25 November 2005

BM-3-6-4

TREASURY'S FORECASTING PERFORMANCE

Introduction

This report is part of a regular process of checking on the performance of forecasting in The Treasury.

In each of the past two years, we have published comprehensive reports on our economic and fiscal forecasting records. As trends in forecast performance tend to evolve slowly over long periods of time, we do not intend to publish a comprehensive report every year.

This report updates the data tables presented in last year's report and provides some additional commentary as necessary. For more comprehensive analysis, refer to the November 2004 report at http://www.treasury.govt.nz/forecasts/performance/ .

Key Results

The key results are largely unchanged from last year's report.

Economic forecasts

- We fail to reject that the forecasts of GDP growth (real and nominal) and CPI inflation are unbiased over the period under examination when each sample is treated as a cross-section, i.e. on average, they tend to be neither too low nor too high. The accuracy measures of these forecasts, e.g. the root mean square errors, compare favourably with the accuracy measures of forecasts made by other forecasting agencies in New Zealand.

Tax forecasts

- We also fail to reject that the forecasts of tax revenue and receipts are unbiased over the period under examination.

- However, one-year-ahead tax forecasts have been too low for the last six years in a row, as have forecasts of nominal GDP. Forecasts in other jurisdictions that we have seen reviewed also show similar tendencies to underforecast in an economic upturn and overforecast in a downturn. This is not an easy problem to correct. Nevertheless, we are looking into ways to remedy the situation.

Fiscal balance

- Adjusting for items that, as a matter of policy, are not explicitly forecast, Crown operating balance forecasts also appear to be unbiased.

Net cash flows

- Current-year forecasts of the net movement in cash tend to be too low. This looks to be chiefly the result of forecasts of operating expenditures, benefit expenditures and physical asset purchases tending to be too high (as actual spending tends to occur later than originally planned).

How can we analyse forecast performance?

Essentially, there are two desirable properties that we would like the forecasts to have. We would like the average forecast errors, i.e. the difference between the actual outcomes and the forecasts, to be zero, which would indicate that, on average, the forecasts are neither too low nor too high. We would also like the variance of the forecast errors to be as small as possible, which would indicate that, on average, the forecasts stay within some relatively small distance of the actual outcomes.

To test the first property, we examine the average of the forecast errors over time and test if this average is 'significantly different' from zero. If so, we would conclude that, in all likelihood, that particular forecast is biased, tending to be either too high or too low, on average. If the average is not significantly different from zero, then we conclude that we do not have enough evidence to suggest that the forecast is biased, which is not quite the same as saying that the forecast is unbiased.

To test the second property, we examine measures of the magnitude of the errors such as mean absolute error (MAE) and root mean square error (RMSE). These measures are described in the report's annex.

Caveats

The analysis covers the period 1991 to 2005 (March years) for the economic forecasts, and 1994 to 2005 (June years) for the fiscal forecasts. We started from 1991 for the economic forecasts as this is the earliest year for which we have reliable historical forecast data and we start from 1994 for the fiscal forecasts as 1994 marked the introduction of accrual accounting for the Crown accounts. This yields relatively small samples of observations and so we are unable to get precise estimates of the mean forecasting errors.

Some of the analysis suggests that there may be correlation of forecasting errors over time. That is, we get runs of persistent under-forecasting or over-forecasting, indicating that a positive (or negative) forecasting error in one year may increase the likelihood of a positive (or negative) forecasting error in the following year. We do not yet have sufficient observations to enable us to reliably test for correlations of errors over time.

Because of these two problems caused by the relatively short data series, it is difficult to draw strong conclusions from the analysis.

A key qualification to Treasury's forecasts is that they assume no change in Government policy and make technical assumptions about the long-run equilibrium values of certain variables (e.g. the exchange rate). This means that the forecasts will not capture all information relevant to future economic and fiscal outcomes.

It is also important to realise that the forecasting models and methods have changed over time as science and technology have advanced; we now have more sophisticated forecasting techniques than we did ten years ago, and desktop computing power has also increased dramatically. Thus, we are not testing the performance of a particular forecasting model, but of an evolving forecast process. This will have had some effect on the forecasting errors, but we are unable to quantify any such effects.

The analysis is restricted to forecast performance and does not consider any policy implications.

