Forecast Evaluation of AveAve Forecasts in the Global VAR Context[1]

Tara M. Sinclair
Department of Economics
George Washington University
Washington DC 20052
tsinc@gwu.edu

H. O. Stekler
Department of Economics
George Washington University
Washington DC 20052
hstekler@gwu.edu

The article "Forecasting Economic and Financial Variables with Global VARs" is on the frontier of current model building based on vector autoregressions. The constructed model links country/region specific models into a unified framework. This global VAR (GVAR) consists of 26 regions and estimates equations for 134 variables. The models for most regions consist of both real and financial endogenous variables. It is thus possible to explore the dynamics and interrelationships of both the real and the financial sectors. The authors explain the issues involved in building a model, evaluating the model and doing the final testing. This paper builds on research by Pesaran, Schuermann, and Weiner (2004) and Dees, di Mauro, Pesaran and Smith (2007) by exploring the forecasting ability of the GVAR.

The paper ties together a large number of diverse literatures. Consider the topics that are covered: VARs, model averaging and combinations, time series benchmarks, cointegration, structural breaks, estimation periods, etc. The model itself takes into account co-trending restrictions, cross country cointegration and trade relations, etc. The paper also develops a panel version of the Diebold-Mariano predictive test statistic. The breath of the topics covered makes

[1] This article was written as an invited commentary and is forthcoming in the *International Journal of Forecasting*.

this paper required reading for any forecaster who uses these types of models. In this comment we will concentrate on the evaluation procedures in section 6 of the paper.

I. Evaluation Procedures

There are 19 different GVAR model versions. Each of these models has been estimated over 10 sample periods, yielding 190 total forecasts. In order to arrive at a single final forecast, the authors average both across models and across sampling windows, developing the average-average (AveAve) forecast. The forecast evaluation is conducted on data for 2004 and 2005, which are the years available beyond the period of fit. In order to have sufficient observations, the models' forecasts of a particular variable are pooled across regions. Comparisons are made against four specific benchmark forecasting models: (1) random walk, (2) random walk plus drift, (3) first order autoregressive and (4) first order autoregressive with drift. The metric for these comparisons is the root mean squared forecast error (RMSFE) and the tests of significance use a panel version of the Diebold-Mariano statistic developed in this paper. The new test is applied only to the one-quarter-ahead forecasts due to a lack of data availability for the 4-quarter ahead horizon.

The forecast evaluation is primarily concerned with three questions: (1) How well does forecast averaging perform? (2) Does averaging across all models and estimation windows (AveAve) beat the benchmarks? and (3) Are the financial variables useful in forecasting real output and inflation?

II. Results

A. Averaged forecasts

GVAR-AveAve has a smaller RMSFE than the underlying individual models a large percentage of the times for all variables (with the exception of oil), for most countries and groupings. The performance of the AveAve is superior to the underlying models for the one-quarter ahead forecasts more than for the four-quarter ahead estimates. While this metric is frequently used, what is the significance of any given percentage?

As an example, suppose that the AveAve is superior to all of the models, i.e. 100% of the time. Then the AveAve RMSFE is smaller than the smallest RMSFE of all the individual models and this result indicates that the errors of the underlying models offset each other. The results are easy to interpret in this case. If, however, we observe other percentages, how should we interpret those results? In order to investigate this issue, we will generate a number of simulations.

These simulations are based on assumptions about the bias of the 190 models.[2] In each simulation we assumed that the forecast errors of the underlying models were independent of each other and were normally distributed with varying degrees of bias. Each simulation was based on 100 trials, each consisting of 190 observations that represent the errors of the underlying models. We realize that these simple distributions abstract from the correlations across forecast errors that result from the same model being estimated for 10 different windows.

Nevertheless, we obtain interesting results about the relationship between the percentage of time that the AveAve beats the underlying models and the degree of bias that we have assumed. This yields some information about the bias/forecasting characteristics of the underlying models. Table 1 shows that if <u>all</u> of the underlying models are unbiased, N(0,1), then AveAve will have a lower RMSFE than all of the underlying models. Similar results obtain if the underlying models are biased, but the bias is offset between the different models, i.e. ½: N(μ,1), ½: N(-μ,1), where μ can be any mean. In this case we cannot determine whether the superiority of AveAve is due to the underlying models all being unbiased or if it is the result of offsetting biases. The remainder of Table 1 shows that as the average bias of the underlying

[2] The authors did not report whether the forecasts of the underlying models were biased or unbiased.

models increases, the superiority of AveAve relative to the underlying models declines monotonically.

