Section III

Alternative forecasting models: a brief overview

Predicting the interest rate is a difficult task since the forecasts depend on the model used to generate them. Hence, it is important to study the properties of forecasts generated from different models and select the "best" on the basis of an objective criterion. The aim of this study is to select the "best" model for each interest rate from a number of alternative models estimated[1] .

The ***benchmark model*** for each interest rate is a ***naïve model*** that implies that the projection for the next period is simply the actual value of the variable in the current period. The naïve model is essentially a random walk as described below:

$$i_t = i_{t-1} + \varepsilon_t$$

with $E(\varepsilon_t)=0$ and $E(\varepsilon_t\varepsilon_s)=0$ for $t \neq s$.

The one-period-ahead forecast is simply the current value as shown below:

$$i^e_{t+1} = E(i_t + \varepsilon_{t+1}) = i_t$$

Similarly the k- period-ahead forecast is:

$$i^e_{t+k} = i_t$$

The next step is to estimate ARIMA models that predict future values of a variable exclusively on the basis of its own past history. These models are then extended to include ARCH/GARCH effects. Clearly, univariate models are not ideal since these do not use information on the relationships between different economic variables. These are, however, a good starting point since predictions from these models can be compared with those from multivariate models.

III.1. ARIMA Models

An ARIMA(p,d,q) can be represented as:

$\varphi(L)(1-L)^d y_t = \delta + \theta(L)\varepsilon_t$ where L = backward shift operator

$\varphi(L)$ = autoregressive operator $= 1-\varphi_1 L- \varphi_2 L^2-\ldots\ldots\ldots- \varphi_p L^p$

$\theta(L)$ = moving average operator $= 1- \theta_1 L- \theta_2 L^2-\ldots\ldots\ldots- \theta_q L^q$

The stationarity condition for an AR(p) process implies that the roots of $\varphi(L)$ lie inside the unit circle, i.e., all the roots of $\varphi(L)$ are less than one in absolute value. Restrictions are also imposed on $\theta(L)$ to ensure invertibility so that the MA(q) part can be written in terms of an infinite autoregression on y. Furthermore, if a series requires differencing 'd' times to yield a stationary

series, then the differenced series is modelled as an ARMA(p,q) process or equivalently, an ARIMA(p,d,q) model is fitted to the series.

Other criteria employed to select the best-fit model include parameter significance, residual diagnostics, and minimization of the Akaike Information Criterion and the Schwartz Bayesian Criterion.

ARIMA-ARCH/GARCH Models

The assumption of constant variance of the innovation process in the ARIMA model can be relaxed following Engle's (1982) seminal paper and its extension by Bollerslev (1986) on modelling the conditional variance of the error process. One possibility is to model the conditional variance as an AR(q) process using the square of the estimated residuals, i.e., the autoregressive conditional heteroscedasticity (ARCH) model. The conditional variance thus follows an MA process, while in its generalized version – GARCH – it follows an ARMA process. Adding this information can improve the performance of the ARIMA model due to the presence of the volatility clustering effect characteristic of financial series. In other words, the errors, ε_t although serially uncorrelated through the white noise assumption, are not independent since they are related through their second moments. Hence, large values of ε_t are likely to be followed by large values of ε_{t+1} of either sign. Consequently, a realisation of ε_t exhibits behaviour in which clusters of large observations are followed by clusters of small ones.

According to Engle's basic ARCH model, the conditional variance of the shock that occurs at time t is a linear function of the squares of the past shocks. For example, an ARCH(1) model is specified as:

$$Y_t = E\left[Y_t \mid \Omega_{t-1}\right] + \varepsilon_t$$

$$\varepsilon_t = v_t \sqrt{h_t} \text{ and } h_t = \alpha_0 + \alpha_1 \varepsilon^2_{t-1}$$

where v_t is a white noise process and is independent of and ε_{t+1} and ε_t has mean of zero and is uncorrelated. For the conditional variance h_t to be non-negative, the conditions $\alpha_0 > 0$ and $\alpha_1 \geq 0$ and $0 \leq \alpha_1 \leq 1$ (for covariance stationarity) must be satisfied. To understand why the ARCH model can describe volatility clustering, observe that the above equations show that the conditional variance of ε_t is an increasing function of the shock that occurred in the previous time periods. Therefore if ε_{t+1} is expected to be large (in absolute value) as well. In other words, large (small) shocks tend to be followed by large (small) shocks, of either sign.

