

THE ROLE OF MODELS AND PROBABILITIES IN THE MONETARY POLICY PROCESS

CHRISTOPHER A. SIMS

ABSTRACT. The primary models in use as part of the policy process in central banks are deeply flawed, both from the point of view of econometric analysis and from the point of view of economic theory. "Subjective" approaches to forecasting play a major role in policy formation in every central bank, and data on the forecasting record of FRB non-model forecasts shows that they are excellent forecasts by several measures. Academic research on econometric method and on macroeconomic theory has not provided much guidance for model builders who need to contribute to policy analysis in real time. Policy discussion at central banks uses the language of Bayesian decision theory — putting post-sample probabilities on models, generating probability distributions for future values of variables that reflect uncertainty about parameter values and subjective judgment, weighing expected losses of alternative courses of action. But the standard toolkit of econometrics does not connect to this way of thinking about probability. There is some reason to hope for improvement before long.

I. INTRODUCTION

This is an essay on the way data relates to decision making in Central Banks.

One component of it is based on a series of interviews with staff members and a few policy committee members of four central banks: The Swedish Riksbank, the ECB, the Bank of England, and the US Federal Reserve Board. These interviews focussed on the policy process and sought to determine how forecasts were made, how uncertainty was characterized and handled, and what role formal models played in the process.

In each of the banks "subjective" forecasting, based on data analysis by sectoral "experts", plays an important role in the process. At the US Federal Reserve Board, there is

Date: October 26, 2002.

©2002 by Christopher A. Sims. Prepared for the Fall 2002 meeting of the *Brookings Panel on Economic Activity*. This research was supported in part by the Princeton Center for Economic Policy Studies.

a 17-year record of model-based forecasts that can be compared with a longer record of subjective forecasts, and a second component of this paper is an analysis of these records.

Two of these central banks (the Riksbank and the Bank of England) have explicit inflation-targeting policies that require them several times a year to publish their forecasts of inflation, for which they set quantitative targets. A third component of the paper discusses the effects of such a policy regime on the policy process and on the role of models within it.

The large models in use in central banks have lost any connection to the simultaneous equations statistical theory that was thought of as the intellectual foundation of their predecessors. The models are now fit to data by ad hoc procedures that have no grounding in statistical theory. A fourth component of the paper discusses how the inference in these models reached this state and why academic econometrics has had so little impact in correcting it.

Despite not providing better forecasts, not having a firm statistical foundation, and having weaknesses in their underlying economic theory, the large models play an important role in the policy process. A final component of the paper discusses what this role is and how the model's performance in it might be improved.

II. THE POLICY PROCESS

At all four banks the policy process runs in a regular cycle, quarterly at all except the Fed, where the cycle is keyed to the FOMC meetings, roughly every 6 weeks. Each bank has a primary model, even though each also has other models. The primary models are the ones used to construct projections of alternative scenarios, conditional on various assumptions about future disturbances or policies, or on various assumptions about the current state. Where there is feedback between models and subjective forecasts, it is generally through the primary model.

The primary models have some strong similarities. There are about 15 behavioral equations in the ECB model,¹ 21 in the Bank of England model,² 27 in the Riksbank model,³ and

¹The ECB model equations are laid out in Fagan, Henry, and Mestre (2001)

²The Bank of England primary model equations are laid out in Quinn (2000)

about 40 in the FRB/US model.⁴ Each has at least some expectational components, with the FRB/US and Riksbank models the most complete in this respect. The banks whose models are less forward-looking describe them somewhat apologetically, suggesting that they are working on including more forward-looking behavior.

The Riksbank and the Bank of England have publicly described “suites” of models of various types, including VAR models, smaller macro models, and optimizing models. Some of these models produce regular forecasts that are seen by those involved in the policy process, but none except the primary model have regular well-defined roles in the process. The other banks also have secondary models with some informal impact on the policy process.

Each policy round proceeds through a number of meetings, through which a forecast is arrived at iteratively, but the number of meetings and the way they order discussions varies. At the Riksbank there is a startup meeting at which forecasts from two large models are presented, followed by another meeting at which the sectoral experts (14 — nearly everyone in the monetary policy staff) present their views and relate them to the model. At a third meeting the staff’s report is put together. Following that meeting, a three to five person editorial committee rewrites the report into a form suitable for issue as a policy statement by the board. At this stage and also earlier, there is some feedback from the policy board, attempting to avoid sharp divergences between policy board views and staff views.

At the Bank of England each policy round involves 6-7 meetings — this after a recent reduction in the number of meetings — and some policy board members attend the meetings from the earliest stages. This may reflect the unusually high proportion of graduate trained economists on the Bank of England policy board (the MPC). All of the discussion

³The Riksbank model is said to be nearly identical to the QPM model of the Bank of Canada, which is described in Poloz, Rose, and Tetlow (1994), Black, Laxton, Rose, and Tetlow (1994) and Black, Laxton, Rose, and Tetlow (1994)

⁴The FRB/US model is described in an html web that provides linked equation descriptions and a set of explanatory discussion papers. This material was made available to me for the research underlying this paper, and is available from the Federal Reserve Board to other researchers on request. Requests for it should be directed to David Reifschneider, dreifschneider@frb.gov.

of projections and policy choices occurs within the framework of the primary model MM. When a section of the model is overridden, that is done via residual adjustments, so large residuals become a check on such model revisions.

At the ECB the process is limited primarily to staff until late stages. It begins with collection of projections and assessments of the current conditions from the national central banks. But as at other banks a major role is played by sectoral experts, and the primary model (the AWM) is used to generate residuals corresponding to the expert forecasts. Twice a year a more elaborate process is undertaken, in which national central bank staff is represented on the forecast committee and the iterations go back and forth to the country banks as well as between sectoral experts.

At the Board of Governors of the Fed, the Green Book process begins with a small group (around four people) meeting to set the forecast “top line”— values for GDP growth and for key financial variables, including the Federal Funds rate. The next stage is generation of forecasts for their variables by the sectoral experts. The expert forecasts are fed through the FRBUS model to generate residuals, and the results of this exercise are considered at a subsequent meeting. Feedback can occur in both directions between model forecasts and subjective forecasts, as well as back to the top line numbers.

The Fed staff emphasized to me (though this may be true at the other banks as well) that the expert subjective forecasters have econometric input well beyond that in the primary model residuals. The sectoral experts generally have one or more small econometric models of their own sectors, and these often are more sophisticated than corresponding equations in the FRBUS model. The Fed has an explicit policy of maintaining the forecast as a pure staff forecast, not allowing any policy board participation in the meetings that go into forecast preparation.

Each of the banks prepares more than just a single forecast. The Fed probably does the most along these lines, with the Green Book recently showing as many as a dozen potential time paths for the economy, under varying assumptions. The Fed staff see these scenarios as a concrete way to give a picture of what uncertainty there may be about their forecast, despite the absence (most, but not all, of the time) of stochastic simulations in their analysis. The Bank of England regularly publishes fan charts of their central forecasts, a

forecast conditioned on a main minority view or views in the FOMC meeting, and a forecast conditioned on the time path for interest rates implicit in the futures market. For the MPC, but not for publication, the Bank staff prepare projections also conditioned on higher and lower values of the policy rate. This gives the MPC a well defined projection, should they decide to set rates higher or lower than the the central forecast.

All the banks discussed here except the US Fed condition their forecasts on a constant interest rate assumption. This is a source of serious analytical difficulty for the Riksbank modelers, because QPM was built around an assumed policy reaction function. If the interest rate is truly left constant, the model explodes. If it is left constant for one or two years, then modeled with the reaction function thereafter, it jumps at the transition date and causes strange behavior. To avoid these problems the Riksbank simply uses the time path of long interest rates that is generated from a model run with the reaction function in place, even though the short rate is set on the constant-rate path. The inflation-targeting banks are no doubt concerned that a non-flat time path for interest rates, if published, might be given too much weight by the markets and might be seen as a commitment by the central bank. But for the ECB, which does not publish its staff forecasts, the constant interest rate assumption is a response to the wishes of the policy board.

III. FED FORECASTS

How well does the Federal Reserve Board Staff forecast? The conclusion, which largely matches that of Romer and Romer (2000), is that the Fed forecasts quite well indeed, especially for inflation. This section goes beyond Romer and Romer by

- extending their sample, which went through 1991, to 1995 or 1996,
- considering data on the Fed's internal model-based forecasts as well as data on their "Green Book" forecasts,
- applying some analytical methods that may give additional insight into the nature of the Fed forecasting advantage, and
- speculating on the implications of these results, in part based on the interviews, along lines that only partially match the Romers' discussion.

III.1. **The Data.** Before each meeting of the Federal Reserve Open Market Committee, the staff prepares a forecast that is presented in the “Green Book”. This forecast is labeled “judgmental”. It included in September 1995, for example, forecasts for 53 variables, though the list of variables included has fluctuated in length, running over 80 in the early 1980’s. The forecasts include estimates of the current quarter numbers and forecasts of future quarters, with the time span of the forecasts varying, in 1979-95, from four to nine quarters into the future. There is also a forecast labeled “model-based” prepared at the same time. Until 1995, these forecasts were based on the MPS model, an economy-wide model originally developed as an academic collaboration, but maintained afterwards by the Federal Reserve Board staff. After 1995, the model used for these forecasts has been the new FRBUS model, created within the Fed. These model forecasts are archived in machine-readable form and were made available to me for this study. Their public use is, as I understand it, restricted only by the same 5-year disclosure rule that governs the Green Book forecasts. The data for the MPS model forecasts that I have used will be posted on the web so that other researchers can access them, though there are apparently no immediate plans for the Board to regularly make available updated data sets as the 5-year horizon advances.

