FORECASTING ECONOMIC ACTIVITY USING THE YIELD CURVE: QUASI-REAL-TIME APPLICATIONS FOR NEW ZEALAND, AUSTRALIA AND THE US

By

Todd Henry and Peter C. B. Phillips

October 2020

COWLES FOUNDATION DISCUSSION PAPER NO. 2259

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Forecasting Economic Activity Using the Yield Curve: Quasi-Real-Time Applications for New Zealand, Australia and the US

Todd Henry[1] and Peter C. B. Phillips[2] *

[1]University of Auckland

[2]Yale University, University of Auckland, University of Southampton, Singapore Management University

14 October, 2020

Abstract

Inversion of the yield curve has come to be viewed as a leading recession indicator. Unsurprisingly, some recent instances of inversion have attracted attention from economic commentators and policymakers about possible impending recessions. Using a variety of time series models and recent innovations in econometric method, this paper conducts quasi-real-time forecasting exercises to investigate whether the predictive capability of the yield curve extends to forecasting economic activity in general and whether removing the term premium component from yields affects forecast accuracy. The empirical findings for the US, Australia, and New Zealand show that forecast performance is not improved either by augmenting simplistic models with information from the yield curve or by making such a decomposition of yields. Results from similar research exercises in previous work in the literature are mixed. The results of the present analysis suggest possible explanations that reconcile these conflicting results.

*Research support from the NSF under Grant No. SES 18-50860 and the Kelly Fund at the University of Auckland is gratefully acknowledged

1 Introduction

In recent times an inverted US Treasury yield curve (hereafter referred to as the yield curve) has come to be seen as a leading recession indicator. Indeed, in the US the last nine recessions were each preceded by the morning star of an atypical "inversion" of the yield curve (Bauer and Merterns (2018)). Curve plots of the yields of US Treasury securities against their respective maturities are shown in Figure 1 illustrating typical and atypical shapes.

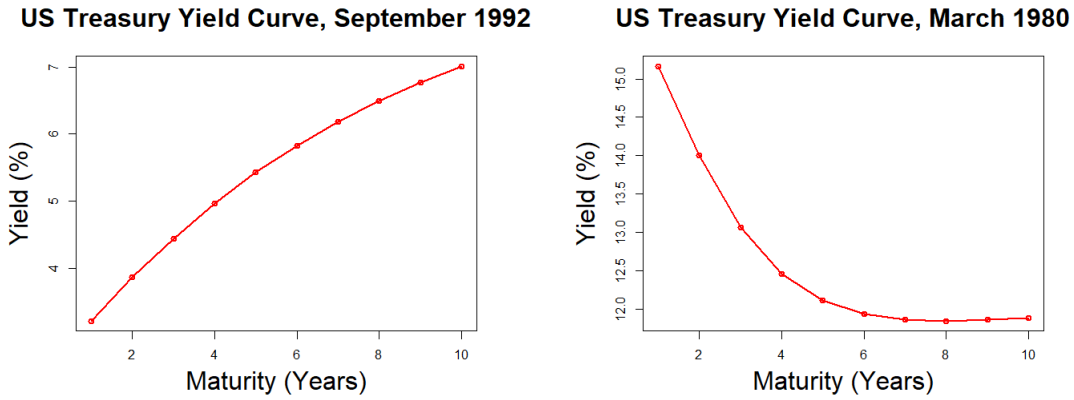

Figure 1: Left panel: a typical upward sloping yield curve following the 1990-1991 US recession. Right panel: an atypical downward sloping 'inverted' yield curve prior to the 1981 US recession. Data source: FRBNY (2013)

An inverted yield curve refers to a scenario where the yield curve slopes downwards. It is commonly believed that this phenomenon may signal an impending recession due to the relationship between short-term, and long-term yields. According to the Pure Expectation Hypothesis (Williams (1938), Hicks (1946), Meiselman (1962)), the price a risk-neutral investor should be willing to pay (or the yield they will accept) for each asset should lead to a balancing equality of the form $(1 + r_l)^n = \prod_{i=1}^{n}(1 + r_{si})$ in which r_l denotes the annual return to the asset with the longer maturity of n years, and r_{si} denotes the expected annual return on assets with a one-year maturity in year i. Intuitively, the expected return on each asset should be equivalent, otherwise the risk-neutral investor would not purchase the asset with the lower expected return.

In the face of an economic slowdown or contraction central banks typically aim to stimulate economic activity through lower policy rates or in severe instances resort to unconventional monetary policy instruments aimed to bring down the cost of borrowing (and thus, the return on saving / investing) in order to increase consumer spending and firm investment activity. If investors anticipate such an event and expect lower short-term yields in future, they will accept relatively lower yields on long-term assets today, all else equal. With a reduction in yields on assets with longer maturities, an upward sloping yield curve flattens

or even inverts. Accordingly, the shape of the yield curve reflects some information about market participant expectations of future events, as recorded in the yields associated with such events. An inverted yield curve may indicate that market participants expect a recession in the foreseeable future, which explains its potential predictive power of an impending recession.

While an inverted yield curve may signal that market participants anticipate a forthcoming economic downturn, the signal is by no means completely reliable. Not all investors are risk-neutral. So long-term yields may reflect more than just investor expectations of short-term yields. Issues such as a premium for re-investment risk, inflation risk, or illiquidity, among other things do figure in decision making. A popular view is that long-term yields comprise two components: an expectations component that reflects expectations about future short-term yields; and a term premium component that reflects everything else.