To model extended persistence, generalizations of the ARCH (1) model such as including additional lagged squared shocks can be considered as in the ARCH (q) model below:

$$h_t = \alpha_0 + \alpha_1 \varepsilon^2_{t-1} + \alpha_2 \varepsilon^2_{t-2} + \ldots\ldots + \alpha_q \varepsilon^2_{t-q}$$

For non-negativeness of the conditional variance, the following conditions must be met: $\alpha_0 > 0$, $\alpha_1 > 0$ and $1 > \Sigma_i \alpha_i \geq 0$ for all $i = 1,2,3, \ldots\ldots, q$.

To capture the dynamic patterns in conditional volatility adequately by means of an ARCH (q) model, q often needs to be quite large. Estimating the parameters in such a model can therefore be cumbersome because of stationarity and non-negativity constraints. However, adding lagged conditional variances to the ARCH model can circumvent this drawback. For example, including h_{t-1} to the ARCH (1) model, results in the Generalized ARCH (GARCH) model of the order (1,1):

$$h_t = \alpha_0 + \alpha_1 \varepsilon^2_{t-1} + \beta_1 h_{t-1}$$

The parameters in this model should satisfy $\alpha_0 > 0$, $\alpha_1 > 0$ and $\beta_1 \geq 0$ to guarantee that $h_t \geq 0$, while α_1 must be strictly positive for to β_1 be identified. Generalising, the GARCH (p,q) model is given by:

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2_{t-i} + \sum_{i=1}^{p} \beta_i h_{t-i}$$

$$h_t = \alpha_0 + \alpha(L)\ \varepsilon^2_t + \beta(L)\ h_t$$

Assuming that all the roots of $1 - \beta(L)$ are outside the unit circle, the model can be rewritten as an infinite-order ARCH model.

As indicated above, univariate models such as ARIMA and ARCH/GARCH models utilize information only on the past values of the variable to make forecasts. We now consider multivariate forecasting models that rely on the interrelationships between different variables.

III.2. VAR and BVAR Modelling

As a prelude to the discussion on multivariate models, it is apposite to note that according to the Statement on Monetary and Credit Policy for 2002-03, short-term forecasts of interest rates need to take cognizance of possible movements in all other macreconomic variables including investment, output and inflation, which are, in turn, susceptible to unanticipated changes emanating from unforseen domestic or international developments. Multivariate forecasting models address such concerns and are often formulated as simultaneous equations structural models. In these models, economic theory not only dictates what variables to include in the model, but also postulates which explanatory variables to use to explain any given independent variable. This can, however, be problematic when economic theory is ambiguous. Further,

structural models are generally poorly suited for forecasting. This is because projections of the exogenous variables are required to forecast the endogenous variables. Another problem in such models is that proper identification of individual equations in the system requires the correct number of excluded variables from an equation in the model.

A vector autoregressive (VAR) model offers an alternative approach, particularly useful for forecasting purposes. This method is multivariate and does not require specification of the projected values of the exogenous variables. Economic theory is used only to determine the variables to include in the model.

Although the approach is "atheoretical," a VAR model approximates the reduced form of a structural system of simultaneous equations. As shown by Zellner (1979), and Zellner and Palm (1974), any linear structural model theoretically reduces to a VAR moving average (VARMA) model, whose coefficients combine the structural coefficients. Under some conditions, a VARMA model can be expressed as a VAR model and as a Vector Moving Average (VMA) model. A VAR model can also approximate the reduced form of a simultaneous structural model. Thus, a VAR model does not totally differ from a large-scale structural model. Rather, given the correct restrictions on the parameters of the VAR model, they reflect mirror images of each other.