This study also considers the forecasts from the Survey of Professional Forecasters (SPF), which was begun in 1968 as a project of the American Statistical Association and the NBER and taken over in 1990 by the Federal Reserve Bank of Philadelphia. Data from this survey are available at the Philadelphia Fed web site.

Because some of the analyses in this section are greatly simplified by having data at a uniform time interval, all the data have been converted to quarterly form. The SPF is quarterly to start with. FOMC meetings occur at least once each quarter, but with non-uniform timing within the quarter. Dean Croushore of the Philadelphia Fed has created and published on the Philadelphia Fed web site a quarterly series of Green Book forecasts, constructed by taking the FOMC meeting date closest to the middle of each quarter. Those data are used in this study. The MPS forecasts have been put in quarterly form by matching their dates to the Croushore quarterly FOMC dates.⁵

The “actual” values used to construct forecast errors in this study are real gdp growth and gdp deflator inflation as it appears in the most recently available chain-weighted data. The Romers instead used the second revision, which appears with about a one-quarter delay. Revisions are often substantial, as are the differences between chain-weighted and fixed-weight series. There is an argument for targeting a near-term revision as “actual”, as the Romers did. The interest in the forecasts and their influence on decisions is highest in the months immediately surrounding the forecasts, so errors as perceived then are probably closest to what enters the forecasters’ own loss functions. It also seems unfair to penalize forecasters for “errors” that arise because an “actual” series is a different accounting concept than the series the forecasters were in fact projecting. On the other hand, the Romers have already considered actual values defined this way, and there is insight to be gained from a different approach.

The most recent revisions should, after all, be the best estimates of the actual historical path of the economy. Arguably we should not penalize a forecaster for failing to forecast a recession that disappears in later revised data, or for anticipating growth that actually occurred, but was not recognized until data were revised a year or two later. The chain-weighted data, though not available at the time most of the forecasts we consider were made, has claims to be more accurate than the fixed-weight data available for most of the historical period we study. On these grounds, then, it is worth knowing whether analysis of forecasting performance is sensitive to whether we measure “actual” outcomes as second revisions or as the latest revisions of the most recently developed accounting concepts. That this study finds results very similar to those of the Romers supports the comforting conclusion that sustained patterns of forecast accuracy or inaccuracy are not sensitive to the details of data definitions.

⁵The Romers chose to convert their data to monthly form instead, and thereby ended up with data sets with non-uniform timing. For their regression analyses this created no great analytical difficulty, and it let them preserve more of the information in the original data set. This paper’s VAR-based analysis of the marginal contribution of Green Book forecasts in the presence of other variables would be made more complicated by non-uniform time intervals in the data.

III.2. Characterizing inflation forecast accuracy. Table 1 shows the root mean square errors of four inflation forecasts over the period (forecasts made in 1979-95) for which all four are available. The “naive” forecasts are no-change forecasts. As can be seen from the first column, though, the forecasts made in real time have substantial error even in determining current quarter inflation, for which data is available only with a delay. The naive forecasts are therefore not naive at all for the current quarter, and are probably an unrealistic standard even one quarter ahead because of the information advantage they reflect. At two or more quarters ahead, all three of the real-time forecasts are better than naive forecasts. The best forecast, uniformly for all horizons 1 through 4, is the Green Book forecast. On the other hand, the differences do not seem large, especially between the MPS model and the Green Book forecasts.

The similarity of the forecasts is also apparent in the correlation matrices shown in tables 2 and 3. The inflation forecasts are highly correlated, and more strongly correlated among themselves than they are with the actual data. We can see the same point in Figure 1, which shows the forecasts and actual data tracking together closely. A similar plot for 1-step ahead forecasts would be even more tightly clustered.

On the other hand, when we follow Romer and Romer in regressing actual inflation on the forecasts, we see from Tables 4 and 5 that we obtain a result similar to theirs: the coefficients on the Green Book forecasts are large and significant, even larger than one at the one year horizon, while those on the other forecasts are insignificant or even negative.

The Romers refer to this sort of regression as measuring the “information content” of forecasts, following Shiller and Fair (1989), who were probably the first to use this language to characterize this sort of regression. While the regression is useful information, if interpreted carefully, it is probably misleading to think of it as characterizing “information content”. Clearly these inflation forecasts in some sense have very nearly the same “content”, since they are so highly correlated.

Consider two different models of how forecasts might be related to each other and to actual outcomes. Let f be the vector of forecasts and y be the outcome. One possible model is

$$y_t = \gamma f_t + \varepsilon_t, \tag{1}$$

with the elements of the f_t vector independent of each other and of ε_t . Then the coefficients in the γ vector, squared, would be direct measures of accuracy of the f_t 's, and they would be estimated properly by a least squares regression.

Another extreme possibility, though, is that all forecasters have noisy observations on a single “forecastable component” of y , which they may or may not use optimally. Then if we let f^* denote the forecastable component of y , we have the model

$$f(t) = \delta + \Lambda f^*(t) + \varepsilon(t) \tag{2}$$

$$y(t) = \phi + \theta f^*(t) + v(t) \tag{3}$$

$$\text{Var} \left(\begin{bmatrix} \varepsilon(t) \\ v(t) \end{bmatrix} \right) = \Omega, \tag{4}$$

with Ω diagonal and f^* orthogonal to ε and v .

In this framework, the quality of a forecast is related inversely to the variance σ_i^2 of its $\varepsilon_i(t)$ and to the deviation of its λ_i coefficient from θ . It can be shown that this model implies that estimated regression coefficients in the regression (1) will all be positive, and will be proportional to λ_i/σ_i^2 . If some forecasts have very small σ_i^2 values, the relative sizes of coefficients can be extreme, then, even though the forecasts have very similar forecast error variances. Note that the coefficients are not proportional to the forecast error variances, which include a perhaps dominant contribution from the variance of v ; the coefficients are inversely proportional to the relative idiosyncratic ε_i variances, even if these are an unimportant component of overall forecast error.

Interpretation of the regression coefficients becomes even more problematic if we admit the possibility of a second component of common variation, a “common error”. This can be allowed for by making f^* have two columns, with the entry in the last row of the second column set to zero. This makes the second column of f able to account for similar fluctuations in the forecasts that are unrelated to the actual outcome. When there is a common component of error, the regression coefficients in models like those of Tables 4 and 5 can be extreme, even though the common component of error is small and very similar among the forecasts. To see this, suppose the idiosyncratic component $\varepsilon(t)$ were negligibly small,

while

$$f_1(t) = f_1^*(t) + \lambda_1 f_2^*(t) \quad (5)$$

$$f_2(t) = f_1^*(t) + \lambda_2 f_2^*(t) \quad (6)$$

$$y(t) = f_1^*(t) + v(t). \quad (7)$$

Assuming the two components of f^* are uncorrelated, the coefficients of f_1 and f_2 in a regression like (1) or Table 4 will be $\lambda_2/(\lambda_2 - \lambda_1)$ and $-\lambda_1/(\lambda_2 - \lambda_1)$. Thus the coefficients will tend to be of opposite sign, with the difference between the coefficients *growing* as the forecasts become more similar ($\lambda_1 \rightarrow \lambda_2$).

It is not a good idea, then, to let our analysis of forecast quality be limited to an examination of this sort of regression. We need to go further in examining the correlation structure of the forecasts.

Despite its simplicity, the model of (2)-(3) provides a good approximation to the actual properties of the forecasts we are examining here. Estimates of it, for one and four quarter ahead forecasts, are shown in Tables 6 and 7. The coefficients on f in the tables have been normalized to make that in the “actual” equation equal to 1, and the constant terms have been converted to deviations from the constant term in the “actual” equation, so they become measures of forecast bias. The model attributes the low RMSE of Green Book forecasts entirely to their low idiosyncratic error. The naive and MPS forecasts both have lower bias at both short and long horizons, with the lower MPS bias particularly pronounced at the one-quarter horizon.

The model fits very well at the four-quarter horizon and fairly well at the one-quarter horizon. For the four-quarter forecasts, the standard likelihood ratio (LR) χ^2 test accepts the null at a marginal significance level of .743, if we ignore the serial correlation inherent in the overlapping forecasts. The serial correlation would almost certainly imply greater variability in the sample covariances, and hence even less evidence against the null. For the one step ahead forecasts, a standard LR test of the model against an unconstrained model rejects the null at a marginal significance level of .013, but the Schwarz criterion, which credits simpler models for having fewer parameters and provides consistent model selection, favors the restricted model. The forecasts have slightly fat-tailed distributions

(two or three residuals of more than 2.5 standard deviations, out of 68); if we accounted explicitly for non-normality, this would probably further weaken any evidence against the model. The tables show, instead of marginal significance levels, the sample size at which the conventional LR test would become “significant” at the .05 level and at which the SC would start to favor the larger model, assuming that the data moment matrices were held constant as we increased sample size. This perhaps provides a better index of how sensitive results might be to serial correlation and non-normality.

