate y_{T+1}^{T+2}. To see that EOS population forecasts are generally biased, note that by construction, $\{\alpha_0^*, \boldsymbol{\alpha}^*\}$ satisfy $E\left(y_{T+1} - \alpha_0 - \boldsymbol{\alpha}'\mathbf{y}_T\right) = 0$, where y_{T+1} and $\mathbf{y}_T = (y_T, \ldots, y_{T-p+1})$ are 'final data'. But forecasts are of necessity conditioned on $\mathbf{y}_T^{T+1} = \left(y_T^{T+1}, \ldots, y_{T-p+1}^{T+1}\right)$, so of interest is the expected value of the forecast error $E\left(y_{T+1}^{T+2} - \alpha_0^* - \boldsymbol{\alpha}^{*\prime}\mathbf{y}_T^{T+1}\right)$. Substituting $y_{T+1}^{T+2} = y_{T+1} - \left(y_{T+1} - y_{T+1}^{T+2}\right)$ and $\mathbf{y}_T^{T+1} = \mathbf{y}_T - \left(\mathbf{y}_T - \mathbf{y}_T^{T+1}\right)$ gives:

$$
\begin{aligned}
E\left(y_{T+1}^{T+2} - \alpha_0^* - \boldsymbol{\alpha}^{*\prime}\mathbf{y}_T^{T+1}\right) &= E\left(y_{T+1}^{T+2} - y_{T+1} - \boldsymbol{\alpha}^{*\prime}\left(\mathbf{y}_T^{T+1} - \mathbf{y}_T\right)\right) \\
&= \left(\mu_{v_l} - \textstyle\sum_{i=1}^{l}\mu_{v_i}\right) + \left(\mu_{\varepsilon_l} - \mu_{\varepsilon_1}\right) - \boldsymbol{\alpha}^{*\prime}\left[\left(\boldsymbol{\mu}_v + \mathbf{i}\mu_{v_l}\right) + \left(\boldsymbol{\mu}_\varepsilon + \mathbf{i}\mu_{\varepsilon_l}\right)\right]
\end{aligned}
$$

since $E\left(y_{T+1}^{T+2}\right) = \mu_{\widetilde{y}} + E\left(v_{T+1}^{T+2}\right) + E\left(\varepsilon_{T+1}^{T+2}\right) = \mu_{\widetilde{y}} - \sum_{i=1}^{l}\mu_{v_i} - \mu_{\varepsilon_1}$, and $E\left(y_{T+1}\right) = \mu_{\widetilde{y}} + E\left(v_{T+1}^{T+1+l}\right) + E\left(\varepsilon_{T+1}^{T+1+l}\right) = \mu_{\widetilde{y}} - \mu_{v_l} - \mu_{\varepsilon_l}$, and similarly for $E\left(\mathbf{y}_T^{T+1}\right)$ and $E\left(\mathbf{y}_T\right)$. Recall that $\boldsymbol{\mu}_\varepsilon \equiv E\left(\left[\varepsilon_{t-1}^t, \ldots, \varepsilon_{t-p}^t\right]\right) = \left[-\mu_{\varepsilon_1}, ..., -\mu_{\varepsilon_p}\right]$, $\boldsymbol{\mu}_v \equiv E\left(\left[v_{t-1}^t, \ldots, v_{t-p}^t\right]\right) = \left[-\sum_{i=1}^{l}\mu_{v_i}, ..., -\sum_{i=p}^{l}\mu_{v_i}\right]$ and $\mu_{\widetilde{y}} = \left(1 - \rho\left(1\right)\right)^{-1}\left[\rho_0 + \sum_{i=1}^{l}\mu_{v_i}\right]$. The expression for this bias will not equal zero unless the means of all the revisions are zero. ■