Analysis of the Economic Forecasts

This section provides measures of Treasury's forecasts of New Zealand's real GDP growth, CPI inflation and nominal GDP growth. Although many other economic variables are also forecast, the focus is on output and inflation because they are the key economic variables of interest and also have a major influence on the fiscal forecasts. Treating the Budget Economic and Fiscal Updates (BEFUs) and December Economic and Fiscal Updates (DEFUs) separately, there are generally 15 data points for the economic forecasts and 12 data points for the fiscal forecasts.

Real GDP growth, CPI inflation and nominal GDP growth forecasts

Table 1 – Performance measures of Treasury's GDP and CPI forecasts

Update	Years ahead	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
Real GDP Growth						
BEFUs	Year-ahead	21-24 mths	-0.31	1.59	2.00	0.63
DEFUs	Year-ahead	16-19 mths	-0.10	1.89	2.15	0.85
BEFUs	Current	9-12 mths	0.01	1.31	1.51	0.55
DEFUs	Current	4-7 mths	0.35	0.69	0.97	0.47
CPI Inflation						
BEFUs	Year-ahead	21-24 mths	0.40	0.80	0.97	0.64
DEFUs	Year-ahead	16-19 mths	0.16	0.80	1.04	0.69
BEFUs	Current	9-12 mths	0.14	0.50	0.68	0.56
DEFUs	Current	4-7 mths	-0.23	0.53	0.70	0.61
Nominal GDP (Expenditure) Growth						
BEFUs	Year-ahead	21-24 mths	-0.48	2.15	2.57	0.65
DEFUs	Year-ahead	16-19 mths	-0.49	2.16	2.83	0.77
BEFUs	Current	9-12 mths	-0.29	1.50	1.82	0.48
DEFUs	Current	4-7 mths	-0.05	0.98	1.10	0.38

Note: The sample period is from the year ended March 1991 to the year ended March 2005. For the years 1993 and 1999, the PREFU (pre-election) forecasts were used in place of the DEFU. Forecast horizon refers to the number of months between the forecasts being finalised and the period the forecast relates to.

Over all the forecast horizons examined, the mean errors are not statistically different from zero (at the 5% significance level), indicating that any given GDP growth or CPI inflation forecast is just as likely to be too high as too low.

The magnitude of the errors generally declines as the forecast horizon gets shorter, which is not surprising given that accuracy is expected to improve the closer we get to the period the

forecast relates to. All the U-coefficient values are less than one, suggesting that the Treasury forecasts perform better on average than a naive model where the naive forecast is the previous observation.

Knowing that Treasury's economic forecasts are superior to a naive model provides somewhat limited comfort. Another benchmark can be obtained by comparing Treasury forecasts against the NZIER[1] Consensus Forecasts. Since the early 1990s, the NZIER has been surveying financial and economic agencies on their forecasts of key economic variables for the next two to three years. The forecasts are averaged and published every quarter.[2] To ensure comparability between the NZIER Consensus Forecasts and Treasury forecasts, the sample period is for the year ended March 1993 to the year ended March 2005. Because of the lag between preparation and publication of Treasury's economic forecasts, Consensus forecasts published in the same quarter as Treasury's forecasts are based on more recent data. To ensure that, as far as possible, we are comparing forecasts that are based on the same information set, we compare Treasury forecasts with NZIER Consensus forecasts that are published at the end of the quarter prior to that in which the Treasury forecasts are published.

Table 2 – Treasury forecasts and NZIER Consensus Forecasts (one quarter prior) - RMSE

			Real GDP Growth		CPI Inflation	
Update	**Years ahead**	**Forecast Horizon**	**Treasury**	**Consensus**	**Treasury**	**Consensus**
BEFUs	Year-ahead	21-24 mths	1.89	1.90	1.12	1.00
DEFUs	Year-ahead	16-19 mths	2.12	1.89	1.28	1.06
BEFUs	Current	9-12 mths	1.48	1.53	1.04	1.10
DEFUs	Current	4-7 mths	1.03	1.08	0.62	0.58

The root-mean-square errors of Consensus real GDP growth forecasts are reasonably close to those of Treasury (perhaps with the exception of DEFU year-ahead forecasts). For CPI inflation forecasts, the Consensus RMSE is lower than that of Treasury for all horizons other than current-year BEFU forecasts. These results are unchanged from last year's report.

[1] New Zealand Institute of Economic Research.