The VAR technique uses regularities in the historical data on the forecasted variables. Economic theory only selects the economic variables to include in the model. An unrestricted VAR model (Sims 1980) is written as follows:

yt	=	$C + A(L)y_t + e_t$, where
y	=	n (nx1) vector of variables being forecast;
A(L)	=	an (nxn) polynomial matrix in the back-shift operator L with lag length p,
	=	$A_1L + A_2L^2 + \ldots\ldots + A_pL^p$
C	=	an (nx1) vector of constant terms; and
e	=	an (nx1) vector of white noise error terms.

The model uses the same lag length for all variables. One serious drawback exists — overparameterization produces multicollinearity and loss of degrees of freedom that can lead to inefficient estimates and large out-of-sample forecasting errors. One solution excludes insignificant variables/lags based on statistical tests.

An alternative approach to overcome overparameterization uses a Bayesian VAR model as described in Litterman (1981), Doan, Litterman and Sims (1984), Todd (1984), Litterman (1986), and Spencer (1993). Instead of eliminating longer lags and/or less important variables, the Bayesian technique imposes restrictions on these coefficients on the assumption that these are more likely to be near zero than the coefficients on shorter lags and/or more important variables. If, however, strong effects do occur from longer lags and/or less important variables, the data can override this assumption. Thus the Bayesian model imposes prior beliefs on the relationships between different variables as well as between own lags of a particular variable. If these beliefs (restrictions) are appropriate, the forecasting ability of the model should improve.

The Bayesian approach to forecasting therefore provides a scientific way of imposing prior or judgmental beliefs on a statistical model. Several prior beliefs can be imposed so that the set of beliefs that produces the best forecasts is selected for making forecasts. The selection of the Bayesian prior, of course, depends on the expertise of the forecaster.

The restrictions on the coefficients specify normal prior distributions with means zero and small standard deviations for all coefficients with decreasing standard deviations on increasing lags, except for the coefficient on the first own lag of a variable that is given a mean of unity. This so-called "Minnesota prior" was developed at the Federal Reserve Bank of Minneapolis and the University of Minnesota.

The standard deviation of the prior distribution for lag m of variable j in equation i for all i, j, and m — $S(i, j, m)$ — is specified as follows:

$$
\begin{aligned}
S(i, j, m) &= \{wg(m)f(i, j)\}s_i/s_j; \\
f(i, j) &= 1, \text{ if } i = j; \\
&= k \text{ otherwise } (0 < k < 1); \text{ and} \\
g(m) &= m^{-d}, \; d > 0.
\end{aligned}
$$

The term ^{s}i equals the standard error of a univariate autoregression for variable i. The ratio s_i/s_j scales the variables to account for differences in units of measurement and allows the specification of the prior without consideration of the magnitudes of the variables. The parameter w measures the standard deviation on the first own lag and describes the overall tightness of the prior. The tightness on lag m relative to lag 1 equals the function $g(m)$, assumed to have a harmonic shape with decay factor d. The tightness of variable j relative to variable i in equation i equals the function $f(i, j)$.

To illustrate, assume the following hyperparameters: $w = 0.2$; $d = 2.0$; and $f(i, j) = 0.5$. When $w = 0.2$, the standard deviation of the first own lag in each equation is 0.2, since $g(1) = f(i, j) = s_i/s_j = 1.0$. The standard deviation of all other lags equals $0.2[s_i/s_j\{g(m)f(i, j)\}]$. For $m = 1, 2, 3, 4$, and $d = 2.0$, $g(m) = 1.0, 0.25, 0.11, 0.06$, respectively, showing the decreasing influence of longer lags. The value of $f(i, j)$ determines the importance of variable j relative to variable i in the equation for variable i, higher values implying greater interaction. For instance, $f(i, j) = 0.5$ implies that relative to variable i, variable j has a weight of 50 percent. A tighter prior occurs by decreasing w, increasing d, and/or decreasing k. Examples of selection of hyperparameters are given in Dua and Ray (1995), Dua and Smyth (1995), Dua and Miller (1996) and Dua, Miller and Smyth (1999).

The BVAR method uses Theil's (1971) mixed estimation technique that supplements data with prior information on the distributions of the coefficients. With each restriction, the number of observations and degrees of freedom artificially increase by one. Thus, the loss of degrees of freedom due to overparameterization does not affect the BVAR model as severely.

