Research Program on Forecasting

# Forecasting Data Vintages

**Tara M. Sinclair**
The George Washington University
Washington, DC 20052
USA

RPF Working Paper No. 2012-001
http://www.gwu.edu/~forcpgm/2012-001.pdf

January 18, 2012

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

# Forecasting Data Vintages

Tara M. Sinclair
George Washington University
Washington, DC

January 18, 2012

JEL: C53

**A revised version paper is forthcoming in the International Journal of Forecasting. It is a comment on Clements and Galvão's "Forecasting with Vector Autoregressive Models of Data Vintages: US output growth and inflation" which is forthcoming in the same issue.**

Keywords: Real time data, Evaluating forecasts, Forecasting practice, Time series, Econometric models

## Abstract

This article provides a discussion of Clements and Galvão's "Forecasting with Vector Autoregressive Models of Data Vintages: US output growth and inflation." Clements and Galvão argue that a multiple-vintage VAR model can be useful for forecasting data that are subject to revisions. Clements and Galvão draw a "distinction between forecasting future observations and revisions to past data," which brings yet another real time data issue to the attention of forecasters. This comment discusses the importance of taking data revisions into consideration and compares the multiple-vintage VAR approach of Clements and Galvão to a state-space approach.

# Forecasting Data Vintages

## Introduction

Over the last several years, a growing body of research has made economic forecasters aware of the need to take into consideration the role of data revisions in their forecasts. The real time data movement has in particular emphasized the importance of evaluating forecasts both in terms of the information forecasters had at the time they made their forecasts (rather than based on later revised data) and in terms of the data the forecasters were aiming to predict (see Croushore, 2011). Clements and Galvão's "Forecasting with Vector Autoregressive Models of Data Vintages: US output growth and inflation" (this issue), addresses yet another concern: forecasters are typically estimating their models using data that are at different points in the revision process. If there are systematic patterns in the revisions, then there are even more issues in choosing which data to use to evaluate both forecasts and early estimates, as well as which data to choose for estimating forecasting models.

Consider a simple benchmark autoregressive model that a forecaster might use to forecast an economic variable $y_t$. In order to estimate the parameters to be used for the forecast, typically the forecaster will obtain the most recently updated data on $y_t$ (i.e. the *vintage* of $y_t$ available at that time) and estimate the model using those data. However, the data in this single time series may in fact be coming from different data generating processes. The data some time back in the series have gone through monthly revisions, annual revisions, and perhaps several benchmark revisions. The most recent data, however, have been only "lightly revised," as Clements and Galvão term it. Therefore, Clements and Galvão argue that the data in a single vintage are of "different maturities." Forecasters may want to forecast future revisions to data as well as exploit any forecastability of data revisions to improve forecasts of future observations.

In their article, Clements and Galvão suggest that a multiple-vintage vector autoregressive model (VAR) is a useful approach for forecasters working with data subject to revisions. This comment discusses the importance of taking revisions into consideration and compares the multiple-vintage VAR approach of Clements and Galvão to a state-space approach.

**The Data Generating Process of Data Subject to Revisions**

A key issue in thinking about data subject to revisions is to decide which data share the same data generating process (DGP). One option would be to assume that data of the same vintage, i.e. the data provided by the statistical agency as a single time series at a particular moment in time, would share a common DGP. This is most often what forecasters take as their dataset for both the left-hand and right-hand sides of their forecasting equations. But, these series are actually composed of data with different amounts of revision – from the recent data which may be not revised at all to data from further back in time which may have undergone substantial revision. Kishor and Koenig (2011) point out that these variations in maturity are particularly a problem for forecasting. Forecasters using a single vintage of data will estimate the parameters of their models based mostly on heavily revised data. The forecasts made from the model, however, depend primarily on recent, more lightly revised data. If data with different numbers of revisions are from different data generating processes, then the forecasts made from these models may not perform well.

Another option would be to group together data that have experienced the same amount of revision. We could, for example, gather all first releases as a single data series.[1] However, historical data in this time series may have been gathered based on different definitions.

[1] These data are often called the "diagonals" due to where they are located in a real time data set. Typically in these datasets each column contains the historical data available at a particular point in time. From these columns, a time series that contains data with the same amount of revision would be on the diagonal.

Furthermore, the revisions may help forecast future data. If the revisions are unpredictable and the early data are efficient estimates of future data, then we may not need to be concerned about the different vintages. For example, it may be appropriate, and more accurate in finite samples, in this case to use first release data as the left-hand-side variable in a forecasting model (Koenig et al, 2003). There is mounting evidence, however, that data revisions are in fact often predictable (Croushore, 2011). Furthermore, even if revisions themselves are unpredictable, it may still be possible for the revisions to serve as predictors of future observations.