B Bivariate Kishor and Koenig (2011) model

This appendix describes the bivariate version of the KK model estimated in section 5.2. This specification allows for the use of a leading indicator which is also subject to data revisions. As in section 4.3, the data are organised by vintage. Similarly to $\mathbf{z}^{t+1}$ and $\mathbf{z}_t$ defined in section 4.3 for the target forecast variable y, the vectors of the $t+1$ vintage and the true data of a leading indicator x are defined by $\mathbf{x}^{t+1} = (x_t^{t+1}, x_{t-1}^{t+1}, ..., x_{t-q+1}^{t+1})'$, and

$\mathbf{x}_t = (x_t, ..., x_{t-q+1})'$. The measurement and state equations are, respectively:

$$\begin{bmatrix} \mathbf{z}^{t+1} \\ \mathbf{x}^{t+1} \end{bmatrix} = \begin{bmatrix} I_{q\times q} & 0_{q\times q} & I_{q\times q} & 0_{q\times q} \\ 0_{q\times q} & I_{q\times q} & 0_{q\times q} & I_{q\times q} \end{bmatrix} \begin{bmatrix} \mathbf{z}_t \\ \mathbf{x}_t \\ \mathbf{z}^{t+1} - \mathbf{z}_t \\ \mathbf{x}^{t+1} - \mathbf{x}_t \end{bmatrix},$$

$$\begin{bmatrix} \mathbf{z}_t \\ \mathbf{x}_t \\ \mathbf{z}^{t+1} - \mathbf{z}_t \\ \mathbf{x}^{t+1} - \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{F} & 0_{2q\times 2q} \\ 0_{2q\times 2q} & \mathbf{K} \end{bmatrix} \begin{bmatrix} \mathbf{z}_{t-1} \\ \mathbf{x}_{t-1} \\ \mathbf{z}^{t} - \mathbf{z}_{t-1} \\ \mathbf{x}^{t} - \mathbf{x}_{t-1} \end{bmatrix} + \begin{bmatrix} v_t \\ \varepsilon_t \end{bmatrix},$$