Since there is uncertainty both about the bias and about the variances of the forecasts, it is interesting to ask how much evidence there is against the hypothesis that the other forecasts have true RMSE's no greater than that of the Green Book forecasts. We can check this by fitting the bivariate mean and covariance matrix parameters for a pairing of the Green Book errors with another model's errors, with and without the constraint that for each forecast i , $\mu_i^2 + \sigma_{ii}$ is the same, where μ_i is the bias (the mean forecast error) and σ_{ii} is the forecast error variance. The result of such tests is shown in Table 8. The evidence for the superiority of the Green Book over SPF is strong, while that for the superiority of the Green Book over the MPS forecasts is weak.

The results here and in the Romer and Romer paper may appear to conflict sharply with the results reported in Atkeson and Ohanian (2001). They claim to find that econometric model forecasts are much worse than, and Green Book forecasts no better than, a simple naive model for forecasting inflation. However, their contrasting results arise entirely from their having restricted their sample period to 1984-99, a period when inflation was very stable. The naive model they consider forecasts average inflation over the next year as the average inflation rate over the preceding year. During the 1979-83 period, such a model, because it has a half-year lag during a period when inflation rose and fell rapidly, performs very badly, worse than a naive model that uses the previous quarter's inflation. They also measure actual outcomes as one-year average inflation rates rather than as one-quarter inflation rates. This is not the source of their contrasting results; I have verified that their type of naive forecast produces almost the same RMSE as the Green Book, when it is applied to forecast one-quarter inflation rates four quarters ahead over 1984-95. But it is substantially

worse, at both short and long horizons, than any of the other forecasts considered in this paper when applied to the entire 1979-95 period.

III.3. Characterizing GDP forecast accuracy. From Table 9 it is already clear that the advantage of the Green Book over other forecasts is much smaller for output growth than for inflation, and that the SPF forecasts look much better for this variable. The forecasts are also substantially less correlated, especially at the longer horizon, as shown in Tables 10 and 11. The Romer and Romer style regressions shown in Tables 13 and 12 still indicate a substantially larger coefficient on the green book than on other forecasts, but now the differences among coefficients, though large, are statistically insignificant. Table 14 shows that, despite its generally top ranked performance, for output growth the Green Book forecast has only a statistically negligible advantage over either the SPF or the MPS model.

We can apparently conclude that the Green Book, though respectable on output forecasting, does not have the advantage over the MPS model and private sector forecasts that it has for inflation forecasting.

III.4. Sources of Fed forecast accuracy. We have confirmed the Romers' conclusion that the Green Book forecasts are very good, in comparison to private and naive forecasts. They have also been historically slightly better, than the Fed's model-based forecasts, though the margin of superiority is statistically thin. Where does this superiority of Fed forecasts come from?

We can provide some evidence on this by embedding the Green Book forecasts in VAR models. We consider three somewhat extreme hypotheses:

- (1) the Fed is simply making better use than other forecasters of the same collection of aggregate time series available to all;
- (2) the Fed's forecasting advantage comes entirely from its knowledge, unavailable from published time series, of the Fed's own likely policy actions;
- (3) the Fed is simply collecting better detailed information about price developments, so that if other forecasters had knowledge of actual future inflation, the Fed's forecasts would not be useful to them.

We can formulate each of these possibilities as restrictions on a VAR model; the latter two gain some support from the data.

The first hypothesis suggests that if Fed forecast data is added to a VAR containing a list of standard quarterly variables known to be useful in forecasting, the Fed forecasts should not contribute substantially to the VAR's fit. Of course since the VAR uses *ex post* data for quarter t in constructing forecasts for $t + 1$, it has an unfair advantage over the Fed forecasts, which are made at t without even preliminary data on many date- t data values. If it turned out that Fed forecasts are indeed insignificant contributors to a VAR, we would have to take careful account of this bias, but in fact the result seems to point in the opposite direction. The Green Book forecasts make substantial contributions to the fit of a standard quarterly VAR, as can be seen from Table 15. The coefficients are highly significant in two of the five equations (GDP and the funds rate); the $\chi^2(5)$ statistic computed from the coefficient estimates and their estimated covariance matrix suggests that correlations among the coefficients strengthen the probability that they are non-zero; and the posterior odds ratio favors the model including the Green Book forecasts with odds of about 5 to 2. This odds ratio is probably the best measure of the strength of the evidence, and it is not decisive in itself. But since the use of *ex post* current data puts the GB forecasts at a disadvantage, the odds ratio is a rather strong indication that they contain information not available in contemporaneous values of the variables in the VAR.

If the Green Book forecasts were simply a better forecast of inflation, with the forecast errors themselves having no influence on the economy, then including actual current inflation on the right-hand side of the VAR should make the Green Book forecasts insignificant in all equations. As can be seen from Table 16, the evidence on this is mixed. The individual coefficients on the GB forecasts in the equations (other than the price equation itself) are all less than 2, and when considered jointly they are just barely "significant" at the .05 level. Posterior odds favor the model without GB forecasts by about 4 to 1, which is not decisive.

If the Green Book forecasts were better solely because they reflected greater knowledge of Fed policy intentions and misperceptions, then we might expect that once the actual

future funds rate were allowed on the right-hand side of the VAR regressions, the contribution of the GB forecast to the fit should disappear. The evidence here is quite similar to what emerges when we include current inflation in the system. The individual t -ratios on GB forecasts are all less than one, the joint χ^2 statistic is at the margin of significance (this time just below the .05 level rather than just above) and the posterior odds favor the model that excludes the GB forecasts, this time by about 10 to 1, which is approaching the decisive range.

Note that in Table 15 the coefficient on the GB forecast in the FF equation is strongly positive, indeed insignificantly different from 1. This means that even when several lagged values of actual inflation are included in the regression, the GB forecast of next period's inflation is a strong predictor of next period's FF rate. This is consistent with the view that the Fed responds to its own forecasts of inflation, and that its forecasts therefore contribute to fit through their contribution to forecasting interest rates.

Despite the caveats about data availability, the pattern of these results is consistent with a view that the superiority of the Fed forecasts arises from the Fed having an advantage in the timing of information — even with the view that this might arise entirely from the Fed having advance knowledge of its own policy intentions. The statistical results don't prove that this view is correct, but they support it as an interesting hypothesis.

IV. THE ROLE OF SUBJECTIVE FORECASTING

The persistence of the system of aggregating the views of sectoral experts to generate forecasts, despite decades of work on formal quantitative models, suggests that the expert system makes a contribution that is not easily duplicated with a formal model. What is this contribution?

One hypothesis is that the models are flawed descriptions of the economy (which is certainly true) and that the expert judgment of seasoned economists allows more subtle and accurate understandings of the economy to be brought to bear. None of the economists I talked to at these central banks expressed this view.

Instead, they claimed that the subjective forecasters mainly provide a more accurate picture of the current state of the economy than can easily be provided by a single quantitative

model. This view was stated most clearly and strongly by those actually involved in making subjective forecasts. They argued that they pay attention to a large amount of data from disparate sources, some of it non-quantitative. They also (this view was stated most clearly at the Fed) have an understanding of how disaggregated bits of data feed in to the preparation of the aggregate numbers that are put forward with some delay by the statistical agencies. This gives them a starting point for their forecasts that is more accurate than can be obtained from a model that is based on data at a fixed monthly or quarterly time interval and uses a slow-changing list of strictly quantitative variables. Because of the high persistence in most economic variables, this advantage in the initial period of the forecast translates into a persistent advantage. However several of those involved in subjective forecasting, in more than one bank, expressed the view that the advantage of subjective forecasts is almost entirely in getting the current and next quarter right. Extrapolations beyond this horizon, they felt, could be done more reliably with the large model.

The historical record of Fed forecasts we have examined in section III is consistent with this view. We saw that Green Book forecasts made at date t for quarter $t + 1$ make a strong contribution to a VAR model's fit if they are introduced on the right-hand-side of VAR regressions, with forecasts made at t competing for explanatory power with other variables dated t . But the contribution to fit from Green Book forecasts became much weaker when either next quarter's inflation or next quarter's interest rate was introduced as a right-hand-side variable. Also, the fact that the MPS model forecasts are very close in accuracy to the Green Book forecasts, together with the strong feedback between model forecasts and subjective forecasts in the policy process (as discussed in section II), is consistent with the view that large model forecasts can be as good as subjective forecasts if given equally good assessments of initial conditions.

If this view is correct, it helps explain why both large scale modeling and subjective, or expert-based, forecasting persist in all these central banks. For the foreseeable future, explicit quantitative models are going to be associated with fixed lists of variables. But of course the list of variables actually in use changes through time. In some cases recognition

of a policy problem or of newly apparent gaps in existing data can lead to new data collection efforts. This is accounting innovation, and it is hard to see how it can fail to include a substantial subjective or expert-based component.

Unusual events — an oil crisis, an attack on the currency, September 11, a data-collection error — can create large disturbances in data series that are not best treated as simple draws from a historical distribution of random disturbances. Because such disturbances are large and have an apparent explanation, it is likely to be possible to predict that their effects will have a different pattern of persistence or influence on other variables than a typical random draw of a disturbance. Analysis of such historically unusual disturbances — including determination of whether they really are historically unusual — will inevitably involve an element of subjective judgement. That is, since they are unique or unusual, extrapolating their effects must rely on more than historical statistical patterns of variation.