[2] The NZIER Consensus Forecast sample includes forecasts made by Treasury. This means they do not represent a strictly "independent" comparator. Averaging across a number of forecasters reduces the materiality of this issue. NZIER Consensus Forecasts may be found at: http://www.nzier.org.nz/SITE_Default/SITE_Publications/consensus_forecasts/default.asp .

Analysis of the Fiscal Forecasts

The sample period for the analysis of fiscal forecasts starts in the year to June 1994, the year in which Treasury first started publishing results on an accrual basis, and ends with June 2005. The analysis begins with the accrual measures before addressing cash flows (where the latter ultimately feeds through to the debt forecasts). Results are presented in percentage terms (percentage of average outturn) to provide an indication of the relative size of the average forecast error.

Tax revenue forecasts

The tax revenue component of the operating balance has the most direct link to the economic forecasts discussed above.

Table 3 – Performance measures of Treasury's tax revenue forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	2.4%	3.9%	4.9%	0.43
1-year-ahead BEFUs	14 mths	1.7%	2.9%	3.6%	0.31
Current year DEFUs	8 mths	0.8%	1.7%	1.9%	0.30
Current year BEFUs	2 mths	0.1%	0.7%	0.9%	0.13

Although there is not enough evidence to suggest that the tax revenue forecasts are biased[3] over the 12 years under examination, the forecasts have been too low for the last six consecutive years over all of the above forecast horizons. We are taking steps to address this situation.

We have conducted reviews, both internal to Treasury and external, of Treasury's tax forecasting process. As a result of these reviews, some changes have been made and more are to follow, including:

- closer integration of the macroeconomic and tax forecasting teams;
- making better use of specialised forecasting software;
- more focus on research and development, particularly with respect to the interaction of tax flows and economic growth; and
- instituting some further quality controls.

[3] For instance, to establish that the forecasts are biased, at the 5% confidence level, the mean error for current year BEFUs would need to be at least 0.5% (positive or negative), rising to 2.7% for 1-year-ahead DEFUs.

Total revenue forecasts

Total Crown revenue is the sum of tax revenue; levies, fees, fines and penalties; sales of goods and services; investment income; and other income.

Table 4 – Performance measures of Treasury's total Crown revenue forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	3.7%*	4.3%	5.5%	0.47
1-year-ahead BEFUs	14 mths	3.4%*	3.7%	4.5%	0.35
Current year DEFUs	8 mths	2.1%*	2.7%	3.2%	0.46
Current year BEFUs	2 mths	0.9%	1.3%	2.1%	0.29

Note: Mean errors marked with an asterisk were found to be significantly different from zero at the five percent significance level.

Forecasts for 1-year-ahead DEFUs, 1-year-ahead BEFUs and current year DEFUs all tend to be too low. This bias seems to be mainly coming through from the forecasts of investment income and other revenue. When calculating the forecast performance statistics for the overall Crown operating balance (below), we exclude investment income as there are some relatively unpredictable elements of investment income that are not forecast as a matter of policy. Over time, these elements have tended to produce positive surprises.

Expense forecasts

Table 5 – Performance measures of Treasury's expense forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	1.7%	3.2%	4.2%	0.51
1-year-ahead BEFUs	14 mths	1.0%	2.6%	4.0%	0.45
Current year DEFUs	8 mths	0.9%	1.8%	3.0%	0.39
Current year BEFUs	2 mths	0.6%	1.4%	2.0%	0.25

The mean expense errors are not statistically different from zero over any of the forecast horizons considered. The absolute error measures are not large and are comparable in magnitude to the errors on the tax revenue forecasts.

Fiscal balance forecasts

To assess the performance of the forecasts of the net components of the operating balance, we assess forecasts for a fiscal balance that excludes investment income, thereby removing the non-forecast items from our analysis. This approach is consistent with international practice. This "adjusted" operating balance is close to, but not exactly the same as, the OBERAC (Operating Balance Excluding Revaluations and Accounting policy Changes).

Non-forecast items are defined as either those items that we did not know about, and could not possibly have known about, at the time of the forecast *or* items that we knew would happen but could not quantify at the time of forecasting. Examples of non-forecast items include Government Superannuation Fund (GSF) pension liability movements, foreign exchange gains and losses, and gains and losses on assets sales.