These considerations raise questions that we seek to address in the current paper. First, if the yield curve has some predictive power as a recession indicator, does the predictive power extend to forecasting economic activity in general? Then, if the predictive power of the yield curve is thought to arise from the information it contains on expectations about future activity, does decomposing yields into the respective expectations and term premium components improve this forecasting performance? To address these questions, a quasi-real-time forecasting exercise is conducted using data from the United States, New Zealand, and Australia.

2 Related literature

The literature investigating the relationship between the yield curve and economic activity is extensive. For example, Stock and Watson (1989) proposed a range of indicators believed to precede, coincide with or follow fluctuations in economic activity, the slope of the yield curve being one of these. Estrella and Hardouvelis (1991) used a probit model to estimate the probability of a recession in any given quarter, finding that an increasing yield spread was associated with a lower probability of recession in the following four quarters. Further examples, with similar results, examining how well yield curves estimate the probability of recessions include Bernard and Gerlach (1998), Anderson and Vahid (2001), Estrella and Trubin (2006), Ahrens (1999), Stock and Watson (2003), Duarte, Payá, and Venetis (2004). The general consensus among the aforementioned studies is that there is empirical evidence of a statistically significant relationship between an inversion of the yield curve, and a forthcoming recession.

Other research has followed time series approaches such as forecasting output using vector autoregressions but with mixed results. Ang, Piazzesi, and Wei (2006) used a predictive regression of real GDP growth on the yield spread and compared this with a vector autore-

gressive model of order 1 (VAR(1)) that jointly models the dynamics between output and the yield curve and imposes a no-arbitrage condition on bond yields. This model gave estimates of the term premia on bonds, allowing for a decomposition of bond yields into expectations and term premia components. Use of this decomposition improved the model fit of regressions relative to those that use yield spreads. The authors computed out-of-sample forecasts for each model and calculated RMSE ratios for each model relative to a simple AR(1). Using Diebold-Mariano tests (Diebold & Mariano, 1995) showed no statistically significant improvement in forecast performance compared with simple predictive regressions based on the VAR(1) for any specification or forecast horizon. Further, many specifications struggled to outperform an AR(1) model in this exercise. Bonser-Neal & Morley (1997) used data from 11 countries to predict economic activity, and found that out-of-sample forecasts are generally improved (relative to an AR(1) process) by incorporating the yield curve. Favero, Kaminska, and Söderström (2005) regressed growth in output on a short-term interest rate, inflation, and the yield spread to forecast future output growth. They estimated a decomposition of yields into their expectations and term premia components, finding that these components in place of the yield spread in the regression decreased forecast RMSE, but no tests for statistical significance were undertaken. Lewis (2015) performed a real-time forecasting exercise similar to that of the present paper but with greater focus on using foreign as well as domestic yield curves in these models to help forecast output, finding that use of the yield curve to forecast output outperformed simple autoregressive forecasts in some countries but not others.

The present paper makes three contributions to the literature. First, much past research investigates the yield curve but does not decompose yields into their respective expectations and term premia components. Papers that have used the decomposition have not compared performance between models that do and those that do not; nor have they carried out statistical tests for significant differences in forecast performance. As indicated earlier, much of the predictive power of the yield curve is believed to come from the expectations component of yields. It is therefore to be expected that the decomposition will lead to forecast improvements. The present work contributes to the literature by providing a systematic econometric exploration of the practical empirical relevance of this decomposition. Understanding how the decomposition affects the inferences we draw from the yield curve is crucial to learning how best to forecast future levels of output and potentially other economic variables through yield curve information.

Second, although many papers recursively estimate models and produce forecasts in an attempt to evaluate the out-of-sample forecast performance of these models, they do not take into account the data revisions that occur over the sample period. Measurements of economic variables such as output, inflation, and unemployment are consistently revised after their initial release. Failure to take this process of revised measurement into account is unrepresentative of the real-time exercise of expectation formation and may exaggerate

forecast performance. The present work specifically addresses this problem by ensuring forecasts are based solely on information that would be available in real-time.

Third, the vast majority of empirical studies of predictability of the yield curve relate to the United States. It is important to consider evidence in other country contexts to assess the external validity of findings for the US. There are some examples in the literature in which data from other countries is used. But this paper appears to be the first to utilise the decomposition of yields into expectations and term premia components outside of the United States.

A final contribution of the paper is to provide a diagnostic econometric approach that helps to reconcile conflicting results that have appeared within the existing literature. Most past research on this topic involves the issue of multiplicity in inference or problems involved in testing multiple hypotheses. No previous works have attempted to adjust for this complication in drawing inferences on the predictability of the yield curve. Correspondingly, the likelihood increases of spurious findings of statistically significant results. To address this risk, the present paper implements recent methods to control the family wise error rate (FWER). With these controls in place, many results that were statistically significant prior to the adjustments turn out to be no longer significant; and several results that claim increased accuracy in out-of-sample forecasts by augmenting models with yield curve information have less credibility.

3 Data and Forecasts

Adrian, Crump and Moench (2013) estimate the yield on US Treasury securities for varying maturities as well as the decomposition into their expectations and term premia components. These estimates are publicly available and were sourced from the Federal Reserve Bank of New York website (FRBNY, 2019). Vintage data, which provide initial estimates of GDP, the date at which they were revised, and the revised figures for the United States, Australia, and New Zealand, were sourced from the Archival Federal Reserve Economic Data website (FRBSL, 2019), the Australian Bureau of Statistics (ABS, 2019), and Adam Richardson (Richardson, 2019), respectively. The model proposed in Adrian, Crump and Moench (2013) was estimated with both Australian data (Jennison, 2017), (sourced from the Australian Office of Financial Management (AOFM, 2019), and New Zealand data (Callaghan, 2019) (provided by Adam Richardson (Richardson, 2019)).