V. A BRIEF AND SELECTIVE HISTORY OF STATISTICAL MODELING FOR MACROECONOMIC POLICY

Tinbergen's 1939 early classic macroeconometric models collected equations estimated by single equation methods. (Haavelmo, 1944) pointed out that the resulting models implied a joint distribution for the data, and that the models should therefore be estimated and assessed as multivariate models. This insight led to further developments in statistical theory, computational methods, and applied macroeconomic modeling. This "simultaneous equations" or "Cowles Foundation" approach to modeling perhaps reached its peak in the MIT-Penn-SSRC (MPS) modeling collaboration, in which many leading academics worked to create a large scale model usable for forecasting and policy analysis.

As we have noted above in discussing the policy process, each of these four banks has a primary model. The models have evolved over time, responding to similar problems and pressures, and the banks clearly keep track of each others' work, so that there are important similarities in the models' current forms.

The MPS model became for over 15 years (of course slowly evolving over time) the main econometric model in use at the Board of Governors of the Federal Reserve. It was retired at the end of 1995, replaced by a new model, developed by Fed staff, that is known as FRBUS.

The Bank of Canada, meanwhile, used a sequence of models called RDX1, RDX2, and RDXF, finally scrapping them for a quite different model called the Quarterly Projection Model (QPM). The QPM model is essentially the same as that used at the Riksbank, and it has been influential on the modeling efforts of other inflation-targeting banks around the world.

Econometrics and macroeconomics were active research areas during the 70's, 80's, and 90's, and one might therefore have hoped that there would be clear progress as we moved from the early simultaneous equations models, to MPS and the RDX's, thence to the current QPM and FRBUS model. But if there is progress, it certainly isn't clear, and my own view is that the changes in these models over time have by and large been more regress than progress. This is not entirely the fault of the central bank researchers who have controlled the model evolution. The model builders have tried to take account of what they have perceived as modern developments in macroeconomics and data analysis. But academic research in these areas has paid very little attention to the central problems of modeling for macroeconomic policy in real time. The three main academic research themes that the modelers have tried to respond to are rational expectations (or, more broadly, dynamic stochastic equilibrium modeling), calibration, and unit root econometrics. The research that has emerged in these areas applies only very awkwardly to policy modeling problems. The attempts of central bank modelers to apply this research to their problems have therefore tended to make matters worse, not better. Another research theme, VAR modeling, was obtaining results that should on the face of it have been more directly applicable to real time policy modeling. However even here there were important mismatches between what was going on in the academic literature and the needs of the policy modelers.

Here is a list of the most serious problems faced by the modelers, with some discussion of the absence of guidance on them from academic research.

- (i) **The econometrics of many variables.** The four central banks that are our focus of discussion use models with from 15 to 40 behavioral equations, with several lags of each variable typically appearing in the model. The classic simultaneous equations toolkit was all based on asymptotic distribution theory, assuming sample size is very large relative to the number of parameters being estimated and the number

of variables being considered. But in these models those conditions do not hold. Two stage least squares, the most widely used and easily implemented estimator suggested by SE theory, degenerates to ordinary least squares when the number of instruments reaches sample size, and in practice gets close to OLS well before that. Econometric theory gave no guidance as to how to truncate the instrument list, or even as to whether it was a good idea to do so. LIML and FIML estimators had reputations as being difficult to implement and unreliable, and of course also have only asymptotic justification. If taking account of simultaneity implied using one of these estimation methods, it seemed to require a lot of work to end up with results that were arbitrary (if based on a truncated instrument list), almost OLS (2SLS with all available instruments), or quirky (FIML and LIML).

- (ii) **The need for decentralization.** Good real-time forecasting and policy analysis requires processing very large amounts of data. Maintaining a good forecasting model of a scale relevant to policy is more than a one-man task. The response to this situation in central banks has been, as we have already discussed, to allocate responsibility for “sectors” (largely identified with particular variables in a model) to “experts” or groups of experts. These experts in all banks are responsible for keeping up to date with the flow of current data in their area. The MPS model had been generated in a decentralized process, in which sectors of the model were assigned to individual economists or groups of economists. The Bank of Canada’s RDXF model differed from RDX1 and RDX2 in pushing this decentralization perhaps to its limit, with the sectoral experts maintaining separate equations of the overall model, with little attention to the properties of the resulting system of equations.⁶ The result was a system that was worse than its predecessors in its long-term simulation properties. A similar history has occurred at the Federal Reserve Board, in the development of its Global model. This model ties together the US model with models for 30 other countries. The list of countries has grown over time, with the need to take account of financial developments in various parts of the world. But as a result, the model can no longer be solved in its “forward-looking” mode (that is,

⁶(Coletti, Hunt, Rose, and Tetlow, 1996, p.9-10)

with model-consistent expectations). This is not a matter mainly of computational time; it reflects the nonexistence of a well-defined solution to the full system.

This is the direction in which the VAR literature veered furthest from policy-modeling reality. The VAR literature began by questioning whether the dissection of models into distinct, manageable equations with small numbers of variables could be justified, and urged treating equations more symmetrically, focusing on properties of the whole system of equations rather than on individual equations. The VAR literature may have been right about this in principle, but if the point is accepted, it still provides no answer as to how to proceed to model 30 or 40 variables jointly in real time, with the human resources available in a bank research department. It also provides no direct answer to the question of how to integrate the subjective input of experts, following more detailed data than is in the model, with the model's own results. The Federal Reserve Bank of Minneapolis did for a number of years maintain a VAR model of a policy-relevant scale, and this model did include a sectoral decomposition. However the model had a 9-variable behavioral core that took no feedback from the sectoral detail, which instead worked off of the core recursively. It could generate forecasts and policy projections with useful detail, but its structure implied that there was no role for sectoral expertise in improving forecasts of the main aggregates.

SE econometrics and rational expectations theory were equally unhelpful on this score, of course. SE theory implies that equations do not have any meaningful distinction between "left-hand-side" and "right-hand-side" variables, so that the practice of naming equations, and sectoral experts, after variables runs counter to SE theory. Rational expectations theory emphasized that exclusion restrictions limiting the list of variables in a behavioral equation were especially dubious, because of the way expectations make every part of the model relevant to behavior in other parts.

- (iii) **Integration of stochastic modeling with decision-making uncertainty.** Textbook econometrics has remained almost entirely non-Bayesian, meaning that it maintains a sharp distinction between "unknown", but non-random, "parameters" on the

one hand, and random “disturbances” on the other. Only estimators, not parameters themselves, are random. But in decision-making under uncertainty, simple axiomatics, and the intuition of most decision-makers, leads to thinking of everything unknown as subject to probability calculations, including parameter values and even which of a list of models is correct. Their conventional econometrics training leads bank staff economists to think of this kind of practical odds calculation as “unscientific”, or as not “econometric”. Here are two examples.

In my discussions with Fed staff, two economists, on separate occasions, brought up the example of the Fed’s wrestling during the 90’s with the question of whether the rate of productivity growth had undergone a permanent shift. They pointed out that evidence on this accumulated slowly, and that even now the conclusion remained uncertain. They suggested that if they had proceeded “scientifically” (the actual word used by one of them), they would have tested the null hypothesis of no change in the productivity growth rate and treated it as true until it was rejected at a 5% significance level. But as a practical matter, they pointed out, policy makers were not interested in that sort of analysis. Policy makers wanted to know what the weight of the evidence was — what were the current probabilities — not whether a test of the null was passed. Furthermore, policy makers were weighing the probability of a change in the growth rate against the costs of erring in either direction — assuming no change when there had in fact been change, or assuming a change when in fact there had not been.

That these elementary applications of the ideas of Bayesian decision theory were seen as “unscientific”, and as in conflict with the use of “econometrics”, is a sad commentary on the way econometrics is now being taught.

A very similar theme occurs in the Bank of Canada documents describing the origins of QPM:

... at policy institutions, the balancing of type I versus type II errors of econometric inference matters less than the balancing of type I versus type II errors in policy advice. Thus, it may be better for a policy model to assume that a particular economic structure exists, even when the evidence

is not overwhelming, if the costs of incorrectly assuming otherwise would be relatively high. (Black, Laxton, Rose, and Tetlow, 1994, p.65)

This passage occurs as part of a justification for considering criteria for model fit that are not “econometric”.