Table 1 Simulation Results (100 per experiment)	
Distribution from which the underlying "forecast errors" were drawn[3]	**Average % of the 190 models that the AveAve beats**
N(0,1)	100.00%
½: N(μ,1) ½: N(-μ,1)	100.00%
N(0.5,1)	99.14%
N(1,1)	89.98%
N(2,1)	73.25%
N(3,1)	65.84%
N(4,1)	62.13%
N(5,1)	59.68%
N(10,1)	54.74%
N(100,1)	50.55%
½: N(0,1) ½: N(-10,1)	50.00%

B. Benchmarks

For output growth, none of the simple benchmarks are statistically superior to AveAve. The averaged output growth forecasts beat the benchmarks 50-75% of the time depending on the regions considered. The results are similar for inflation, except when the Latin America region is included in generating the inflation forecasts. In this case, the AveAve forecasts for inflation are significantly beaten by the two random walk benchmarks. This result is attributable to the high inflation that occurred in that region. When the forecasts of the financial variables are evaluated,

[3] The results are similar for N(-μ,1) for all cases.

the model average does not perform as well against the benchmarks as it did in predicting output and inflation.

The benchmarks considered in this paper are much easier to construct than the GVAR and are often hard to beat. One question that arises is whether the gain from the improved forecasts is worth the cost in terms of increased complexity. We therefore suggest an additional benchmark to consider as an alternative: market forecasts. As an example, consider the mean forecast of the Survey of Professional Forecasters (SPF) for the US. These data come from the Federal Reserve Bank of Philadelphia (for a description of this survey, see Croushore, 1993).

The SPF and AveAve forecasts are not reported on the same basis. The former are reported as quarterly changes at annual rates, while the latter are not annualized. It was, therefore, necessary to convert the SPF forecasts to be comparable to those in the article. For GDP we constructed the comparison forecast errors by taking the natural log of the one quarter ahead SPF forecast (forecasts made from 2003Q4-2005Q3) for real GDP minus the natural log of the SPF forecast for the current quarter. We constructed the actual values as the log changes in current data (as of August, 2008). For CPI inflation, the SPF reports only annualized inflation rates in percentage points, so we constructed the forecast errors as the difference in the actual and forecasted annualized inflation rates divided by 400. We then used the formula from section 5.3 of the paper where h=1 to construct the RMSFE.

We compare the results from Figures 1a and 2a (values provided by the article's authors) for the US with the mean SPF forecasts for the same period and present the results in Table 2 below. Our findings indicate that the AveAve errors are smaller than those of the SPF for US real GDP, but not for US CPI inflation. These results are merely suggestive because the SPF forecasts are truly ex ante. The individual forecasters in the SPF survey make their forecasts in the middle of each quarter, when only the first estimates for the previous quarter were available. The AveAve forecasts, however, are made using the latest available data. The comparison does not take into account any issues involved in using real-time data versus revised data as in this analysis.[4]

Table 2 Comparison of the US AveAve RMSFE with the SPF RMSFE		
	US Real GDP	**US CPI Inflation**
AveAve RMSFE	0.173	0.516
SPF RMSFE	0.273	0.473

C. Include financial variables?

Given that the forecasts of the model's financial variables were not significantly better than those of the benchmarks, the authors next determined the effect that excluding two of the financial variables would have on the accuracy of the output, inflation, and short run interest rate estimates. They showed that including the equity and long run interest variables does not appear to increase the accuracy of the predictions of the other variables.

As the authors mention in the conclusion, it would also be useful to perform a similar exercise removing the foreign variables from the country-specific equations to evaluate the contribution of the "global" aspect of the VAR.

[4] Along these lines, the authors note at the end of Section 5.3 that for "one-quarter ahead forecasts no adjustment for serial correlation is needed." That is true if the data for the current quarter are known. In real time, however, the current quarter data are also forecasts, because not all the data for that quarter are yet available.

III. Conclusion

The evaluation in the article is state of the art. It demonstrates that GVARs have the potential for generating output and inflation forecasts that are generally superior to those produced by simple benchmarks. Nevertheless, there are some topics that this evaluation has not considered. For example, are the forecasts biased? In practice, how do data revisions affect the estimates? How much, if at all, has accuracy improved by constructing a GVAR as compared to doing VARs for each region? Answers to these questions would show whether GVARs enhance our ability to predict macroeconomic variables.

References

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Business Review,* November/December, 3-13.

Dees, S., di Mauro, F., Pesaran, M.H. & Smith, L.V. (2007). Exploring the international linkages of the euro area: A global VAR analysis. *Journal of Applied Econometrics*, 22, 1-38.

Pesaran M.H., Schuermann, T. & Weiner, S.M. (2004). Modeling regional interdependencies using a global error correcting macroeconometric model. *Journal of Business and Economic Statistics*, 22, 129-162.