Unfortunately, vintage data on these estimates of yields, expectations, and term premia are not available. So a strict real-time forecasting exercise is not possible. Instead, the results of the study can be interpreted in the following manner.

- When comparing models that do not make use of the yield curve with models that do, any improvement in forecast performance should be treated as an upper bound on the

improvement in forecast accuracy one could reasonably expect in practice.

- When comparing models which make use of total yields to models which use the decomposition of yields, any difference in forecast performance likely reflects what one would expect to see in practice. This is because estimates of yields and estimates of their decomposition are based on the same information available at the same time. So neither model has an informational advantage over the other in this regard and may therefore be a minor issue in assessing the practical import of the following findings. Indeed, as noted in Lewis (2015), estimates of yields derived from the Nelson-Siegel model (Nelson and Siegel, 1987) do not tend to change substantially after re-estimating the model when data revisions have been implemented. As estimates of yields used in Adrian, Crump & Moench (2013), Jennison (2017), and Callaghan (2019) are all based on the Nelson-Siegel-Svensson model (Svensson, 1994), it seems reasonable to assume that these estimates do not change in the face of data revisions and reflect the information that would be available to practitioners in real time. But it is unclear how data revisions may affect estimates of the decomposition of yields.

In conducting this forecasting exercise the data were split into two time periods – the model selection period and the forecast period. The division is somewhat arbitrary. Since model selection and forecasts are performed using only data that would be available in real time, any increase in the sample period for model selection comes at the expense of a reduction in the sample period for forecasting, and vice versa. This trade-off produces a corresponding trade-off between the statistical power of tests used for model diagnostics and the statistical power of tests used to compare forecasts. The sub-periods are therefore chosen so that roughly half the data in each case is used for model selection and half the data for forecast evaluation. The precise dates of the sub-periods involved are detailed in the following table.

Country	Model selection period	Forecast period
United States	1971Q1 to 1993Q3	1994Q2 / 1996Q1 to 2019Q1
New Zealand	1992Q2 to 2003Q4	2004Q2 / 2006Q1 to 2019Q1
Australia	1992Q3 to 2005Q3	2006Q1 / 2007Q4 to 2019Q3

Table 1: Model selection and forecast periods

The gaps that appear in the table between the end of each selection period and the forecast period are due to the lags that occur in the release of the data. For example, in Q1 of 2004, NZ GDP data for 2003Q4 was released. So if we were interested in forecasting NZ output in the next period, we would be producing a forecast for 2004 Q2.

Policymakers are typically concerned with the change in output from one period to another rather than the level of output itself. Accordingly, two forecasts are produced for each

case – a "short-term" and a "long-term" forecast. A "short-term" forecast will be a forecast of GDP growth in the following quarter from when the forecast is made. A "long-term" forecast will be defined as a forecast of GDP growth two years from the quarter of the forecast.

4 Model Selection

In each case, the selected model is first compared to a relatively simple benchmark. If the model incorporating yield curve information fails to outperform benchmarks as simple as an autoregressive process, the yield curve evidently has little predictive power in out-of-sample forecasts of economic activity. Next, VARs are estimated with (i) variables GDP growth and some measure of the yield curve, and (ii) variables GDP growth and some measure of the expectations component of the yield curve. The measures employed in these regressions for the yield curve and expectations components are the spread between the two and 10 year yields and expectations components. A key limitation of this approach is that it omits some information in the yield curve, such as yields at different maturities or the level of the yield spread. The latter may require a nonlinear formulation as a 100 basis point drop in the yield spread from 3% to 2% may not be associated with the same decline in economic activity as a 100 basis point drop from 0.5% to -0.5%. Nonetheless, this measure offers a parsimonious and convenient way of introducing yield curve information into the models. As is well known and as noted in this context by Ang, Piazzesi and Wei (2006), parsimonious time series models often lead to superior out-of-sample forecast performance.

4.1 The benchmark

An autoregression is used as a benchmark and lag selection is performed using BIC. Residuals from the model with the lowest BIC value are assessed for serial correlation. If the test outcome supports martingale difference errors, the model with this lag structure is taken as the forecasting model. If the test indicates that the residuals exhibit serial correlation, the test will be performed again on the next model with the lowest BIC, until one satisfies the necessary diagnostic checks. If no model satisfies the test for serially correlated residuals at the 5% level of significance, the procedure is to perform the tests again with a significance level of 10%, and so on incrementing the level until one does pass the test. As seen below, each case has at least one model which satisfies the necessary diagnostic checks at the 5% level of significance, so this aspect of the procedure turns out not to be of practical importance in the empirical application.

These tests for serial correlation are conducted using the robust test developed in recent work by Dalla et al. (2020), hereafter, referred to as the robust test. Tests that are most commonly used to detect serial or cross-correlation rely on i.i.d. innovation assumptions.

In practical work with economic and financial data, i.i.d innovation conditions are often too strong and such tests frequently find spurious evidence of such correlation. Moreover, optimal forecasting procedures typically rely on martingale difference residuals. Dalla et al. (2020) propose a test that allows for very general martingale difference errors. Their robust test reduces size distortion, helps to avoid spurious inference, and is better suited to finding an optimal forecasting model.

Details of the lag selection choices and results of the robust test are summarised below in Table 2, and Figures 2 to 4.