Another advantage of the BVAR model is that empirical evidence on comparative out-of-sample forecasting performance generally shows that the BVAR model outperforms the unrestricted VAR model. A few examples are Holden and Broomhead (1990), Artis and Zhang (1990), Dua and Ray (1995), Dua and Miller (1996), Dua, Miller and Smyth (1999).

The above description of the VAR and BVAR models assumes that the variables are stationary. If the variables are nonstationary, they can continue to be specified in levels in a BVAR model because as pointed out by Sims et. al (1990, p.136) '……the Bayesian approach is entirely based on the likelihood function, which has the same Gaussian shape regardless of the presence of nonstationarity, [hence] Bayesian inference need take no special account of nonstationarity'. Furthermore, Dua and Ray (1995) show that the Minnesota prior is appropriate even when the variables are cointegrated.

In the case of a VAR, Sims (1980) and others, e.g. Doan (1992), recommend estimating the VAR in levels even if the variables contain a unit root. The argument against differencing is that it discards information relating to comovements between the variables such as cointegrating relationships. The standard practice in the presence of a cointegrating relaionship between the variables in a VAR is to estimate the VAR in levels or to estimate its error correction representation, the vector error correction model (VECM). If the variables are nonstationary but not cointegrated, the VAR can be estimated in first differences.

The possibility of a cointegrating relationship between the variables is tested using the Johansen and Juselius (1990) methodology as follows.

Consider the p-dimensional vector autoregressive model with Gaussian errors

$$y_t = A_1 y_{t-1} + ... + A_p y_{t-p} + \Psi . D_t + A_0 + \varepsilon_t$$

where y_t is mx1 an vector of I(1) jointly determined variables, D is a vector of deterministic or nonstochastic variables, such as seasonal dummies or time trend. The Johansen test assumes that the variables in y_t are I(1). For testing the hypothesis of cointegration the model is reformulated in the vector error-correction form:

$$\Delta y_t = -\Pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + A_0 + \Psi D_t + \varepsilon_t$$

where

$$\Pi = I - \sum_{i=1}^{p} A_i, \qquad \Gamma_i = -\sum_{j=i+1}^{p} A_j, \qquad i = 1, \ldots\ldots, p-1.$$

Here the rank of Π is equal to the number of independent cointegrating vectors. Thus, if the rank (Π)=0, then the above model will be the usual VAR model in first differences. Similarly, if the vector y_t is I(0), i.e., if all the variables are stationary, then all characteristic roots will be greater than unity and hence Π will be a full rank m x m matrix. If the elements of vector y_t are I(1) and cointegrated with rank (Π)=r, then $\Pi = \alpha\beta'$ where α and β are m x r full column rank matrices and there are $r < m$ linear combinations of y_t. The model can easily be extended to include a vector of exogenous I(1) variables.

Suppose the m characteristic roots of Π are $\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_m$. If the variables in y_t are not cointegrated, the rank of Π is zero and all these characteristic roots will be equal zero. Since ln(1)=0, $\ln(1-\lambda_i)$ will be equal to zero if the variables are not cointegrated. Similarly, if the rank of Π is unity, then if $0 < \lambda_1 < 1$ so that ln $(1-\lambda_1)$ will be negative and $\lambda_i = 0$ $(\forall i > 1)$ so that $\ln(1-\lambda_i) = 0$ $(\forall i > 1)$.

λ_{trace} and λ_{max} tests can be used to test for the number of characteristic roots that are significantly different from unity.

$$\lambda_{trace}(r) = -T \sum_{i=r+1}^{n} \ln\left(1 - \hat{\lambda}_i\right)$$

$$\lambda_{max}(r, r+1) = -T \ln\left(1 - \hat{\lambda}_{r+1}\right)$$

where $\hat{\lambda}_i$ = the estimated values of the characteristic roots of Π

T = the number of usable observations

λ_{trace} tests the null hypothesis that the number of distinct cointegrating vectors is less than or equal to r against a general alternative. If $\lambda_i=0$ for all i, then λ_{trace} equals zero. The further the estimated characteristic roots are from zero, the more negative is $\ln(1-\lambda_i)$ and the larger the λ_{trace} statistic. λ_{max} tests the null that the number of cointegrating vectors is r against the alternative of $r+1$ cointegrating vectors. If the estimated characteristic root is close to zero, λ_{max} will be small. Since λ_{max} test has sharper alternative hypothesis, it is used to select the number of cointegrating vectors in this study.