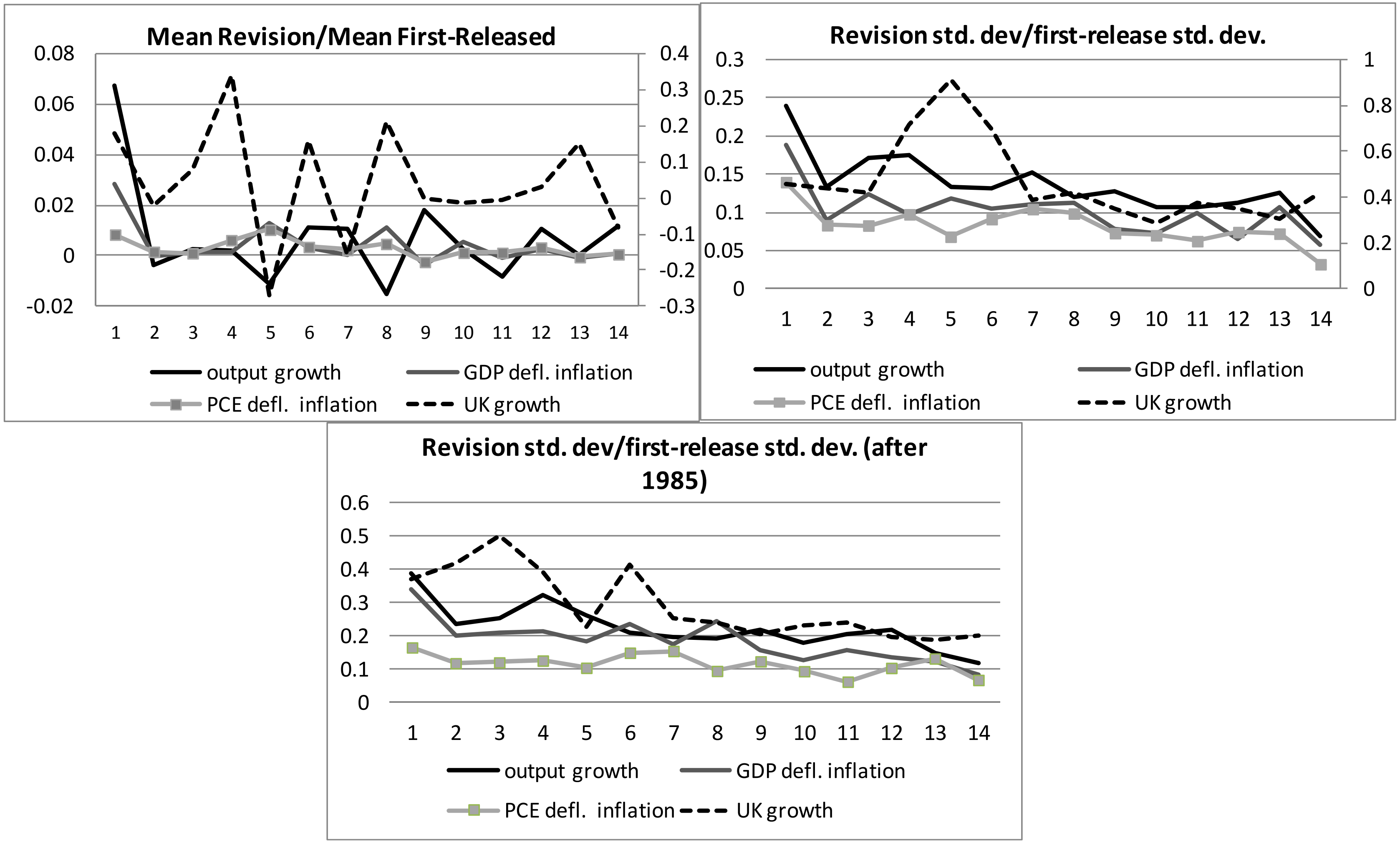
where the matrices $\mathbf{c}_1, \mathbf{c}_2$ are now of dimension $2q \times 1$, $2q \times 1$, and $\mathbf{F}$ and $\mathbf{K}$ are of dimension $2q \times 2q$. We define matrix $\mathbf{F}$ such that the number of lags of z and x (p_y and p_x) may differ, while the same RHS variables are included in each equation of the bivariate VAR (assuming that maximum number of lags is smaller than $(q-1)$). $\mathbf{K}$ is defined such that neither past data revisions to z help predict data revisions to x, and nor do revisions to x help predict revisions to z. This assumption is justified by the data being released by different statistical agencies with different operating procedures, but in any case allows for greater parsimony.

As in the case of the univariate model, the coefficients in $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{F}$, $\mathbf{K}$ and the disturbances variance-covariance matrix are estimated by SURE using observations up to $T-q+1$ of vintages up to $T+1$. Then based on the filtered estimates of z and x, forecasts of $y_{T+1,...}, y_{T+4}$ are computed by iteration of the estimated state equation. Forecasts of first-released data, $y_{T+1}^{T+2}, ..., y_{T+4}^{T+5}$, are obtained by combining the predictions of post-revision values (using forecasts of $\mathbf{z}_t$) with the predictions for data revisions (using forecasts of $(\mathbf{z}^{t+1} - \mathbf{z}_t)$).

When nowcasting z_t, we assume we have information on z up to vintage $t+1$ and on x up to vintage $t+2$, since the first estimate of indicators such as employment and industrial production are released before the first estimate of real GDP, GDP and PCE deflators.

Note also that the quarterly vintage on the indicator refers to the vintage published in the first month of the quarter. Then, the vector of observables becomes $\left(\mathbf{z}^{t+1}, \mathbf{x}^{t+2}\right)'$, while the state vector is $\left(\mathbf{z}^{t+1}, \mathbf{x}^{t+2}, \mathbf{z}^{t+1} - \mathbf{z}_t, \mathbf{x}^{t+2} - \mathbf{x}_{t+1}\right)'$. The $h = 0$ forecast is computed using the estimates of the state equation conditioned on the filtered values of $z_{t|t}$ and $x_{t+1|t+1}$.