Lags	United States	New Zealand	Australia
1	**244.12**	119.26	**107.91**
2	248.54	**119.00**	111.77
3	252.98	122.16	114.59
4	257.36	121.08	114.02
5	261.66	124.67	114.67
6	265.96	127.72	117.40
7	269.04	131.19	121.00
8	260.51	127.33	123.67

Table 2: BIC for autoregressive processes with varying lags

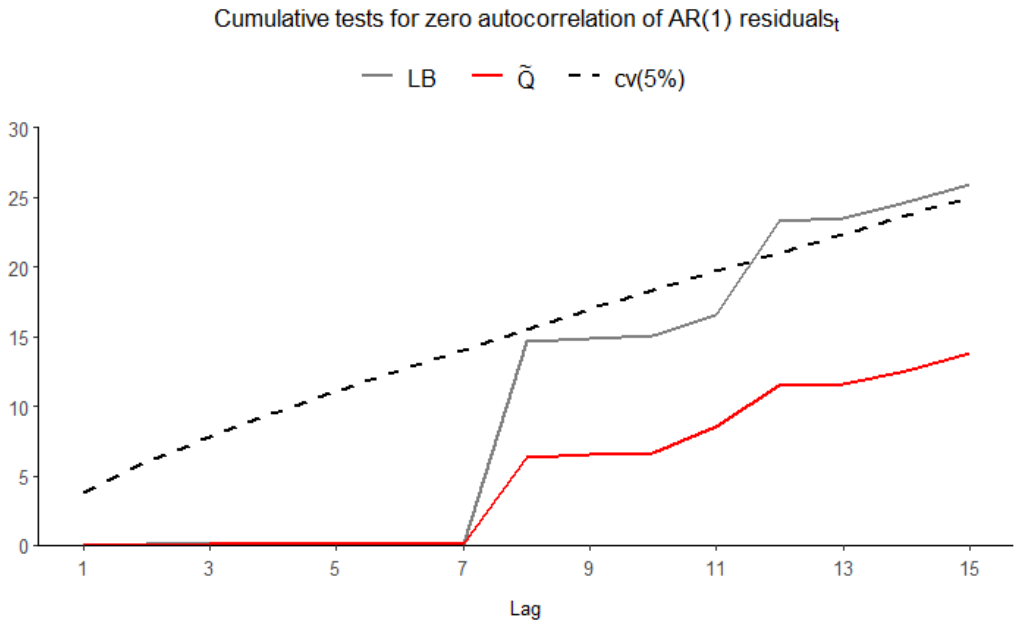

Figure 2: Results of the tests for serially correlated residuals of an AR(1) in the US case

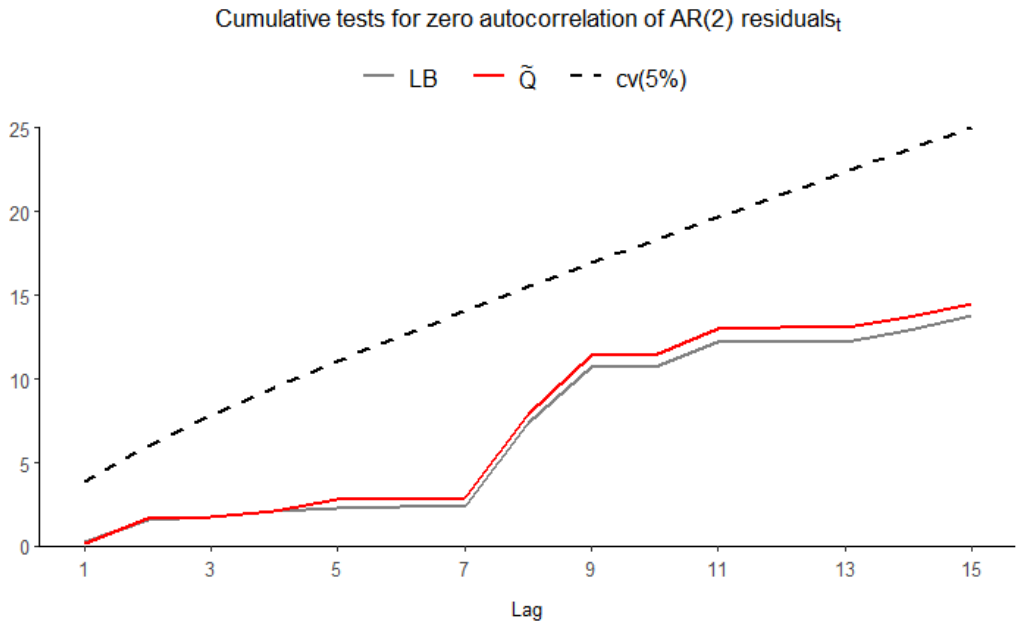

Figure 3: Results of the tests for serially correlated residuals of an AR(2) in the NZ case