Inflation-targeting banks have generally (and this applies to both the Riksbank and the Bank of England) published their inflation forecasts in the form of “fan charts” — time series plots that show not a single line, but a fan of differently shaded colored regions, with darker regions meant to be considered more probable than lighter ones. Policy boards are concerned to make it clear, when they publish inflation forecasts, that they do not have a firmly held single number in mind, but instead a distribution of possible outcomes. This makes it clear that the policy board has not made a mistake, or failed to deliver on a commitment, when outcomes deviate from their most likely values by about the expected absolute amount. When I realized this, my initial thought was that this must imply an increased role for stochastic economic models, which would be used to generate the distributions needed for these charts. But in the Riksbank and the Bank of England, econometric models are not used at all in preparing these charts. There are two reasons for this. One is that these charts represent policy choices and commitments by the policy board. Their subjective judgment is an essential element in preparing the charts. Non-Bayesian approaches to econometrics have no conceptual framework for combining stochastic simulation of an econometric model with subjective judgment. Perhaps more important, everyone understands that the main source of uncertainty about forecasts generated by a model is not the disturbance terms in the model, but errors of estimation in the coefficients and uncertainty about whether this model, as opposed to a number of others from which forecasts are available, is closer to the truth. Combining subjective judgment, uncertainty about parameter values, and uncertainty across models with uncertainty about equation disturbance terms would be a technically demanding task in any case. But Bayesian thinking provides a clear conceptual starting point, whereas the classical apparatus of confidence intervals, tests, and distributions of estimators (as opposed to parameters),

provides no useful conceptual framework for these issues. The bank staff who work on these fan charts understand very well the importance of judgmental input, and that they are working with a version of subjective probability as they prepare the charts. But they have not seen a practical way to make econometric models useful as they do it.⁷

A senior staff member of the Bank of England, described an incident in which he was asked by a reporter who had just viewed a fan chart for output growth, what was the probability of two successive quarters of negative output growth over the span of the forecast. He had no answer, because the simple process of extracting judgmental probability distributions that generates the charts applies only to the terminal values the charts display. The evolution of the probability bands over time that the charts display is generated entirely by interpolation. This is the kind of useful extension of the existing analysis that could be produced if models and methods that can handle all the sources of uncertainty and merge it with subjective judgment were available.

(iv) **Modeling policy choice**

There are difficult conceptual issues surrounding the use of a statistical model fit to historical data to project the effects of a disturbance of known type, including when the disturbance is a change in policy. These issues were confronted and analyzed clearly early in the SE literature, perhaps best by Hurwicz (1962). He explained why it is by definition essential to have a structural model in order to analyze interventions. The early SE modelers understood this point in principle, though the fact that it applied strongly to macroeconomic policy interventions was not widely recognized early on in econometric practice. The most common procedure was to assume exogeneity of policy variables in estimation, and then to model policy changes as changes in the time path of policy variables. More recently, thanks mainly to the structural VAR literature, it has been recognized that separating policy behavior from other influences on policy variables (like the interest rate,

⁷Blix and Sellin (2000) provide a clear description of how such charts are prepared and of some of the underlying conceptual issues.

or money stock) is a non-trivial problem. It has become common for models to include an equation characterizing monetary policy behavior as setting the short term interest rate in response to the state of the economy. Policy changes are modeled as temporary or permanent changes in this reaction function, including as a special case setting time paths for the interest rate or (equivalently) for disturbances to the policy reaction function.

This is all straightforward, and practical policy modeling has never shown much confusion about it. However the rational expectations critique of econometric policy evaluation was interpreted by many economists as implying that this way of modeling policy choices had deep conceptual flaws. This viewpoint was perhaps best expressed by Sargent (1984). The idea is that setting time paths of policy equation disturbances does not change the unconditional joint distribution of the time series implied by the model, and that since (unconditional) expected welfare is determined by this joint distribution, nothing important is affected by such choices of random disturbances. Many economists, even on central bank staffs, claim to hold this view even now.⁸ Actual policy projections are still done, however, and in the same way as before for the most part. Policy simulations in QPM and FRBUS can be carried out with model-consistent expectations, and considerable effort went in to making these models capable of distinguishing such projections from the usual kind. However such projections, for policies that are not actually permanent changes, do not respond to Sargent's objections. Furthermore, it has turned out, as we will discuss below, that projections with model-consistent expectations are not the type most commonly used in practice.

- (v) **Rigor vs. Fit.** Most economists would agree that ideally a policy model should be derived from a theory of the behavior of economic agents who interact to generate

⁸Despite my having explained its fallaciousness with some frequency (Sims, 1987, e.g.). The basic idea is that, though unconditional distributions don't change, conditional distributions as of the current date do change, and this matters a great deal. Furthermore, the permanent change in "rule" that Sargent would have us limit attention to is actually only a special case of choosing "shocks" to policy. A change in rule, when it occurs, is the realization of a random variable, from the point of view of a rational public.

an equilibrium. Models that meet this criterion, though, are generally nonlinear and difficult to solve. If soluble with reasonable time and effort, such models tend to be fairly small and to be built around a small number of sources of stochastic disturbance. For these reasons, they tend not to fit the data nearly as well as more loosely restricted models, like VAR's. This situation has been a source of major concern at central banks for some years, and the banks have research under way to try to overcome the problems. Academic researchers have paid little attention to these issues, however. Those building dynamic stochastic general equilibrium models (DSGE's) have tended to stick to small models, with little apparent interest in expanding to the scale needed by central banks. This probably reflects the fact that many of those most active in working with DSGE's have the view that real-time monetary policy formation is not very important to economic welfare. Those building structural VAR models have written papers aimed at academic audiences, engaging in disputes about the effects of monetary policy and deliberately leaving unspecified the detailed interpretation of the private sector components of their models.

Despite concern about this issue at the banks, there has been little progress on it. The International Finance section of the Fed research department has a project underway to construct a DSGE model that might serve as a partial substitute for their Global model, which has (as we have already noted) become unwieldy. They are not at all sure how much of the Global model's function the new model can take over, however. The Bank of England is in the midst of a two year project to construct a new primary model. After considerable thought and investigation, they have concluded that they will not be able to construct a DSGE model that fits the data well enough to be used for forecasting. They plan to construct a DSGE model nonetheless, and append to it ad hoc stochastic elements that can bring it into contact with the data. This perceived tension between rigor and fit is discussed explicitly in the Bank of Canada's documentation for QPM (Coletti, Hunt, Rose, and Tetlow, 1996).

Both at the Bank of Canada and at the Fed, the existing primary models were seen at the beginning of the 90's as inadequate. At the Bank of Canada this reflected the results of radical decentralization and focus on single-equation fit, which had produced what was seen as excessive fluctuation in the form of the model and also made the model sometimes misbehave when used for longer term projections. I am not sure whether these considerations were as important at the Fed. However in both banks there was a desire to make the model more modern by introducing "forward-looking" elements, so that a distinction between anticipated and unanticipated policy changes was formally possible.

The new models, QPM and FRBUS, have a tiered structure. In a first tier long run, static relationships among variables are postulated. In the QPM, this tier is based on an overlapping generations growth model whose parameters are set by calibration. This means in practice simply that no measures of uncertainty are attached to parameter estimates. The parameter estimates emerge from a mixture of subjective prior information and informal matching of model properties to some summary statistics generated from the data. In the FRBUS model the static relationships are in many cases generated from static regression equations. The static regression equations are labeled "cointegrating equations", which allows invocation of asymptotic theory that implies that uncertainty in their coefficients is negligible and thereby also justifies estimating them by straightforward methods, independently of the second tier's complicated dynamics.

The second tier of the models describes adjustment to the long run equilibrium. In almost every behavioral equation, there is a single designated left-hand-side variable, and a "target" for it is generated from the first-tier static relations. An equation is then estimated describing the dynamics of adjustment of the left-hand-side variable to its target value.

This breaking of the model into static and dynamic tiers is common to all four of the models considered here in detail. It can be seen as a reaction to the tendency of models built up from decentralized sectors to display inconvenient long run simulation properties. However, whether via calibration ideology or cointegration asymptotic theory, it also insulates the long run properties of the model from any serious interaction with the data.

The QPM and FRBUS models introduce widespread expectational dynamics using a standard mechanism. The left-hand-side variable is assumed to be determined as the outcome of an optimization problem in which the variable tracks its target value subject to adjustment costs that depend on squared difference in the variable of order up to some k . Such an optimization problem implies that the current value of the variable will be determined by lagged and by expected future values of the target, with symmetry restrictions connecting the pattern of coefficients on future and on past values.

This approach to equation specification does introduce expectational terms pervasively in the model, but it does not respond convincingly to the critiques of the older SE models from the rational expectations and real business cycle perspectives. Those critiques emphasized the interdependence of model specification across equations and the derivation of dynamics and steady states from the same internally consistent, multiple-equation, equilibrium model. That research program does not lead to models that are collections of single-equation, adjustment-to-target specifications. A clear exposition of how this approach works in the investment sector of FRBUS is in Kiley (2001). This sector explains, with separate equations, demand for four kinds of investment goods — high tech equipment, other equipment, inventories, and non-residential structures. The possibility that these four highly interrelated variables might have dynamics that interact was apparently not considered. Equation estimates are presented separately, with no check for whether the implied restrictions on the cross-variable dynamics are consistent with the data.

In the QPM, the dynamic equations in the second tier are calibrated, just as are the steady state equations. The model is therefore probably unreliable for forecasting. The FRBUS model, on the other hand, is fit, largely by OLS, equation-by-equation. OLS estimation is possible because the expected future values in the dynamic equations are replaced by forecasts from a reduced form VAR. In principle it is possible that simultaneity would make the resulting collection of single equations perform badly as a system. However, the use of flexible lag structures, and the modest amount of simultaneity in the system, apparently make it a reasonable approximation to a VAR. The process of model-building included comparison of the model's impulse responses to those of a reduced form VAR, with adjustments undertaken if the deviations were too sharp.

The Bank of England and ECB models (MM for Macroeconometric Model and AWM for Area-Wide Model) contain much more limited forward-looking components. Their dynamic equations do not have the constrained symmetry of FRBUS and QPM, so that the appearance of lags in an MM or AWM equation does not force the appearance of corresponding lead terms. In fact, they both have explicit expectational terms only in the form of expected inflation, which enters mainly the wage-price block and is proxied in estimation by a weighted average of two previous quarters' inflation.