Figure 1: The mean and standard deviation of the revisions relative to the first-released data.



Note: Revisions are defined as $r_t^{(i)} = y_t^{t+1+i} - y_t^{t+i}$ for i = 1,...,*l*. So that *i=1* denotes the revision between the first and second estimate, etc. The series are for the US, unless otherwise specified, and the dual axes in the first two plots reflect the very different magnitudes of revisions in the two countries: the left scale refers to the US, the right scale to the UK.

Table 1: Parameter values of the data generating process described in Section 3.1.

DGP n.	ρ_0	ρ_1	ρ_2	R_1	μ_{r_1},μ_{r_5}	$\sigma_{y_t^{t+1}}$	$\sigma_{r_1}/\sigma_{y_t^{t+1}}$	$\sigma_{r_2,\ldots,r_{l-1}}/\sigma_{y_t^{t+1}}$	$\sigma_{r_l}/\sigma_{y_t^{t+1}}$
1	.4	.2	.2	.5	.06, .03	.589	.4	.2	.1
2	.4	.2	.2	.5	.20,.10	.589	.4	.5	.3
3	.4	.5	.3	.5	.12, .06	.749	.4	.2	.1
4	.4	.5	.3	.5	.20,.10	.749	.4	.5	.3

Table 2: Small-sample forecast accuracy findings: RMSFEs for the DGPs of Table 1 (*l = 14*)

	Forecasting y_{T+1}^{T+2}					Forecasting y_{T+1}^{T+1+14} (bias-corrected)					Forecasting y_{T+1}^{T+1+14} (with y_t^{t+14} in the LHS)			
DGP n.	Asym. Ratio	T=50	T=100	T=200	T=500	Asym. Ratio	T=50	T=100	T=200	T=500	T=50	T=100	T=200	T=500
Pure Noise														
1	*0.996*	1.003	1.001	0.997	*0.999*	*1.000*	1.005	1.000	1.000	*1.000*	1.062	1.008	0.999	*0.999*
2	*0.980*	0.979	0.982	0.977	*0.982*	*0.998*	1.002	0.998	0.999	*0.999*	1.052	1.004	0.999	*0.999*
3	*0.990*	1.000	0.993	0.993	*0.991*	*0.988*	1.004	0.993	0.991	*0.993*	1.186	1.012	0.993	*0.986*
4	*0.940*	0.925	0.940	0.930	*0.943*	*0.952*	0.960	0.958	0.961	*0.952*	1.121	0.975	0.960	*0.953*
Pure News														
1	*0.996*	0.988	0.990	0.992	*0.995*	*0.999*	1.005	1.001	0.999	*1.000*	1.082	1.013	1.003	*1.001*
2	*0.963*	0.937	0.951	0.954	*0.962*	*0.997*	1.007	1.002	0.999	*0.997*	1.097	1.016	1.002	*0.999*
3	*0.986*	0.971	0.982	0.987	*0.983*	*0.989*	0.998	0.996	0.989	*0.989*	1.233	1.024	0.999	*0.990*
4	*0.918*	0.900	0.908	0.922	*0.918*	*0.959*	0.978	0.963	0.964	*0.959*	1.227	1.003	0.972	*0.962*

Note: Entries are ratios of RMSFE: values smaller than one indicate that RTV reduces the RMSFE. The first columns under “Bias Diff” and “RMSFE ratio” report the population analytical values. The autoregressive order of all the AR models is 2.

Table 3: Comparing RMSFEs with alternative data generating processes: VB-VAR and KK models (q=14).

Table 3A. Using the VB-VAR model (q=14) as data generating process.