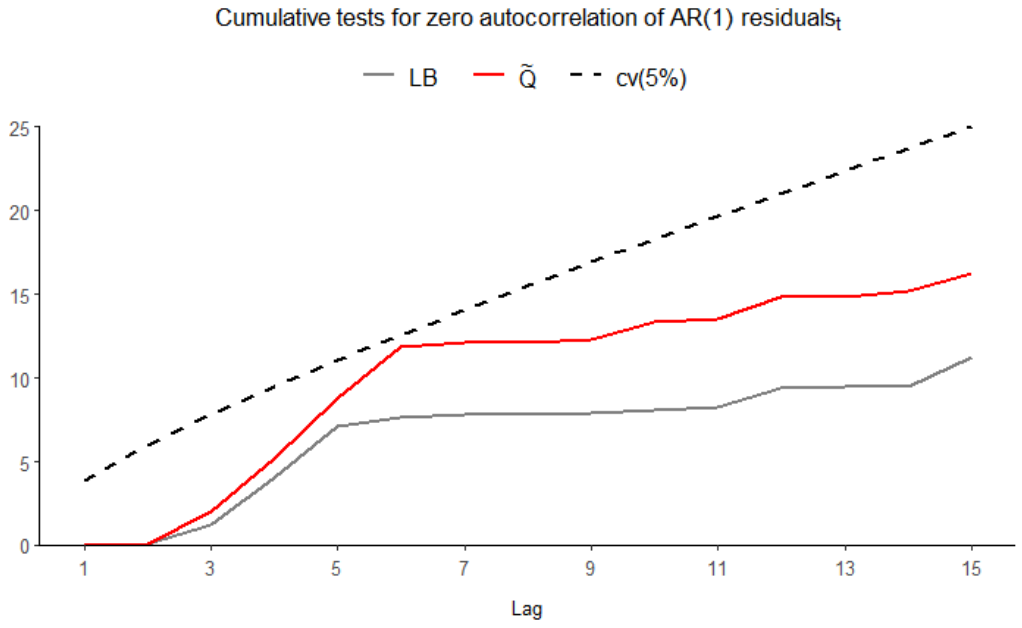

Figure 4: Results of the tests for serially correlated residuals of an AR(1) in the Australian case

In the figures above, the results of the standard Ljung-Box test for serial correlation are presented alongside those of the robust test. The results are largely consistent with

each other, generally indicating that none of the models which achieve the lowest BIC have serially correlated residuals. Therefore, the models chosen to serve as a benchmark take the following form

$$\text{GDP Growth}_t = \alpha_0 + \sum_{i=1}^{p} \text{GDP Growth}_{t-i} + \epsilon_t,$$

where $p = 1$ in the United States and Australian cases, and $p = 2$ in the New Zealand case.

4.2 The model with yields

In each case a vector autoregression (VAR) or vector error correction model (VECM) with GDP growth and the yield spread are used. Augmented Dickey-Fuller (Dickey and Fuller, 1979) and Phillips-Perron (Phillips and Perron, 1988) tests were performed for each variable with the outcomes summarised in Table 3 below.

Country	GDP Growth	Yield Spread	Expectations Spread
United States	I(0)	I(0)	I(1)
New Zealand	I(0)	I(0)	I(0)
Australia	I(0)	I(1)	I(1)

Table 3: Results of the unit root tests

Lags	United States	New Zealand	Australia
1	**-1.91**	-2.39	**-2.87**
2	-1.68	**-2.53**	-2.45
3	-1.31	-1.72	-1.74
4	-0.63	-0.64	-0.76
5	0.32	0.79	0.45
6	1.44	2.64	2.00
7	2.81	4.97	4.18
8	4.35	7.75	6.76

Table 4: MSC for VARs with varying lags (minima bolded)

In view of the unit root test findings an appropriate way to proceed in a model with the yield spread is to estimate a VAR of the form

$$y_t = A_0 + \sum_{j=1}^{p} A_j y_{t-j} + \epsilon_t$$

Where $y_t = (\text{GDP Growth}_t, \text{Yield Spread}_t)^\top$ in the United States and New Zealand case, and $(\text{GDP Growth}_t, \Delta\text{Yield Spread}_t)^\top$ in the Australian case. An information criterion approach is again adopted for model selection and the resulting multivariate Schwarz criteria (MSC) are presented in Table 3.

In both the US and New Zealand cases the robust tests do not indicate issues with auto or cross-correlation among the residuals for the model with the lowest MSC. A VAR(1) is therefore fitted for the US and a VAR(2) for New Zealand. While the MSC favours a VAR(1) in the Australian case, there appears to be serial correlation among the residuals at lag 4 onwards. Estimating a VAR(4) resolves this issue, as does the following restricted VAR (RVAR)

$$\begin{aligned} y_t &= A_0 + \sum_{j=1}^{4} A_j y_{t-j} + \epsilon_t, \quad \text{such that } A_2 = A_3 = 0, \\ &= A_0 + A_1 y_{t-1} + A_4 y_{t-4} + \epsilon_t, \end{aligned}$$

where $y_t = (GDP\ Growth_t, \Delta Yield\ Spread_t)^\top$. Further, this RVAR(4) achieves a MSC of -2.76, which is lower than the MSC of both the VAR(1) and the unrestricted VAR(4). This model is therefore selected for practical implementation.

4.3 The model with expectations

Following the empirical results of the unit root tests, a VAR(p) is estimated, with $\text{y}_t = (\text{GDP Growth}_t, \text{Expectations Spread}_t)^\top$ for New Zealand, and $(\text{GDP Growth}_t, \Delta\text{Expectations Spread}_t)^\top$ for Australia and the US. The respective findings for model selection are shown in Table 4 below. The MSC criteria indicate a VAR(1) is suited to the US and Australian data, and a VAR(2) to New Zealand. Further testing shows that none of these models have residuals exhibiting serial or cross-correlation. So these models are selected for implementation in the forecast exercises.