VI. ASSESSMENT OF THE PRIMARY MODELS

In giving up on the SE statistical theory that seemed to be providing no guidance to models of the scale actually in use, central bank modelers have ended up back at the level of the original Tinbergen model or worse. The models they use have no claim to be probability models of the joint behavior of the data series they are meant to explain, and they are not being checked against competitors by well-defined criteria of fit. It is unlikely that the models can start systematically improving without changing this situation.

These models have given up on serious effort to fit the data in good part as a tradeoff for an apparently more rigorous theoretical foundation. But the improvement in rigor is largely illusory. The single equation "target-tracking" specifications in FRBUS and QPM do not accord with modern DSGE theory, and they seem unlikely to give anything close to an accurate rendition of the contrast between anticipated and unanticipated policy changes. Indeed the FRBUS model is seldom used in its "model consistent expectations" mode. Most uses of the model are for monetary policy, and with a horizon of up to two years or so. In this time frame, it is not reasonable to suppose that the public would quickly perceive and act on a shift in policy behavior. The VAR forecasts are therefore likely to be more reasonable approximations to actual expectations. But further, even when accurate anticipation is more plausible, the model in rational expectations mode is said generally to imply unrealistically strong and quick responses to policy actions.

I used to argue that the big policy models, though not structural in the sense they claimed to be, were still useful summaries of statistical regularities, as they were not far from being simply big VAR's with quite a few exclusion restrictions. The QPM can't be rationalized

this way, because it is entirely calibrated. The FRBUS may still be a good data summary, though its tiered structure makes it difficult to assess this. Because data are available only with a five year delay, and because there was a six-month hiatus in the preparation of model forecasts during the 1996 transition to FRBUS from MPS, it will not be possible to assess the performance of the FRBUS model for a few more years — unless there is in fact a change in the Fed’s policy of keeping forecasts secret for 5 years.

VII. DIRECTIONS FOR IMPROVEMENT

Bayesian statistical inference is sometimes mistakenly thought of as a collection of “techniques” for doing the same sorts of things that can also be done by other “techniques”. But this is a mistake. Bayesian inference is a perspective, a way of thinking about statistical techniques, not in itself a collection of techniques. The Bayesian perspective on inference, were it widely understood by those working on policy models, would ease the connection between modeling and decision making. The non-Bayesian (sometimes, imprecisely, called “classical”) perspective, which is more appropriate (if anywhere) in the natural sciences, imposes on itself the rule that parameters and competing models do not have probabilities attached to them. Only potentially observable data have probabilities attached to them. It is not possible, within the classical perspective, to make a statement like, “given the data observed so far, the probability that β lies between 1.2 and 2.7 is .95”; or a statement like “given the data observed so far, and taking account of our uncertainty about model parameters, the probability that next quarter’s gdp growth rate will be between 1.1 per cent and 2.1 per cent is .95”; or a statement like, “given the data observed so far, the probabilities of the three models we have been using are .2, .1, and .7, respectively”. Economists well-trained in the classical perspective know this, and often claim not to be bothered by it. But when a decision maker, confronted with results from three models that conflict, asks for what the data imply about uncertainties across the models, he or she does not want to be told that no probability weights can be given for the models. Weighting uncertain prospects to compare the expected consequences of different courses of action is the essence of decision making.

Because the need for such probabilities, conditional on observed data, is so clear, classical statistics does attempt to produce substitutes for them. Confidence intervals, which are

not probability intervals conditional on data, are often interpreted as if they were. There is a classical literature on “model selection”, though it does not yield probabilities on models and produces a bewildering variety of procedures. There are ways, from a classical perspective, to create hybrids of confidence intervals and probability intervals that in some sense incorporate parameter uncertainty into measures of forecast uncertainty — but nonetheless do not result in probability intervals conditioned on the data.

Since these substitutes are often in practice interpreted as if they were actual probability statements conditioned on the data, the distinction between the Bayesian and classical perspectives is sometimes thought to be philosophical hairsplitting. But there are some real costs to the persistence of classical thinking among econometricians. There is an important difference, for example, between the situation where the data does not discriminate sharply among several competing models, but one fits slightly better, and the situation where a single model stands out as best fitting. Econometric analysis should be able to give quantitative guidance as to which type of situation we are in, rather than insisting on the impropriety of putting probabilities on models. Inference about unit roots and cointegration is, from a classical perspective, extremely complex — so much so that the academic literature provides little practical guidance about how to handle such phenomena in models of the scale of central bank primary models. The literature has also emphasized asymptotic results that in many cases appear to justify multi-stage inference, “testing” for unit roots and estimating cointegrating relationships, then treating the results of this first stage of inference as if they were free of uncertainty. But in actual, finite samples the uncertainty surrounding these first stage inferences is often high. Ignoring it produces unbelievable results. Despite not applying unit root theory at the full system level, because of its impracticality, several of the central bank models do apply it in exactly this multi-stage way, equation by equation. A Bayesian perspective would allow setting aside the multistage complexity of this inference and thereby would allow a more accurate and believable characterization of uncertainty about low frequency aspects of the model. Analysis of the possibility that a model’s parameters have shifted at discrete points in time is another situation, central to decision-making use of models, that is much more straightforward to analyze from a Bayesian than from a classical perspective.

The Bayesian perspective recognizes that all decision-making depends on decision-maker judgment as well as on inference from the data, and it formalizes the interaction of judgment and data, in contrast to the classical perspective, which avoids the subject. When the role of judgment is kept under the table, it becomes more difficult to discuss it and more difficult to recognize bad judgment.

But the most persuasive argument for the Bayesian perspective is its increasing ability to make apparently intractable inference problems tractable. Within the last 5 years or so, economists have become aware of Markov-chain Monte Carlo methods for Bayesian analysis of econometric models. These methods are being applied rather widely, particularly in finance, and as economists see them produce insights in otherwise intractable problems, they are likely to learn how to use these methods and understand their rationale. In fact, a recent paper (Smets and Wouter, 2002) apparently represents the first example of a DSGE that has been fit to data and produces a fit that is competitive with that of a Bayesian reduced form VAR. The paper accomplishes this in a Bayesian framework, using Markov chain Monte Carlo methods, and thereby produces a model that should be directly usable for realistic stochastic simulation and that should be directly comparable in fit to models with different specifications. While it explains just 9 variables, it was put together by two researchers in a relatively short span of time. With the resources of a central bank research staff and computational equipment, the same methods should work on models of the scale of current central bank primary models. On the face of it, this makes obsolete the widespread belief that rigorous dynamic theoretical modeling and good statistical fit are incompatible.

So the problem of conflict between rigor and fit in modeling may be on its way to resolution. Since it looks like the resolution will involve increased awareness of Bayesian ideas, along the way there may be progress in formalizing the connection between subjective judgment and model forecasts. But the problem of decentralizing modeling effort seems likely to be difficult for some time.

A model like that of Smets and Wouter (2002) has firm and household sectors, but each of these sectors generates several tightly related behavioral equations. It does not seem much more suited than a VAR to equation-by-equation decentralization. Given the nature

of central bank subjective forecasting, though, it might be possible to combine variable-by-variable decentralization of expertise input with a more integrated approach to modeling. It might be worth exploring a structure in which judgmental forecasters focus almost entirely on the current and next quarter. Their forecasts could be treated as noisy observations on the data that enter a Bayesian structural model. Model forecasts could then incorporate these observations, as well as generate measures of how implausible they are.

VIII. CONCLUSION

As I see it the most important component of inflation-targeting is the regular reporting of forecasts and of policy analyses by the central bank that the targeting regime entails.⁹ This supports bank policy by making it easier to preserve credibility in the face of shocks that create temporary increases in the inflation rate. By announcing a policy path and corresponding inflation path, the central bank may be able to convince people that the inflation will end without having to generate a recession. This regular reporting of forecasts also encourages probabilistic thinking and creates a demand, yet unsatisfied, for policy models that can generate realistic measures of uncertainty about their results.

Some apparently unnecessary barriers to transparency of monetary policy persist. Other banks around the world are regularly reporting inflation and output forecasts without ill consequences, indeed with apparently good consequences. The Fed's Green Book forecasts of inflation are of very high quality. They could be useful to the private sector and if published in timely way they could contribute to the effectiveness of Fed policy actions. It seems that at least for these forecasts the Fed's 5 year embargo should be dropped.

In the inflation targeting countries, internal consistency of forecasts and the level of discussion of policy would be elevated by switching to a practice of publishing forecasts in which inflation, output, and interest rates all appear and have been derived from a model to be mutually consistent. The usual objection is that this asks too much of policy boards that already have difficulty agreeing on just the current level of the policy rate. However the task of choosing among and adjusting several proposed time paths for the policy rate does not seem much more difficult than the problem of generating a skewed fan chart distribution

⁹For an analysis of the value of this type of transparency that supports my views here, see Geraats (2001).

that policy boards are already solving. So here again is an area where considerable progress could be had cheaply.