Under cointegration, the VECM can be represented as

$$\Delta y_t = -\alpha\beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-1} + A_0 + \Psi D_t + \varepsilon_t$$

where α is the matrix of adjustment coefficients. If there are non-zero cointegrating vectors, then some of the elements of α must also be non-zero to keep the elements of y_t from diverging from equilibrium.

The concept of Granger causality can also be tested in the VECM framework. For example, if two variables are cointegrated, i.e. they have a common stochastic trend, causality in the Granger (temporal) sense must exist in at least one direction (Granger, 1986; 1988). Since Granger causality is also a test of whether one variable can improve the forecasting performance of another, it is important to test for it to evaluate the predictive ability of a model.

In a two variable VAR model, assuming the variables to be stationary, we say that the first variable does not Granger cause the second if the lags of the first variable in the VAR are jointly not significantly different from zero. This concept is extended in the framework of a VECM to include the error correction term in addition to lagged variables. Granger-causality can then be tested by (i) the statistical significance of the lagged error correction term by a standard t-test; and (ii) a joint test applied to the significance of the sum of the lags of each explanatory variables, by a joint F or Wald c^2 test. Alternatively, a joint test of all the set of terms described in (i) and (ii) can be conducted by a joint F or a Wald X^2 test. The third option is used in this paper.

III.3. Testing for Nonstationarity

Before estimating any of the above models, the first econometric step is to test if the series are nonstationary or contain a unit root. Several tests have been developed to test for the presence of a unit root. In this study, we focus on the augmented Dickey-Fuller (1979, 1981) test, the Phillips-Perron (1988) test and the KPSS test proposed by Kwiatkowski et al. (1992).

To test if a sequence y_t contains a unit root, three different regression equations are considered.

$$\Delta y_t = \alpha + \gamma y_{t-1} + \theta t + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \qquad (1)$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \qquad (2)$$

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \qquad (3)$$

The first equation includes both a drift term and a deterministic trend; the second excludes the deterministic trend; and the third does not contain an intercept or a trend term. In all three equations, the parameter of interest is *y*. If $y = 0$ the y_t sequence has a unit root. The estimated t-statistic is compared with the appropriate critical value in the Dickey-Fuller tables to determine if the null hypothesis is valid. The critical values are denoted by ^{t}t, ^{t}m and t for equations (1), (2) and (3), respectively.

Following Doldado, Jenkinson and Sosvilla-Rivero (1990), a sequential procedure is used to test for the presence of a unit root when the form of the data-generating process is unknown. Such a procedure is necessary since including the intercept and trend term reduces the degrees of freedom and the power of the test implying that we may conclude that a unit root is present when, in fact, this is not true. Further, additional regressors increase the absolute value of the critical value making it harder to reject the null hypothesis. On the other hand, inappropriately omitting the deterministic terms can cause the power of the test to go to zero (Campbell and Perron, 1991).

The sequential procedure involves testing the most general model first (equation 1). Since the power of the test is low, if we reject the null hypothesis, we stop at this stage and conclude that there is no unit root. If we do not reject the null hypothesis, we proceed to determine if the trend term is significant under the null of a unit root. If the trend is significant, we retest for the presence of a unit root using the standardised normal distribution. If the null of a unit root is not rejected, we conclude that the series contains a unit root. Otherwise, it does not. If the trend is not significant, we estimate equation (2) and test for the presence of a unit root. If the null of a unit root is rejected, we conclude that there is no unit root and stop at this point. If the null is not rejected, we test for the significance of the drift term in the presence of a unit root. If the drift term is significant, we test for a unit root using the standardised normal distribution. If the drift is not significant, we estimate equation (3) and test for a unit root.