5 Forecast Comparisons

Each of the selected models was used to produce short-term and long-term forecasts and forecast errors were calculated using the most recent vintage version of the data. In conducting forecast comparisons we need to take account of multiplicity in testing. Hypothesis testing is conventionally conducted using a statistic whose distribution (obtained by finite sample, asymptotic, or bootstrap analysis) under the null hypothesis is used to assess significance, leading to rejection of the null if the critical value is exceeded. However, in multiple and sequential hypothesis testing, the likelihood of exceeding the critical value typically increases as more hypotheses are tested. In the present application, testing multiple hypotheses is an

issue as several different models are being compared and these at various forecast horizons. There are methods designed to adjust critical values or p-values to address such issues, including the Dunn-Bonferroni or Holm-Bonferroni corrections. These methods rely on the assumption that test statistics are independent and are therefore unlikely to be appropriate in the present application. For example, with dependent data if model A outperforms model B and model B outperforms model C, then relative to the case of independent statistics, it is more likely there is statistical significance when comparing Model A to Model C. Failing to account for this and applying Dunn-Bonferroni or Holm-Bonferroni corrections may penalise testing the third hypothesis involving Model C more harshly than is necessary, thereby increasing the probability of a type II error (or failure to reject a false null hypothesis). A similar argument applies when considering tests that involve multiple forecast horizons.

To address dependence among the test statistics in the present application a bootstrap procedure that closely follows those developed in White (2000) and Romano and Wolf (2005) is implemented. Although the context of this application differs slightly from that of White (2000) and Romano & Wolf (2005), their methods are still suited to resolve the issue of multiplicity. In applying these methods some limitations are worthy of note.

(i) The methods for dealing with multiplicity were developed in the context of recursive estimation of models but not in recursions that involve updates to vintage data. Using the most recent revisions of the observations may result in estimates of the sampling distribution of the test statistics of interest that differ slightly from the true distribution of interest. But differences of this type may not be of great concern as White (2000) established that re-estimating parameters in a recursion is not required to achieve consistent estimates of the sampling distributions of the relevant test statistics. If we accept that we do not need to decrease the variance of our estimates of parameters by updating them as more information becomes available in order to achieve a consistent estimate of the desired sampling distributions, we may reasonably expect a similar result to hold for data revisions.

(ii) The methods do not specifically deal with data that mixes different numbers of observations. Ideally, one would implement the StepM algorithm for the US, NZ, and Australian cases simultaneously; otherwise the correction will not be conservative enough. As an illustration, suppose 10 hypotheses are to be tested and suppose the hypotheses are split into two groups of 5 hypotheses, with the Bonferroni correction applied within each group. For hypotheses with p-values in the interval $\frac{\alpha}{10} \leq p < \frac{\alpha}{5}$, significance is declared at the α level, even though the correct inference is the opposite. A similar argument applies to the StepM algorithm. Although it may affect the resulting estimate of the joint sampling distribution of each test statistic, data is merged to avoid this difficulty. The bootstrap is performed using the final vintage observations, where observations at the beginning of the US and NZ data, and at the end of the Australian

data have been dropped in order to make the range of observation dates equivalent across each country. The final table used for the bootstrap contains observations ranging from 1992Q3 to 2019Q1, for a total of 107 observations.

Notwithstanding these limitations, corrections to address the issue of multiplicity were implemented in the following way.

The bootstrap method

Many conventional bootstrap methods return a sample with i.i.d. observations by construction. But in the case of time series data this construction seldom matches the true generating process due to the presence of serial dependence in the observations. Several methods to address this concern have been proposed and the approach adopted here is the stationary bootstrap procedure of Politis & Romano (1991), where bootstrap resampling has the following steps.

1. Select M,the number of times to resample the data $(Y_1, Y_2, ..., Y_T)$.

2. Draw $G_{i,m} \sim$ Geometric$(\frac{1}{q})$, and $U_{i,m} \sim$ Discrete Uniform$(1, T)$ as i.i.d. random variables for $m = 1, 2, ..., M$, and $i = 1, 2, ...$

3. Construct the blocks $B_{i,m} = (Y_{U_i,m}, Y_{U_{i,m}+1}, ..., Y_{U_{i,m}+G_{i,m}-1})$ for $m = 1, 2, ..., M$, $i = 1, 2, ...$ If $j > T$, define $Y_j = Y_{j(mod\ T)}$. For example, if $T = 100$, and $j = 341$, $341(mod\ 100) = 41$, so $Y_{341} = Y_{41}$. Thus, rather than trying to resample from a time period we do not observe, the modulus function allows the time period we do resample to be bounded above by the latest observation in the sample.

4. Construct the resample $B_m = (Y^*_{1,m}, Y^*_{2,m}, ..., Y^*_{T,m}) = (B_{1,m}, B_{2,m}, ...)$ by combining the individual blocks $B_{i,m}$ until T observations are attained in the resample for $m = 1, 2, ..., M$. If the final block contains more than enough observations to reach a length of T observations in B_m, discard the remainder that are unnecessary. For example, if $T = 50$, $B_{1,m}$ had a length of 30, $B_{2,m}$ had a length of 15, and $B_{m,3}$ had a length of 10, only the first 5 observations in block $B_{m,3}$ would be used.

After generating the resampled observations, the resampled observations are used to estimate each model and produce the forecasts and forecast errors in the same way as the first set of forecast errors. The StepM algorithm described in Romano and Wolf (2005) is then performed in the following way.

1. Formulate $H_{0,k}$ as the hypothesis that model k fails to outperform the benchmark model, for $k = 1, 2, ..., K$. For each of these K hypotheses, calculate the corresponding test statistics, $\omega_1, \omega_2, ..., \omega_K$.