Finally, there are the central challenges of policy modeling. Is it realistic to hope that we can find fully identified models, allowing story-telling about not only the effects of policy but also about where major shocks have originated, and that also fit the data? Is it realistic to expect that economists can learn to understand the Bayesian perspective on inference and how it dovetails with decision theory? Is it practical to continue to do policy modeling in teams, while also improving the models along these lines? I've argued that the answer to the first two of these questions is yes, with important developments in these areas likely soon. The third question, about decentralizing modeling, does not have as clear an answer, though there is some reason for hope.

The academic branch of the economics profession has not been contributing much to the progress of policy modeling. Increased attention by scholars to these interesting problems might fairly quickly resolve many of the problems I have cited with central bank practice. However, academic interest in these issues has been low for years, and I am not sure how this can be changed.

REFERENCES

- ATKESON, A., AND L. E. OHANIAN (2001): "Are Phillips Curves Useful for Forecasting Inflation?," *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), 2–11.
- BLACK, R., D. LAXTON, D. ROSE, AND R. TETLOW (1994): "The Bank of Canada's New Quarterly Projection Model Part 1, The Steady-State Model: SSQPM," Technical Report 72, Bank of Canada.
- BLIX, M., AND P. SELLIN (2000): "A Bivariate Distribution for Inflation and Output Forecasts," Working Paper 102, Sveriges Riksbank.
- COLETTI, D., B. HUNT, D. ROSE, AND R. TETLOW (1996): "The Bank of Canada's New Quarterly Projection Model Part 3 The Dynamic Model: QPM," Technical Report 75, Bank of Canada.
- FAGAN, G., J. HENRY, AND R. MESTRE (2001): "An Area-Wide Model (AWM) for the Euro Area," Working Paper 42, European Central Bank, Frankfurt.

- GERAATS, P. M. (2001): "Why Adopt Transparency? The Publication of Central Bank Forecasts," Working Paper 41, European Central Bank.
- HAAVELMO, T. (1944): "The Probability Approach in Econometrics," *emet*, 12(Supplement), iii-vi+1-115.
- HURWICZ, L. (1962): "On the Structural Form of Interdependent Systems," in *Logic, Methodology and Philosophy of Science*, pp. 232-239. Stanford University Press, Stanford, CA.
- KILEY, M. (2001): "Business Investment in the Federal Reserve Board's U.S. Model (FRB/US): Specification and Implications," Frbus support document, Board of Governors of the Federal Reserve.
- POLOZ, S., D. ROSE, AND R. TETLOW (1994): "The Bank of Canada's new Quarterly Projection Model (QPM): An introduction," *Bank of Canada Review*, pp. 23-38.
- QUINN, M. (2000): *Economic Models at the Bank of England: September 2000 Update*. Bank of England, London.
- ROMER, C. D., AND D. H. ROMER (2000): "Federal Reserve Information and the Behavior of Interest Rates," *American Economic Review*, 90(3), 429-457.
- SARGENT, T. J. (1984): "Autoregressions, Expectations, and Advice," *American Economic Review*, 74, 408-15, Papers and Proceedings.
- SHILLER, R. J., AND R. FAIR (1989): "The Informational Content of Ex Ante Forecasts," *restat*, 71(2), 325-331.
- SIMS, C. A. (1987): "A Rational Expectations Framework for Short Run Policy Analysis," in *New Approaches to Monetary Economics*, ed. by W. Barnett, and K. Singleton, pp. 293-310. Cambridge University Press.
- (1989): "Modeling Trends," in *Proceedings, American Statistical Association Annual Meetings*, <http://eco-072399b.princeton.edu/yftp/trends/asa889.pdf>.
- (1993): "A 9 Variable Probabilistic Macroeconomic Forecasting Model," in *Business Cycles, Indicators, and Forecasting*, ed. by J. H. Stock, and M. W. Watson, vol. 28 of *NBER Studies in Business Cycles*, pp. 179-214.

- (revised 1996): “Inference for Multivariate Time Series with Trend,” Discussion paper, presented at the 1992 American Statistical Association Meetings, <http://eco-072399b.princeton.edu/yftp/trends/asaper.pdf>.
- SIMS, C. A., AND T. ZHA (1998): “Bayesian Methods for Dynamic Multivariate Models,” *International Economic Review*.
- SMETS, F., AND R. WOUTER (2002): “An Estimated Stochastic Dynamic General Equilibrium Model of the Euro Area*,” working paper, European Central Bank and National Bank of Belgium.
- TINBERGEN, J. (1939): *Business Cycles in the United States of America, 1919-1932*, vol. II of *Statistical Testing of Business Cycle Theories*. League of Nations, Geneva.

APPENDIX A. VAR PRIORS

The estimation of the VAR models in this paper uses priors meant to aid in interpreting the likelihood shape by concentrating our attention on the most reasonable parts of the parameter space. This is necessary in the first place because time series models, especially models where the number of parameters is a relatively large fraction of sample size, easily produce spuriously precise results, attributing much of the observed sample behavior to unusual initial conditions, if no prior is used. This problem is the Bayesian counterpart of the cumbersome unit root theory in classical inference (Sims, revised 1996, 1989).

In addition, we are comparing restricted and unrestricted versions of the VAR models here, and there are well known difficulties in deciding how properly to penalize overparameterized models that have good in-sample fits. The widely applied Schwarz criterion is one approach to such penalization, and it has a Bayesian justification as a decision procedure in large samples, but its justification when unit roots are possibly present is problematic. Since there is a widely used, standardized family of priors for VAR’s that can easily be made proper (that is, made to integrate to one), using this prior directly rather than the Schwarz asymptotic approximation seems like a good idea, and we have done so here.

The prior pdf is of the form

$$|\Sigma|^{-n-\frac{1}{2}} |\Sigma|^{-k/2} |X_d' X_d|^{n/2} e^{-\frac{1}{2}(y_d - (I \otimes X_d)\beta)' (\Sigma^{-1} \otimes I)(y_d - (I \otimes X_d)\beta)}, \quad (8)$$

where y_d and X_d are dummy right-hand-side and left-hand-side variables, k is the number of columns in X_d , n is the number of equations, and β is the vector of parameters from all equations, stacked on each other. We have omitted some factors that vary with n , but not with the data or with k .

This is just convenient notation for a prior that, conditional on Σ , is Gaussian for β and with the covariance of the β 's across equations fitting a $\Sigma \otimes \Omega$ form, where $\Omega = (X_d' * X_d)^{-1}$. The marginal prior on Σ is proportional to $|\Sigma|^{-n-\frac{1}{2}}$. This is not a proper prior. It is the limit of an inverse-Wishart prior with n degrees of freedom as the scale matrix shrinks toward zero. A proper prior with very small scale matrix would give nearly identical results.

This prior is in conjugate form, so that, combined with the likelihood, it can be integrated analytically to provide posterior odds on models.

The dummy observations actually used were in three sets.

- (1) A ‘‘cointegration’’ dummy. This is a single observation in which all exogenous and all current and lagged endogenous variables in the system are set equal to their means over the period of the initial conditions. If given very high weight, it forces the appearance of at least one unit root affecting all variables, in which case it also enforces zero coefficients on exogenous variables, or else it forces the system towards stationarity with a steady state equal to the initial condition means. This observation is weighted by the hyperparameter λ .
- (2) A set of n ‘‘unit root’’ dummies. These, if weighted very strongly, force the appearance of n unit roots in the system, one in each variable. The i 'th of these is an observation in which the i 'th variable and its lags are all set equal to initial condition means, and all other variables, including the constant term, are set to zero. These observations are weighted by the hyperparameter μ .
- (3) A version of the Minnesota prior. These are $nq + r$ dummy observations, where q is the number of lags in the system. For the j 'th variable's k 'th lag, the ‘‘x’’ dummy observation is nonzero only for that lag of that variable and is set equal to $\sigma_j k^\theta$, where σ_j is the standard deviation of the changes in the variable over the period of the initial conditions. There are also dummy observations corresponding to the

other right-hand-side variables, sized simply at the standard deviation of the “x-variable” itself over the period of the initial conditions. One can think of these as even applying to the constant term, but since the constant term has zero variance, the weight on the corresponding dummy observation is zero. This whole set of observations is given weight ζ . The y_d variable is non-zero only for the lag q of 1, in which case both the left-hand-side and right-hand-side y 's are set to σ_j . This makes the prior mean emerging from these dummy observations be 1 for the first own lag, zero for all other own lags and for all other variables on the right hand side.

The VAR results reported in the tables use $\lambda = 5$, $\mu = 2$, and $\zeta = 3$. The $\zeta = 3$ value is close to maximizing the posterior pdf, conditional on $(\lambda, \mu) = (5, 2)$. The values of λ and μ were set based on experience with similar models. They cannot be chosen to maximize the posterior pdf without reintroducing the problem of estimates implying spuriously unusual initial conditions.

The family of priors from which this one is drawn is discussed in more detail in Sims (1993); Sims and Zha (1998).