	Forecasting y_{T+1}^{T+2}				Forecasting y_{T+1}^{T+15}			
	T=50	T=100	T=200	T=500	T=50	T=100	T=200	T=500
DGP is estimated with:	Comparing AR_RTV/AR_EOS							
Output growth	0.987	0.992	0.992	0.991	1.002	1.005	1.004	1.002
GDP deflator	0.989	0.983	0.986	0.979	0.994	0.999	0.990	0.996
PCE deflator	0.988	0.983	0.975	0.971	1.004	0.999	0.994	0.987
UK growth	0.967	0.961	0.962	0.967	1.007	1.014	1.013	1.019
	Comparing AR_RTV_KDP/AR_EOS							
Output growth					1.021	1.001	0.999	0.999
GDP deflator					1.325	1.006	0.989	0.985
PCE deflator					1.361	1.014	0.988	0.978
UK growth					1.115	0.986	0.976	0.972
	Comparing VB-VAR/AR_EOS							
Output growth	1.214	1.011	0.972	0.949	1.222	1.022	0.975	0.961
GDP deflator	1.172	0.984	0.951	0.920	1.176	0.987	0.959	0.935
PCE deflator	1.192	1.014	0.974	0.952	1.210	1.022	0.989	0.973
UK growth	1.108	0.957	0.913	0.898	1.019	0.955	0.959	0.958

Table 3B. Using the KK model (q=14) as data generating process.

	Forecasting y_{T+1}^{T+2}				Forecasting y_{T+1}^{T+15}			
	T=50	T=100	T=200	T=500	T=50	T=100	T=200	T=500
DGP is estimated with:	Comparing AR_RTV/AR_EOS							
Output growth	0.984	0.989	0.995	0.998	0.997	0.996	0.997	0.999
GDP deflator	1.004	0.999	0.992	0.990	1.003	0.990	0.993	0.979
PCE deflator	1.016	1.004	1.000	0.996	1.011	0.997	0.994	0.990
UK growth	0.948	0.950	0.948	0.961	0.968	0.958	0.985	0.979
	Comparing AR_RTV_KDP/AR_EOS							
Output growth					1.016	1.002	0.999	0.999
GDP deflator					1.290	1.054	0.998	0.984
PCE deflator					1.216	1.029	0.994	0.985
UK growth					0.968	0.969	0.967	0.975
	Comparing KK/AR_EOS							
Output growth	1.064	1.005	0.999	0.999	1.045	1.009	1.001	1.001
GDP deflator	1.221	1.011	0.986	0.976	1.212	1.011	0.977	0.967
PCE deflator	1.178	1.011	0.994	0.989	1.203	1.011	0.990	0.978
UK growth	0.979	0.946	0.951	0.952	0.969	0.945	0.962	0.969

Note: Entries are RMSFE ratios of the indicated comparison. AR_RTV is the autoregressive model estimated with the indicated order (left column) using RTV data, and using bias-correction for forecasting post-revision values. AR_RTV_KDP is an autoregressive model estimated with RTV data on the RHS, but the post-revision values on the LHS. The entries are computed based on 10000 replications. KK model has a diagonal **K** matrix. DGPs are estimated with the full sample of quarterly vintages (US data vintages: 1965Q4-2009Q1, and UK data vintages: 1976Q1-2009Q1). The autoregressive order of the AR and KK models is 1 for US output growth and 4 for all the other variables. The autoregressive order of the VB-VAR is always 1.

Table 4: Characteristics of data releases during the estimation (1965-1985) and out-of-sample forecast periods (1985-2006).