We also conduct the Phillips-Perron (1988) test for a unit root. This is because the Dickey-Fuller tests require that the error term be serially uncorrelated and homogeneous while the Phillips-Perron test is valid even if the disturbances are serially correlated and heterogeneous. The test

statistics for the Phillips-Perron test are modifications of the t-statistics employed for the Dickey-Fuller tests but the critical values are precisely those used for the Dickey-Fuller tests. In general, PP test is preferred to the ADF tests if the diagnostic statistics from the ADF regressions indicate autocorrelation or heteroscedasticity in the error terms. Phillips and Perron (1988) also show that when the disturbance term has a positive moving average component, the power of the ADF tests is low compared to the Phillips-Perron statistics so that the latter is preferred. If, however, a negative moving average term is present in the error term, the PP test tends to reject the null of a unit root and, therefore, ADF tests are preferred.

In both the ADF and the PP test, the unit root is the null hypothesis. A problem with classical hypothesis testing is that it ensures that the null hypothesis is not rejected unless there is strong evidence against it. Therefore these tests tend to have low power, that is, these tests will often indicate that a series contains a unit root. Kwiatkowski et al. (1992), therefore, suggest that, based on classical methods, it may be useful to perform tests of the null hypothesis of stationarity in addition to tests of the null hypothesis of a unit root. Tests based on stationarity as the null can then be used for confirmatory analysis, that is, to confirm conclusions about unit roots. Of course, if tests with stationarity as the null as well as tests with unit root as the null, both reject or fail to reject the respective nulls, there is no confirmation of stationarity or nonstationarity.

KPSS test with the null hypothesis of difference stationarity

To test for difference stationarity (DS), KPSS assume that the series y_t with T observations (t=1,2,…,T) can be decomposed into the sum of a deterministic trend, random walk and stationary error:

$$y_t = \delta t + r_t + \varepsilon_t$$

where r_t is a random walk

$$r_t = r_{t-1} + \mu_t$$

and μ_t is independently and identically distributed with mean zero and variance σ^2_μ. The initial value r_0 is fixed and serves the role of an intercept. The stationarity hypothesis is $\sigma^2_\mu=0$. If we set $\delta = 0$, then under the null hypothesis y_t is stationary around a level (r_0).

Let the residuals from the regression of y_t on an intercept be e_t, t=1,2,....,T. The partial sum process of the residuals is defined as:

$$S_t = \sum_{i=1}^{t} e_t.$$

The long run variance of the partial error process is defined by KPSS as

$$\sigma^2 = \lim_{T\to\infty} T^{-1}E(S^2_T).$$

A consistent estimator of σ^2, $s^2(l)$, can be constructed from the residuals e_t as

$$s^2(l) = T^{-1}\sum_{t=1}^{T} e^2_t + 2T^{-1}\sum_{s=1}^{l} w(s,l) \sum_{t=s+1}^{T} e_t e_{t-s}$$

where $w(s,l)$ is an optional lag window that corresponds to the selection of a spectral window. KPSS employ the Bartlett window, $w(s,l) = 1-s/(l+1)$ as in Newey and West (1987), which ensures the non-negativity of $s^2(l)$. The lag operator l corrects for residual

serial correlation. If the residual series are independently and identically distributed, a choice of $l = 0$ is appropriate.

The test statistic for the DS null hypothesis is

$$\hat{\eta}_{\mu} = T^{-2}\sum_{t=1}^{T} S^2_t/s^2(l).$$

KPSS report the critical values of $\hat{\eta}_{\mu}$ (p. 166) for the upper tail test.

Thus, three tests, augmented Dickey-Fuller, Phillips Perron and KPSS tests, are used to test for the presence of a unit root. The KPSS test, with the null of stationarity, helps to resolve conflicts between ADF and PP tests. If two of these three tests indicate nonstationarity for any series, we conclude that the series has a unit root.

In sum, the study proceeds as follows. First, the series are tested for the presence of a unit root using the augmented Dickey-Fuller, Phillips Perron and KPSS tests. If the interest rate series are nonstationary, univariate models, i.e. ARIMA without and with ARCH-GARCH effects, are fitted to differenced, stationary series.