2. Order these test statistics in descending order, $\omega_{r_1}, \omega_{r_2}, ..., \omega_{r_K}$, such that $\omega_{r_1} \geq \omega_{r_2} \geq ... \geq \omega_{r_K}$ to obtain indices $r_1, r_2, ..., r_K$

3. Let $j = 1$, and $R_0 = 0$.

4. For a suitably chosen critical value, c_j (discussed below), construct the confidence region:
$$C_j = \bigtimes_{R_{j-1}+1 \leq k \ \leq K} [\omega_{r_k} - c_j, \infty)$$

5. For $R_{j-1} + 1 \leq k \leq K$, if $0 \notin [\omega_{r_k} - c_j, \infty)$, reject H_{0,r_k}.

6. Stop upon failure to reject any of the remaining hypotheses. Otherwise, let R_j be the number of hypotheses rejected so far and set $j \mapsto j+1$. Calculate a new critical value, c_j, then repeat steps 3 to 6.

The critical value c_j

To obtain the appropriate critical value at each step of the algorithm described above, bootstrap procedures are used in the following way:

1. Given observations $(Y_1, Y_2, ..., Y_T)$, use the bootstrap method described above to re-sample $(Y^*_{1,m}, Y^*_{2,m}, ..., Y^*_{T,m})$ for $m = 1, 2, ..., M$.

2. Calculate the test statistics $\omega^*_{k,m}$ to test the null hypothesis that within bootstrap sample m model k fails to outperform the benchmark model, for $k = 1, 2, ..., K$, and $m = 1, 2, ..., M$.

3. Set $\omega^+_{j,m} = \max_{R_{j-1}+1 \leq k \leq K}(\omega^*_{r_k,m} - \omega_{r_k})$.

4. Compute c_j from the $(1-\alpha)$th quantile of the distribution of $(\omega^+_{j,m})^M_{m=1}$

This procedure delivers an estimate of the joint distribution of all test statistics, enabling estimation of the probability that the maximum among K test statistics will exceed a certain value, so that the critical values can be adjusted accordingly.

The Diebold-Mariano test statistics for forecast comparisons using the loss function $L_t = (y_t - \hat{y}_t)^2$ are presented in Table 5 below.

DM test statistics using the loss function $L_t = (y_t - \hat{y}_t)^2$		
Country	Short-term test statistic	Long-term test statistic
United States	-0.957 (p = 0.829)	-1.482 (p = 0.928)
New Zealand	1.201 (p = 0.118)	1.836 (p = 0.036)
Australia	2.543 (p = 0.007)	2.58 (p = 0.007)

Table 5: Diebold-Mariano test statistics at varying forecast horizons

The confidence region constructed using an average block length of 30 is therefore

$$C_1 = [-3.17, \infty) \times \cdots \times [-7.24, \infty)$$

Where the critical value at the first step of the algorithm is $c_1 = 5.76$.

Noting that $0 \in [\omega_{rj} - c_1, \infty)$ for $j = 1, 2, ..., 18$ (the total number of hypotheses being tested), the process stops at the first step of the StepM algorithm after failing to reject any null hypotheses. We conclude that the empirical evidence is insufficient to claim that any improvement in forecast performance has been achieved by augmenting the models with yield curve information or through the decomposition of yields into their expectations and term premia components. This conclusion continues to hold under various adjustments to the tests that include: (i) assessing hypotheses at the 10% level of significance; (ii) changing the average block length in the stationary bootstrap to 20 or 40 (from 30) observations; (iii) using differences between average forecast losses for the models as the test statistic in place of the Diebold-Mariano test statistic; or (iv) using the loss function $L_t = |y_t - \hat{y}_t|$ instead of $L_t = (y_t - \hat{y}_t)^2$. Overall and notwithstanding the fact that the methodology employed here has some limitations as noted earlier, the conclusions reached appear robust to a wide range of adjustments, including multiplicity, the particular test statistics used in the comparisons, and significance levels employed in the tests.

6 Discussion and Future Research

Our empirical findings reveal no cases in which forecasting with a model that includes yield curve information leads to more accurate forecasts than those made using a simple autoregressive process. Further, decomposing yields into expectations and term premia components does not improve forecast accuracy relative to an autoregression or a model with total yields. Overall, these results suggest that whilst the yield curve may act as a leading recession indicator, this property does not translate materially into better out-of-sample forecast performance of economic activity, at least for the countries and periods studied here. In contrast to these findings, results of previous research tend to be mixed and no general consensus on predictability has emerged within the existing literature. Since our findings contrast with some of the present evidence, potential reasons for the difference are worth discussing.

First, the predictive content of yield curve information has been examined in different contexts that may justify different conclusions. Previous studies may use data from different countries or different time periods and estimates of yields vary across papers as there are various ways they may be estimated and decomposed into expectations and term premia. Second, many empirical exercises do not perform quasi-real-time forecasting exercises or compensate for the effects of data revisions. Such revisions may well enhance with up-

dated information the capability of the yield curve as an aid in forecasting economic activity, whereas no strong evidence is found here using quasi-real-time forecasting methods. Third, the use of vintage data may lead to decreased statistical power. Forecasting with more recent data is likely to enhance accuracy and the use of vintage data introduces noise in the forecast errors, making small improvements more difficult to distinguish from randomness. In addition to data updating, there are issues associated with methodology. Best practice has generally been followed in model specification. But there is no general consensus on performing specification searches. Papers may apply different information criteria or sequential testing algorithms and reach different conclusions on specification with the same data; and the same criteria could lead to different outcomes when data for different time periods is used. Similarly, results may differ depending on the specific tests used to assess serially correlated residuals or the specific variables included in a model.