		Mean			Standard Deviation			Autocorrelation (1st)			H_0: Mean = 0		H_0: News		H_0:Noise	
Variable:	Period:	y_t^{t+1}	y_t^{t+14}	y_t^{09Q1}	y_t^{t+1}	y_t^{t+14}	y_t^{09Q1}	y_t^{t+1}	y_t^{t+14}	y_t^{09Q1}	r_{14}	r_{09Q1}	r_{14}	r_{09Q1}	r_{14}	r_{09Q1}
Output growth	1965Q3-1985Q2	0.76	0.79	0.81	1.08	1.07	1.08	0.28	0.29	0.29	[.03]	[.01]	[.08]	[.02]	[.09]	[.01]
	1985Q3-2006:Q4	0.68	0.68	0.75	0.43	0.52	0.50	0.33	0.33	0.23	[.92]	[.06]	[.28]	[.03]	[.00]	[.00]
GDP deflator	1965Q3-1985Q2	1.45	1.43	1.41	0.59	0.59	0.59	0.70	0.75	0.80	[.17]	[.03]	[.48]	[.36]	[.10]	[.03]
	1985Q3-2006:Q4	0.58	0.65	0.60	0.28	0.27	0.23	0.57	0.55	0.56	[.00]	[.30]	[.00]	[.00]	[.00]	[.47]
PCE deflator	1965Q3-1985Q2	1.40	1.39	1.38	0.63	0.64	0.65	0.83	0.83	0.84	[.25]	[.13]	[.52]	[.23]	[.26]	[.04]
	1985Q3-2006:Q4	0.64	0.69	0.64	0.38	0.35	0.30	0.49	0.60	0.56	[.00]	[.76]	[.00]	[.00]	[.00]	[.14]

Note: *p*-values are computed for *F*-statistics (news/noise) and *t*-statistics (mean) using Newey-West standard errors, and are displayed in [].
The revisions are defined as $r_{14} = y_t^{t+14} - y_t^{t+1}$ and $r_{09Q1} = y_t^{09Q1} - y_t^{t+1}$.

Table 5: Comparing the RMSFEs from RTV and EOS in a recursive out-of-sample forecasting exercise (for the period 1985Q3-2008Q4; $n = 94$ quarters).

5A. Forecasting output growth

	h = 1			h = 4		
Model:	y_{T+1}^{T+1+1}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+4}^{T+4+1}	y_{T+4}^{T+4+14}	y_{T+4}^{09Q1}
AR(1)	0.998	1.008	1.001	**0.970**	**0.978**	**0.975**
VB-VAR, q=5	0.995	0.986	0.994	**0.967**	0.983	0.984
VB-VAR, q=14	1.044	1.028	0.999	1.017	1.025	1.001
KK, q=5, p=1	1.040	1.052	1.034	**0.976**	**0.975**	0.982
KK, q=14, p=1	1.065	1.085	1.056	**0.972**	**0.972**	**0.972**

5B. Forecasting GDP deflator inflation

	h = 1			h = 4		
Model:	y_{T+1}^{T+1+1}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+4}^{T+4+1}	y_{T+4}^{T+4+14}	y_{T+4}^{09Q1}
AR(4)	**0.962**	0.992	1.018	**0.956**	1.034	1.035
VB-VAR, q=5	0.988	0.987	0.993	1.034	1.074	1.070
VB-VAR, q=14	**0.955**	0.987	1.011	**0.973**	1.058	1.056
KK, q=5, p=4	0.997	1.008	1.026	1.057	1.096	1.098
KK, q=14, p=4	1.047	1.056	1.147	1.221	1.338	1.334

5C. Forecasting PCE deflator inflation

	h = 1			h = 4		
Model:	y_{T+1}^{T+1+1}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+4}^{T+4+1}	y_{T+4}^{T+4+14}	y_{T+4}^{09Q1}
AR(4)	**0.974**	0.992	0.996	**0.964**	0.990	0.992
VB-VAR, q=5	1.000	1.012	1.009	1.017	1.027	1.022
VB-VAR, q=14	1.068	1.063	1.071	1.054	1.056	1.052
KK, q=5, p=4	1.001	1.011	1.009	1.003	1.014	1.015
KK, q=14, p=4	1.008	1.018	1.039	1.113	1.157	1.170

Note: Entries with values smaller than one indicate that RTV reduces the RMSFE relative to EOS. For VB-VAR and KK models, ratios are relative to the AR(p) with EOS data (same benchmark model for each panel). Forecasts of post-revisions values with AR+RTV are bias-corrected first release forecasts. The correction is based on the difference between the unconditional means of y_t^{t+14} and y_t^{t+1} with data up to the forecast origin. The VB-VAR and KK models are described in the text. The KK model is estimated with a full **K** matrix. Emboldened entries indicate RMSFE reductions larger than 2%. Multiple step-ahead forecasts are computed by iteration. Models are estimated with increasing windows of data. In-sample period starts in 1965Q3.