Multivariate models include VAR, VECM, and BVAR models. To estimate VAR models, if all the variables are nonstationary and integrated of the same order, the Johansen test is conducted for the presence of cointegration. If a cointegrating relationship exists, the VAR model can be estimated in levels. Tests for Granger causality are also conducted in the VECM framework to evaluate the forecasting ability of the model. Lastly, Bayesian vector autoregressive models are estimated that impose prior beliefs on the relationships between different variables as well as between own lags of a particular variable. If these beliefs (restrictions) are appropriate, the forecasting ability of the model should improve.

The forecasting ability of each model is evaluated by examining the performance of out-of-sample forecasts and the "best" forecasting model is selected.

III.4. Evaluation of Forecasting Models

The "best" forecasting model is one that produces the most accurate forecasts. This means that the predicted levels should be close to the actual realized values. Furthermore, the predicted variables should move in the same direction as the actual series. In other words, if a series is rising (falling), the forecasts should reflect the same direction of change. If a series is changing direction, the forecasts should identify this.

To select the best model, the alternative models are initially estimated using weekly data over the period April 1997 to December 2001 and tested for out-of-sample forecast accuracy from January 2002 to September 2002. In other words, by continuously updating and reestimating, we conduct a real world forecasting exercise to see how the models perform. The model that produces the most accurate one- through thirty-six-week-ahead forecasts is designated the "best" model for a particular interest rate.

To test for accuracy, the Theil coefficient (Theil, 1966), is used that implicitly incorporates the naïve forecasts as the benchmark. If A_{t+n} denotes the actual value of a variable in period (t+n), and ${}_tF_{t+n}$ the forecast made in period t for (t+n), then for T observations, the Theil U-statistic is defined as follows:

$$U = [\Sigma(A_{t+n} - {}_tF_{t+n})^2/\Sigma(A_{t+n} - A_t)^2]^{0.5}.$$

The U-statistic measures the ratio of the root mean square error (RMSE) of the model forecasts to the RMSE of naive, no-change forecasts (forecasts such that ${}_tF_{t+n} = A_t$). The RMSE is given by the following formula:

$$RMSE = [\Sigma(A_{t+n} - {}_tF_{t+n})^2/T]^{0.5}.$$

A comparison with the naïve model is, therefore, implicit in the U-statistic. A U-statistic of 1 indicates that the model forecasts match the performance of naïve, no-change forecasts. A U-statistic >1 shows that the naïve forecasts outperform the model forecasts. If U is < 1, the forecasts from the model outperform the naïve forecasts. The U-statistic is, therefore, a relative measure of accuracy and is unit-free.

Since the U-statistic is a relative measure, it is affected by the accuracy of the naïve forecasts. Extremely inaccurate naïve forecasts can yield U <1, falsely implying that the model forecasts are accurate. This problem is especially applicable to series with trend. The RMSE, therefore, provides a check on the U-statistic and is also reported.

To evaluate the forecast performance, the models are continually updated and reestimated. The models are estimated for the initial period April 1997 through December 2001. Forecasts for up to 36-weeks-ahead are computed. One more observation is added to the sample and forecasts up to 36-weeks-ahead are again generated, and so on. Based on the out-of-sample forecasts for the period January through September 2002, the Theil U-statistics and RMSE are computed for one- to 36-weeks-ahead forecasts and the average of successive four U-statistics and RMSE are also computed. The overall average of the U statistic and the RMSE for up to 36-weeks-ahead forecasts is also calculated to gauge the accuracy of a model.

---

[1] Fauvel, Paquet and Zimmermann (1999) provide a survey of major methods used to forecast interest rates as well as a review of interest rate modelling. Examples of studies that examine forecasting of interest rates are as follows: Ang and Bekaert (1998); Barkoulas and Baum (1997); Bidarkota (1998); Campbell and Shiller (1991); Chiang and Kahl (1991); Cole and Reichenstein (1994); Craine and Havenner.(1988); Deaves (1996); Dua (1988); Froot (1989); Gosnell (1997); Gray (1996); Hafer, Hein and MacDonald (1992); Holden and Thompson (1996); Iyer and Andrews (1999); Jondeau and Sedillot (1999); Jorion and Mishkin (1991); Kolb and Stekler (1996); Park and Switzer (1997); Pesando (1981); Prell (1973); Roley (1982); Sola and Driffil (1994); and Throop (1981).