Table 6 – Performance measures of "adjusted" operating balance forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	-8.2%	101%	129%	0.95
1-year-ahead BEFUs	14 mths	22.0%	86%	110%	0.74
Current year DEFUs	8 mths	27.1%	75%	85%	0.42
Current year BEFUs	2 mths	-2.5%	55%	69%	0.32

All of the error statistics are quite large, especially when compared to the results for total tax revenues and total expenses, for example. This is because the adjusted operating balance is mostly the relatively small difference between two very large numbers (tax revenues and expenses). Since operating balances tend to be relatively small, comparing error statistics with the average operating balance can produce some quite large percentages.

A more meaningful way of presenting the statistics is to express them as a percentage of total revenues, as in Table 7 below. This is consistent with overseas practice and also assists with comparisons between jurisdictions.

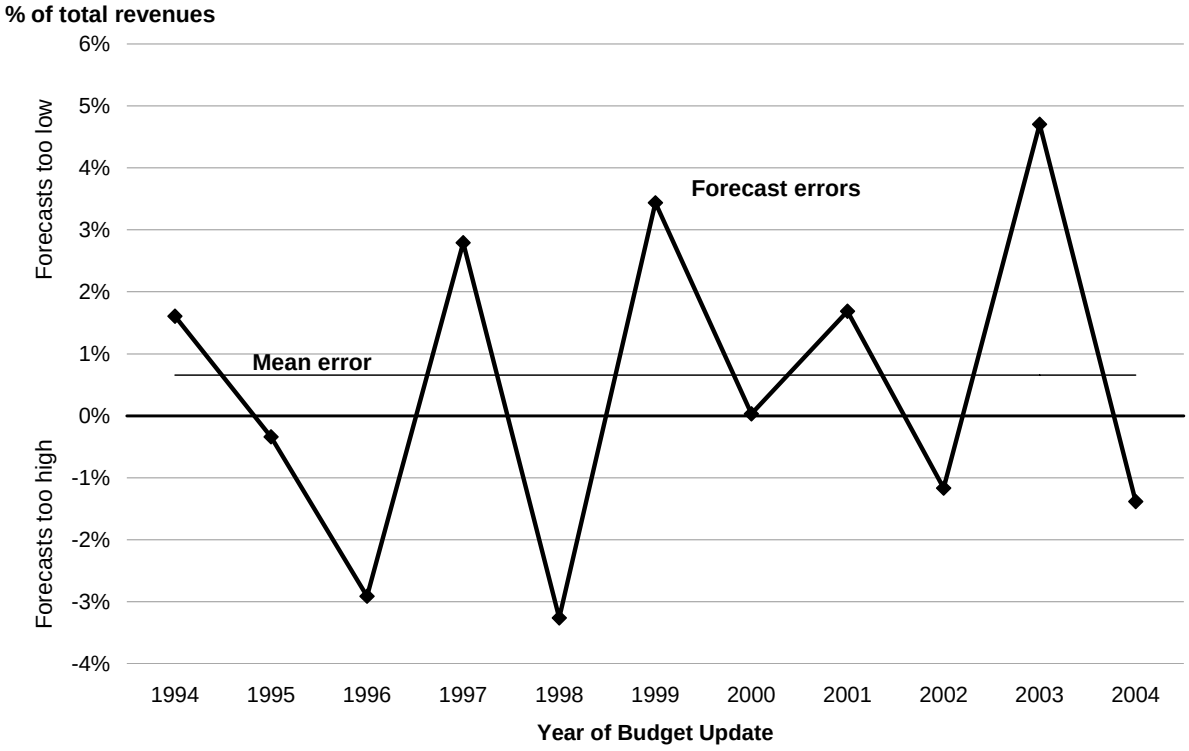
Table 7 – Performance measures of "adjusted" operating balance forecasts, expressed as a percentage of total revenues

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
1-year-ahead DEFUs	20 mths	-0.4%	2.7%	3.4%
1-year-ahead BEFUs	14 mths	0.5%	2.1%	2.5%
Current year DEFUs	8 mths	0.7%	1.8%	2.1%
Current year BEFUs	2 mths	-0.2%	1.2%	1.4%

The mean forecast errors for this adjusted operating balance are not statistically different from zero. The mean absolute errors indicate that forecasts over these horizons tend to be out by between 1% and 3% of revenues, on average. To put this in context, total crown revenue for the year to June 2005 was $67 billion.