Table 6: Comparison of ADL models with RTV data and KK models with ADL models with EOS data for forecasting output growth and inflation with activity variables (for the period 1985Q3-2008Q4; *n = 94* quarters).

6A. Forecasting output growth

		h = 0				h = 1				h = 4			
					Ratio AR(1)_RTV				Ratio AR(1)_RTV				Ratio AR(1)_RTV
Model:	sample starts:	y_{T+1}^{T+2}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+1}^{09Q1}	y_{T+1}^{T+2}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+1}^{09Q1}	y_{T+4}^{T+5}	y_{T+4}^{T+19}	y_{T+4}^{09Q1}	y_{T+4}^{09Q1}
		With industrial production											
ADL (1,2)	1965:Q3	1.057	1.033	**0.976**	0.829*	1.020	1.023	0.989	0.956*	0.988	0.997	1.013	1.056
KK (q=5, p=(1,1))		1.223	1.119	1.060	0.902*	1.030	1.019	1.003	0.969*	**0.963**	**0.956**	**0.959**	0.999*
ADL (1,2)	1979:Q1	**0.933**	**0.947**	**0.893**	0.826*	**0.973**	**0.980**	**0.952**	0.958*	**0.944**	0.982	1.000	1.054
KK (q=5, p=(1,1))		1.141	1.042	**0.979**	0.908*	1.247	1.118	1.125	1.136	**0.961**	**0.958**	**0.967**	1.005
		With employment											
ADL (1,2)	1965:Q3	**0.968**	0.983	**0.933**	0.845*	0.995	1.047	0.995	1.006	1.046	1.067	1.073	1.132
KK (q=5, p=(1,1))		1.043	1.034	0.997	0.902*	1.015	1.068	0.999	1.007	**0.931**	**0.934**	**0.948**	1.000
ADL (1,2)	1979:Q1	**0.965**	0.984	**0.941**	0.853*	**0.932**	0.986	**0.948**	1.004	**0.954**	0.996	1.007	1.052
KK (q=5, p=(1,1))		1.684	1.299	1.385	1.256	1.005	1.040	**0.956**	1.012	1.123	1.026	1.018	1.064

6B. Forecasting GDP deflator inflation

		h = 0				h = 1				h = 4			
					Ratio AR(4)_RTV				Ratio AR(4)_RTV				Ratio AR(4)_RTV
Model:	Sample starts:	y_{T+1}^{T+2}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+1}^{09Q1}	y_{T+1}^{T+2}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+1}^{09Q1}	y_{T+4}^{T+4+1}	y_{T+4}^{T+19}	y_{T+4}^{09Q1}	y_{T+4}^{09Q1}
		With industrial production											
ADL (4,2)	1965:Q3	**0.970**	0.980	1.007	0.967*	**0.972**	**0.969**	**0.978**	0.970*	1.064	1.094	1.094	1.074
KK(q=5, p=(4,1))		1.099	1.023	1.059	1.016	1.083	0.994	1.030	1.022	1.078	1.079	1.100	1.080
ADL (4,2)	1979:Q1	**0.940**	**0.911**	**0.932**	0.997*	**0.935**	**0.906**	**0.920**	0.996*	0.984	0.999	**0.977**	1.036
KK(q=5, p=(4,1))		1.963	1.281	1.406	1.496	1.096	**0.948**	**0.969**	1.043	0.982	**0.931**	**0.927**	1.002
		With employment											
ADL (4,2)	1965:Q3	**0.968**	**0.971**	0.981	0.953*	**0.972**	**0.974**	**0.977**	0.978*	1.047	1.108	1.097	1.154
KK(q=5, p=(4,1))		1.025	**0.960**	0.996	0.968*	1.048	0.986	0.999	1.005	1.049	0.994	1.034	1.088
ADL (4,2)	1979:Q1	**0.939**	**0.896**	**0.920**	0.991*	**0.939**	**0.906**	**0.916**	0.991*	0.981	0.981	**0.964**	1.057
KK(q=5, p=(4,1))		1.321	1.042	1.018	1.110	2.015	1.138	1.231	1.342	1.106	1.048	1.119	1.222