**Forecasting Models for Data Subject to Revisions**

Given modern computer capacity, it now makes sense to keep track of data revisions so that we can include all the data in the model and keep track of the amount of revision. One way to do this is the multiple-vintage VAR approach advocated by Clements and Galvão. This model is a generalization of a single-vintage autoregressive (AR) model to include multiple vintages. Thus a single variable is included in the VAR but the vector aspect captures different vintages of the data. The model Clements and Galvão use can also allow for seasonal and benchmark revisions. The multiple-vintage VAR directly produces forecasts of future vintage values of the variable of interest. This future vintage represents revised values of past data as well as new observations. Thus the forecasts produced are an intuitive collection of forecasts of data revisions and forecasts of new observations.

An alternative model, however, would be a state space model, for example the one proposed in Kishor and Koenig (2011). State space models using the Kalman filter are an obvious choice when the "truth" needs to be filtered out of the data. In the model proposed by Kishor and Koenig, the forecasts can be made by first estimating a VAR using only data that has been "fully" revised. Then the parameters of that VAR can be substituted into a state space

model and the Kalman filter can be applied to obtain an estimate of the current state vector and forecasts for the future. In this way they first convert the end-of-sample, not yet fully revised data into fully revised data before substituting them into the model. This approach keeps the data generating processes consistent. This approach also easily allows for other variables to be included, so that they can jointly forecast GDP growth and the output-consumption ratio, for example.

Both the VAR and the state space assume that the releases are efficient with some finite lag to make them feasible to estimate. Both approaches also generally require that the process of the data or of the revisions to the data be characterized by an AR (or moving average) process (Croushore, 2006). Finally, both approaches can produce forecasts of future vintages of data, although the multiple-vintage VAR approach produces these directly whereas the state space approach requires that state vector forecasts be substituted back into the model to produce the forecasts of data with different amounts of revision.

The key question is whether the data revision process can be modeled based only on observables (as in the VAR) or if it is better modeled with unobserved components (as in state space models). The state space models are focused on forecasting data with a constant amount of revision (most often the fully-revised data). They address the combination of different data generating processes by converting data into a common one through the Kalman filter. On the other hand, the multiple-vintage VAR approach produces forecasts of vintages of data taking advantage of the potential forecastability of data revisions. Each forecasted vintage consists of forecasts of revisions to past observations as well as forecasts of new observations. This intuitive approach may particularly be of interest to forecasters who may want to predict what a

future vintage of data may look like. Policymakers may also be interested in the predicted data revisions as well as the predictions of future observations.

In the end, selecting between forecasting models comes down to forecasting performance. Clements and Galvão have shown that their method appears to perform well, particularly for inflation, most likely because inflation has predictable revisions based on observables. A comparison of the forecasting ability across the different approaches would be useful, although the results most likely depend on both the particular data being forecasted as well as the objectives of the forecaster.

**Conclusion**

Clements and Galvão's article contributes to the literature on the impact of data revisions on forecasting. Forecasters have become aware of the need to keep track of the vintage of data they have available at the time they make their forecasts. The issue of different maturities of data in each vintage, however, is less well known. Clements and Galvão's article is an excellent addition to the literature drawing attention to this important subject.

One drawback of Clements and Galvão's article is that it focuses on models of a single variable. Next steps in this literature, as Clements and Galvão note in their introduction, should include extending the model to incorporate additional explanatory variables that often play a role in more complex forecasting models. In those cases, data revisions may arrive at different times for the different variables. Knowledge of the revision process for the explanatory variables may, in some cases, improve the forecasts of the target variable. Then a full comparison with the Kishor and Koenig (2011) state space approach would also be possible. Alternatively, as suggested by Croushore (2006), a factor model approach might avoid the problems of data

revisions. It might, however, also miss out on both the usefulness of data revisions for forecasting and on providing forecasts of the data revisions that may be useful for policymakers.

**Acknowledgements**

The author thanks Michael Clements, Dean Croushore, Fred Joutz, Roberto Samaniego, H. O. Stekler, Simon Van Norden, and the participants in the 7th Workshop of the International Institute of Forecasters in Verbier, Switzerland, for helpful discussions and comments.

**References**

Clements, M. P. and A. B. Galvão (this issue). Forecasting with Vector Autoregressive Models of Data Vintages: US output growth and inflation. *International Journal of Forecasting.*

Croushore, D. (2006). Forecasting with real-time macroeconomic data. In Elliott, G., Granger, C., and Timmermann, A. (eds.), *Handbook of Economic Forecasting*, Volume 1. Handbook of Economics 24, pp. 961.982: Elsevier, North-Holland.

Croushore, D. (2011). Forecasting with real-time data vintages. In: M. P. Clements and D. H. Hendry (Eds.) *The Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press.

Koenig, E. F., S. Dolmas, and J. Piger (2003). The Use and Abuse of Real-Time Data in Economic Forecasting. *The Review of Economics and Statistics*, 85(3): 618-628.

Kishor, N. K., and Koenig, E. F. (2011). VAR estimation and forecasting when data are subject to revision. *Journal of Business and Economic Statistics*. Forthcoming.