Figure 1 shows the forecast errors for one-year-ahead Budget Update forecasts of the adjusted operating balance going back as far as the 1992 Budget. So, the observation for 2004 is the 2004 BEFU one-year-ahead forecast error i.e. the error for the year ended June 2005, the most recently-completed fiscal year for which we have final data. Also shown is the mean of these errors, about 0.7% of the revenue outturn. The 95%-confidence interval includes zero, which leads us to the conclusion that forecasts of the adjusted operating balance are just as likely to be too low as too high.

Figure 1 – Forecast errors for one-year-ahead Budget Update forecasts of the adjusted operating balance



Cash fiscal forecasts

The total cash flow balance (as distinct from net cash flow from operations) can be expressed as:

Cash provided from operations (mostly tax receipts)
less **Cash disbursed to operations**
plus **net cash flows from investing activities**
plus **net cash flows from financing activities.**

All of these components, with the exception of financing activities, are forecast separately and combined to produce a forecast for cash flow. Net cash flows from investing activities are dominated by the purchase of physical assets. Financing activities (e.g. issuance of currency and repayment of debt) are not forecast as they are either beyond our control or are effectively balancing items resulting from Debt Management Office (DMO) activity.

Tax receipt forecasts

Table 8 – Performance measures of Treasury's tax receipt forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	2.0%	3.4%	4.4%	0.39
1-year-ahead BEFUs	14 mths	1.8%	2.7%	3.5%	0.31
Current year DEFUs	8 mths	0.9%	1.7%	1.9%	0.29
Current year BEFUs	2 mths	0.3%	0.5%	0.8%	0.11

Mean tax receipt errors are not significantly different from zero and the absolute errors gradually reduce as the forecast horizon shortens. Adjusting forecasts to allow for policy change makes only small differences to the performance statistics.

Operating expenditure forecasts

This line item forms the major part of *Cash disbursed to operations*. It is principally the Crown's personnel and operating payments.

Table 9 – Performance measures of Treasury's operating expenditure forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	4.7%*	4.7%	5.3%	0.43
1-year-ahead BEFUs	14 mths	0.9%	1.3%	1.6%	0.13
Current year DEFUs	8 mths	-1.4%*	1.4%	1.8%	0.27
Current year BEFUs	2 mths	-0.6%	1.5%	2.0%	0.31

Note: Mean errors marked with an asterisk were found to be significantly different from zero at the five percent significance level.

The 1-year-ahead DEFU forecasts tend to be too low, whereas the current year DEFU forecasts tend to be too high. In last year's analysis, current year BEFU forecasts also tended to be too high. This is no longer the case, thanks to personnel and operating payments coming in $1 billion above forecast in the year to June 2005.

Benefit expenditure forecasts

Benefit expenditures are the second biggest contributor (after operating expenditures) to *Cash disbursed to operations.*

Table 10 – Performance measures of Treasury's benefit expenditure forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	-0.9%	2.6%	3.3%	0.55
1-year-ahead BEFUs	14 mths	-1.6%*	1.7%	2.5%	0.47
Current year DEFUs	8 mths	-1.5%*	1.6%	2.2%	0.71
Current year BEFUs	2 mths	-1.3%*	1.3%	1.8%	0.66

Note: Mean errors marked with an asterisk were found to be significantly different from zero at the five percent significance level.

The current-year forecasts are significantly different from zero being, on average, too high. Absolute errors are not large in relative terms and generally decline as the forecast horizon shortens.

As with last year's analysis, the forecast bias is mostly coming through via the Family Support forecasts. See last year's report for further detail.

Purchase of physical asset forecasts

The purchase of physical assets is usually the largest item in *Investment activities*.

Table 11 – Performance measures of Treasury's purchase of physical assets forecasts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	10.5%*	11.2%	16.9%	0.59
1-year-ahead BEFUs	14 mths	-7.0%	10.6%	12.5%	0.43
Current year DEFUs	8 mths	-16.3%*	16.3%	17.5%	0.88
Current year BEFUs	2 mths	-12.1%*	12.6%	15.9%	0.79

Note: Mean errors marked with an asterisk were found to be significantly different from zero at the five percent significance level.

For most of the forecast horizons, mean forecast errors look likely to be non-zero. Generally, the forecast errors do not reduce as we get closer to the outturn. The 'most accurate' forecasts appear to be the ones made about one year out.