6C. Forecasting PCE deflator inflation

Model:	Sample starts:	h = 0			Ratio AR(4)_RTV	h = 1			Ratio AR(4)_RTV	h = 4			Ratio AR(4)_RTV
		y_{T+1}^{T+2}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+1}^{09Q1}	y_{T+1}^{T+2}	y_{T+1}^{T+15}	y_{T+1}^{09Q1}	y_{T+1}^{09Q1}	y_{T+4}^{T+4+1}	y_{T+4}^{T+19}	y_{T+4}^{09Q1}	y_{T+4}^{09Q1}
		With industrial production											
ADL (4,2)	1965:Q3	0.985	0.994	1.004	0.951*	**0.974**	0.988	0.998	0.955*	1.028	1.062	1.070	1.034
KK(q=5, p=(4,1))		1.058	1.056	1.074	1.017	1.032	1.031	1.043	1.016	1.042	1.042	1.060	1.053
ADL (4,2)	1979:Q1	**0.977**	0.983	0.981	0.986*	**0.972**	0.981	**0.969**	0.993*	**0.978**	**0.979**	**0.968**	1.029
KK(q=5, p=(4,1))		1.414	1.071	1.086	1.091	**0.971**	**0.977**	0.982	1.006	**0.938**	**0.934**	**0.939**	0.997*
		With employment											
ADL (4,2)	1965:Q3	**0.978**	0.980	0.982	0.934*	**0.970**	**0.978**	**0.975**	0.972*	1.028	1.036	1.039	0.980*
KK(q=5, p=(4,1))		0.996	0.989	1.001	0.952*	1.003	0.998	1.008	1.001	1.046	1.039	1.058	0.998*
ADL (4,2)	1979:Q1	0.997	0.997	1.005	0.999	0.981	0.988	**0.978**	1.006	**0.975**	0.987	**0.970**	1.007
KK(q=5, p=(4,1))		**0.970**	**0.953**	**0.947**	0.941*	1.064	1.002	1.002	1.046	**0.944**	**0.956**	**0.953**	0.998*

Note: Entries with values smaller than one indicate that either the ADL model estimated with RTV data or the KK model reduces the RMSFE relative to the ADL model estimated with EOS data, except entries in the column "Ratio AR(p)_RTV" where values smaller than one indicate that the model reduces the RMSFE relative to the AR model estimated with RTV data and bias-corrected to predict post-revision values. ADL models are estimated for each forecast horizon (specification as described in the first column) and forecasts are computed `directly' with increasing windows of data. The variable we aim to forecast is the quarterly difference at an annualized rate *h*-quarters ahead, namely, $y_{t+h} = 400[\ln(Z_{t+h}) - \ln(Z_{t-h-1})]$, where *Z* is the level of real GDP, the GDP deflator or the PCE deflator. The explanatory variables are also annualized quarterly differences. We use as our quarterly vintage of industrial production and employment the vintage published in the first month of each quarter, and calculate a quarterly series by using the observation from the last month of the quarter, before computing first differences. When estimating with RTV data, the vintage of the RHS variables depends on the forecasting horizon as described in the text. Forecasts of post-revision data computed with RTV data are bias corrected. KK (q=5,p=(4,1)) means that the bivariate KK model uses five estimates of both the target variable and the indicator, such that y_t^{t+5} and x_t^{t+5} are efficient estimates of the true value; four is the number of lags of *y* and one is number of lags of *x* in the VAR with true values (see Appendix B for details). A * denotes forecasts are more accurate than those of the AR model. Emboldened entries indicate an increase in accuracy of more than 2% from using RTV data (or the KK model) relative to using EOS data.