FIGURE 1. Four-quarter-ahead Inflation Forecasts, 1979-1975

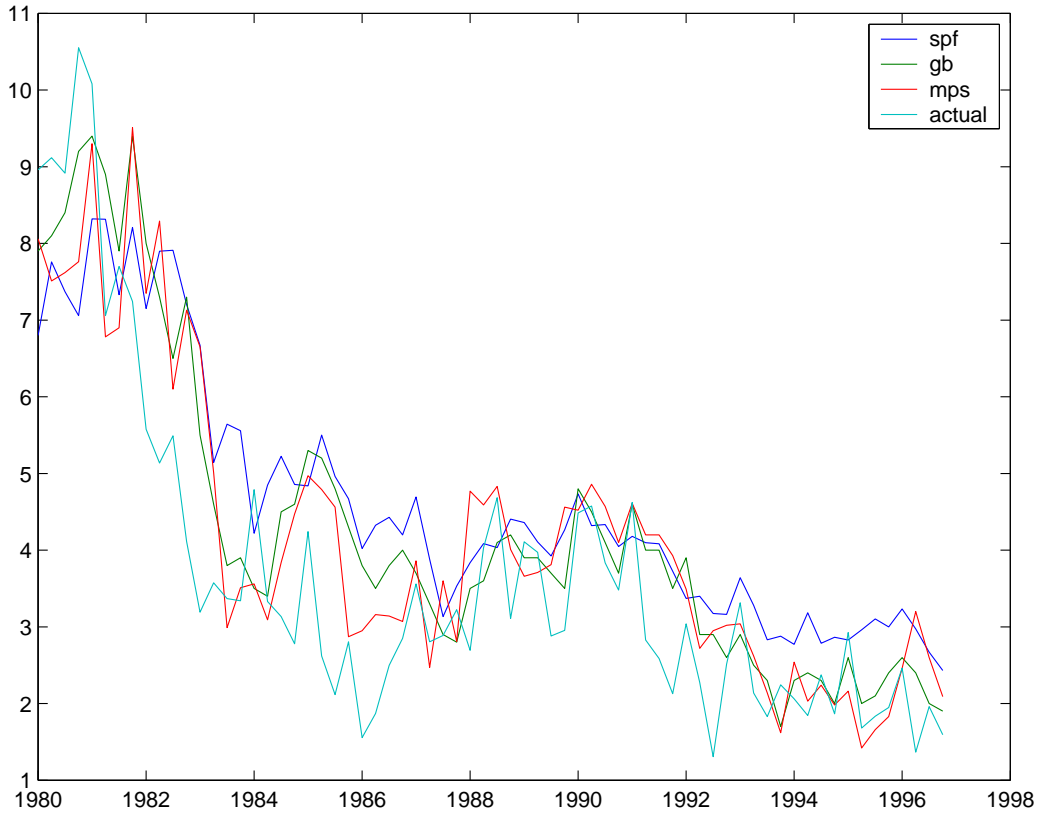


TABLE 1. RMSE's of inflation forecasts over 1979-1995

forecast	quarters ahead				
	0	1	2	3	4
naïve	0.00	0.94	1.15	1.14	1.35
spf	0.80	1.02	1.22	1.41	1.54
gb	0.96	0.91	0.92	0.99	1.16
mps	1.10	1.08	1.16	1.10	1.24

TABLE 2. Correlation of 4-quarter inflation forecasts

naïve	1	0.9079	0.9336	0.8937	0.8324
spf	0.9079	1	0.953	0.9106	0.8117
gb	0.9336	0.953	1	0.9528	0.8877
mps	0.8937	0.9106	0.9528	1	0.8486
actual	0.8324	0.8117	0.8877	0.8486	1

TABLE 3. Correlation of 1-quarter inflation forecasts

naïve	1.0000	0.9488	0.9327	0.9091	0.9170
spf	0.9488	1.0000	0.9539	0.9282	0.9212
gb	0.9327	0.9539	1.0000	0.9494	0.9458
mps	0.9091	0.9282	0.9494	1.0000	0.8963
actual	0.9170	0.9212	0.9458	0.8963	1.0000

TABLE 4. Regression of actual inflation on forecasts, 4-step ahead, 79-95

forecast	coefficient	std error
spf	-0.4882	0.30683
gb	1.2564	0.3234
mps	0.0444	0.23511
constant	0.3447	0.61875
R^2	0.8009	
SEE	0.9751	

Standard errors account for MA(3)
serial correlation

TABLE 5. Regression of actual inflation on forecasts, 1-step ahead, 79-95

forecast	coefficient	std error
spf	0.2677	0.1596
gb	0.7750	0.1578
mps	-0.0652	0.1231
constant	-0.4221	0.2553
R^2	0.9750	
SEE	0.7472	

TABLE 6. One-factor model for four quarter inflation forecasts

	f	σ^2	bias
naive	1.1505	0.6740	0.3738
spf	0.8235	0.2372	0.9222
gb	1.0794	0.0094	0.6143
mps	0.9818	0.3377	0.5058
actual	1	0.9511	0
<hr/>			
log LH		-6.3542	
Unconstrained log LH		-6.3061	
<hr/>			
T for $\chi^2 = .05$		115	
T for SC		438	
<hr/>			
σ, μ for actual		2.12	3.6768

TABLE 7. One-factor model for one quarter inflation forecasts

	f	σ^2	bias
naive	1.0195	0.4493	0.0765
spf	0.8714	0.2028	0.499
gb	1.0566	0.1712	0.4906
mps	1.0609	0.5021	0.0309
actual	1	0.4729	0
<hr/>			
log LH		-6.4929	
Unconstrained log LH		-6.3868	
<hr/>			
T for $\chi^2 = .05$		53	
T for SC		99	
<hr/>			
σ, μ for actual		2.2811	3.9741

TABLE 8. Assessing evidence against equal inflation RMSE's

	4-quarter		1-quarter	
	mps/gb	spf/gb	mps/gb	spf/gb
log LH difference	0.007900	0.160500	0.026400	0.283700
T for .05	243	12	73	7
T for SC	534	26	160	15

TABLE 9. RMSE's of output growth forecasts over 1979-1995

forecast	quarters ahead				
	0	1	2	3	4
naïve	0.00	3.61	4.09	4.39	4.58
spf	2.46	2.93	3.09	3.37	3.12
gb	2.38	2.89	3.07	3.20	3.02
mps	2.69	3.05	3.16	3.43	3.24

TABLE 10. Correlation of 4-quarter output forecasts

naïve	1.0000	-0.1444	0.2605	0.2437	0.0180
spf	-0.1444	1.0000	0.2871	0.2819	0.2498
gb	0.2605	0.2871	1.0000	0.8230	0.4191
mps	0.2437	0.2819	0.8230	1.0000	0.3216
actual	0.0180	0.2498	0.4191	0.3216	1.0000

TABLE 11. Correlation of 1-quarter output forecasts

naïve	1.0000	0.5365	0.5114	0.4004	0.3850
spf	0.5365	1.0000	0.8863	0.7503	0.4664
gb	0.5114	0.8863	1.0000	0.8210	0.5047
mps	0.4004	0.7503	0.8210	1.0000	0.4411
actual	0.3850	0.4664	0.5047	0.4411	1.0000

TABLE 12. Regression of actual inflation on forecasts, 4-step ahead, 79-95

forecast	coefficient	std error
spf	0.4853	0.4130
gb	1.4133	0.6112
mps	-0.1857	0.4054
constant	-1.5002	1.3170
R^2	0.1967	
SEE	3.0278	

Standard errors account for MA(3)
serial correlation

TABLE 13. Regression of actual output growth on forecasts, 1-step ahead, 79-95

forecast	coefficient	std error
spf	0.3289	0.4664
gb	0.4744	0.4813
mps	0.1224	0.3057
constant	0.7366	0.5523
R^2	0.2626	
SEE	2.8802	

TABLE 14. Assessing evidence against equal output growth RMSE's

	4-quarter		1-quarter	
	mps/gb	spf/gb	mps/gb	spf/gb
log LH difference	0.024094	0.009275	0.010729	0.000000
T for .05	80	207	179	∞
T for SC	175	455	393	∞

TABLE 15. 1-step GB forecasts in a VAR

	coefficient	std. Error	t-stat
GNP/GDP	-0.2558	0.1585	-1.61
GNP/GDP deflator	0.1852	0.0735	2.52
FFR	0.7145	0.2279	3.13
Comm.PI	-0.5663	2.2832	-0.25
gs3	0.0481	0.6836	0.07
$\chi^2(5)$		23.7	
log posterior odds against exclusion		20.1	

This VAR and the restricted version on which the last row of the table is based were estimated using a combination of a “Minnesota prior” symmetric across own and other variables with “unit root” and “cointegration” dummy observations. Parameters were $\lambda = 5$, $\mu = 2$, $\zeta = 3$, $\theta = 1.5$. Details in the appendix. The VAR used 4 lags and included the previous period’s one-quarter-ahead GB forecast as an explanatory variable.

TABLE 16. 1-step GB forecasts in a VAR with current inflation

	coefficient	std. Error	t-stat
GNP/GDP	-0.3609	0.2148	-1.68
FFR	0.5600	0.3089	1.81
Comm.PI	-3.3499	3.0942	-1.08
gs3	-0.4064	0.9264	-0.44
$\chi^2(4)$		10.02	
P[GB helps]		.2024	

TABLE 17. 1-step GB forecasts in a VAR with current funds rate

	coefficient	std. Error	t-stat
GNP/GDP	-0.3011	0.1710	-1.76
GNP/GDP deflator	0.3026	0.2258	1.34
Comm.PI	-0.4614	2.4671	-0.19
gs3	-0.1947	0.7364	-0.26
$\chi^2(4)$		9.26	
P[GB helps]		.0720	

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY

E-mail address: sims@princeton.edu