Net movement in cash

Table 12 – Performance measures of Treasury's total cash movement forecasts, expressed as a percentage of total receipts

Update	Forecast Horizon	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Theil's U-Statistic (U)
1-year-ahead DEFUs	20 mths	0.2%	0.4%	0.6%	1.03
1-year-ahead BEFUs	14 mths	0.3%	0.5%	0.6%	1.08
Current year DEFUs	8 mths	0.8%*	0.8%	1.0%	1.54
Current year BEFUs	2 mths	0.7%*	0.7%	0.8%	1.46

Note: Mean errors marked with an asterisk were found to be significantly different from zero at the five percent significance level.

As with the operating balance analysis, we have expressed cash flow balance forecast errors against income, in this case, total receipts.

Although the mean errors for one-year-ahead forecasts are not significantly different from zero, the current year forecasts tend to be too low. The mean errors do not decrease as we get closer to the outturn. Indeed, all of the performance statistics generally *increase* as the forecast horizon shortens. The U statistics are all greater than one, indicating that using the last known actual as the forecast may be a more accurate way of forecasting than the current method.

The total cash flow balance outturn is typically a fairly small number; it has averaged less than $200 million over the period we are looking at. As with the accrual fiscal balance, the total cash flow balance is mostly the relatively small difference between two very large numbers (receipts and expenditure). Small forecasting errors in the large components can lead to relatively large errors in the forecasting of the net cash movement, although they look small in Table 12 as they are expressed as percentage of total receipts.

Annex: Details of the Measures Used

Summary measures of forecast performance and their interpretation

The forecasts are analysed using four measures commonly used in the forecast performance literature. Forecast "error" is defined as the actual outturn for a particular period less the forecast. The four measures can be summarised as follows:

Mean Error (ME)

$$ME = \frac{1}{n}\sum_{t=1}^{n}\left(A_t - F_t\right)$$

where: n = number of observations
F_t = forecast for period t
A_t = actual outturn for period t

A positive ME indicates that the forecasts are, on average, too low and a negative ME indicates that they are, on average, too high.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|A_t - F_t\right|$$

The MAE is the average of the forecast errors disregarding their sign.

Root Mean Square Error (MAE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(A_t - F_t\right)^2}$$

The RMSE is computed by taking the square root of the average squared errors. It penalises large forecast errors relative to smaller errors.

Theil's U coefficient (U)

$$(4) \qquad U = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(A_t - F_t\right)^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(A_t - A_{t-1}\right)^2}}$$

U is computed by taking the ratio of the RMSE from the set of forecasts being evaluated to the RMSE from a naive forecast of no change. This measure offers a way of evaluating forecasting performance relative to an alternative or benchmark model, in this case a naive model. A Theil's U coefficient value of less than one indicates that the forecast is more accurate than a naive model, which in this case is the previous observation.

Forecast bias

We also conducted the following *t*-test to assess whether the forecast errors have a mean of zero:

$$t = \frac{ME}{\sqrt{s^2/n}}$$

where: s^2 = sample variance

For an ideal forecasting model, the expected mean error would be zero. Having errors that are close to zero, on average, suggests that the forecasts are not biased in the sense that they are not “one-sided”. Given a sample of forecast errors, we want to determine whether the sample is consistent with coming from a "population" of errors with zero mean. To test for this, we need to know what the probability distribution of sample means from a population with a true mean of zero would look like (for a given sample size). By the central limit theorem, we can say that if our sample is large enough[4], the mean of samples from a population with zero mean would be ~N(0,s^2/n), i.e. normally distributed with a mean of zero and a variance of s^2/n. We then look at where our sample mean falls in this probability distribution and see how likely it is that our sample came from this population. If the probability that our sample came from this distribution is less than a significance level that we choose, we reject that the true mean is zero. If this is the case, we conclude that the model is likely to be *biased*, that is, it tends to under/over-forecast. If the probability is greater than this significance level, we do not think that we have enough evidence to reject that the model is unbiased.

We have chosen to use a small significance level of 0.05. This is the probability of rejecting that the true mean is zero when it actually is. There is a trade-off between this and failing to reject that the true mean error is zero when it isn't. We have chosen to use a low significance level because it is costly to change forecasting techniques and because we have quite a small sample size. We would want to be reasonably sure that a model is biased before adjusting it.

[4] The central limit theorem states that the minimum sample size needed to assume normality is 30.