Finally, the present findings and methodology employed in conducting tests tell a cautionary tale concerning empirical model building. Without attention to the issue of multiplicity in the present exercise several differences in forecast performance would be judged statistically significant. Best practice in testing multiple hypotheses does not sustain this interpretation. Much previous research has concluded that the yield curve helps to forecast output and other studies reported mixed results across models, forecast horizons and countries. No previous studies to our knowledge have addressed the effects of multiplicity on inferential validity. Some past conclusions may hold up under more robust methods but only a complete re-analysis of the relevant data with appropriate methodology attending to multiplicity would reveal how many statistically significant findings would be sustained and how many would not.

Several avenues for future research seem promising beyond replication studies with more robust methods of inference. Forecast performance may be time-varying: if central banks improve ability to forecast recessions and offset them, the correlation between an inverted yield curve and a subsequent recession may attenuate over time; and similar effects may apply to the forecast performance of the yield curve in general. The methodology employed here may be used in forecasting other variables. Given the role of interest rates and the yield curve in many macroeconomic models, financial models, and general policy relevance, their potential in forecasting other variables than GDP growth seems worthy of investigation. Finally, to better reflect the small open economy features of New Zealand and Australia and their susceptibility to shocks in large economies such as the US and China, it seems appropriate for models to be augmented with term spreads from other countries to examine whether these variables may enhance forecast performance.

References:

Adrian, T., Crump, R.K., Moench, E. (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics*. 110(1), pp 110-138.

Ahrens, R. (1999). Predicting recessions with interest rate spreads: A multicountry regime-switching analysis. CFS Working Paper Series, 1999/15.

Anderson, H.M., Vahid, F. (2001). Predicting the Probability of a Recession with Nonlinear Autoregressive Leading-Indicator Models. Macroeconomic Dynamics, 5(4), pp 482-505.

Ang, A., Piazzesi, M., Wei, M. (2006). What does the yield curve tell us about GDP growth?. *Journal of Econometrics*. 131, pp 359-403.

AOFM. (2019). Data Hub: Term Premium Estimates.

ABS. (2019). Australian National Accounts: National Income, Expenditure and Product.

Bauer, M.D., Mertens, T.M. (2018). Economic Forecasts with the Yield Curve. FRBSF Economic Letter.

Bernard, H.J., Gerlach, S. (1998). Does the Term Structure Predict Recessions? The International Evidence. *International Journal of Finance and Economics*, 3(3), pp 195-215.

Bosner-Neal, C., Morley, T.R. (1997). Does the yield spread predict real economic activity?: a multicountry analysis. *Economic Review*, qiii, pp 37-53.

Callaghan, M. (2019). Expectations and the term premium in New Zealand long-term interest rates. Reserve Bank of New Zealand Analytical Note Series. ISSN 2230-5505.

Dalla, V., Giraitis, L., Phillips, P.C.B. (2020). Robust Tests for White Noise and Cross-Correlation. Cowles Foundation Discussion Paper No. 2194; *Econometric Theory* (forthcoming).

Dickey, D.A., Fuller, W.A., (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root, *Journal of the American Statistical Association*, 74(336), pp 427-431.

Diebold, F.X., Mariano, R.S., (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13(3), pp 253-263.

Duarte, A.& Payá, I., Venetis, I.A. (2004). Predicting Real Growth and the Probability of Recession in the Euro Area Using the Yield Spread. IVIE working paper V-3361-2004.

Estrella, A., Hardouvelis, G.A. (1991). The Term Structure as a Predictor of Real Activity. *The Journal of Finance.* XLVI(2), pp 555-576.

Estrella, A., Mishkin, F.S. (1996). The Yield Curve as a Predictor of U.S. Recessions. FRBNY Current Issues in Economics and Finance. 2(7).

Estrella, A., Trubin, M.R. (2006). The Yield Curve as a Leading Indicator: Some Practical Issues. FRBNY Current Issues in Economics and Finance. 12(5).

Favero, C.A., Kaminska, I.,& Söderström, U. (2005). The Predictive Power of the Yield Spread: Further Evidence and A Structural Interpretation. CEPR Discussion Papers 4910.

FRBNY. (2019). Treasury Term Premia.

FRBSL. (2019). Real Gross Domestic Product (GDPC1).

Hicks, J.R. (1946). Value and Capital, *Oxford: Clarendon Press.*

Jennison, F. (2017). Estimation of the Term Premium Within Australian Treasury Bonds. Australian Office of Financial Management Working Paper.

Lewis, M. (2015). Forecasting with Macro-Finance Models: Applications to United States and New Zealand. Masters thesis, Victoria University of Wellington, New Zealand.

Meiselman, D. (1962). The Term Structure of Interest Rates. *The American Economic Review*, 54(6), pp 1149-1151.

Nelson, C., Siegel,A.F. (1987). Parsimonious Modeling of Yield Curves. *The Journal of Business.* 60(4), pp 473-469.

Phillips, P.C.B., Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika.* 75(2), pp 335-346.

Politis, D.N., Romano, J.P. (1994). The Stationary Bootstrap *Journal of the American Statistical Association*, 89(428), pp 1303 - 1313.

Richardson, A. (2019). Personal communication, 4th December 2019.

Romano, J.P., Wolf, M. (2005). Stepwise Multiple Testing Formalized as Data Snooping. *Econometrica*, 73(4), pp 1237 - 1282.

Stock, J.H., Watson, M.W. (1989). New Indexes of Coincident and Leading Economic Indicators. NBER Macroeconomics Annual 1989, Volume 4, pp 351-409.

Stock, J.H., Watson, M.W. (2003). How Did Leading Indicator Forecasts Perform During the 2001 Recession? Federal Reserve Bank of Cleveland *Economic Quarterly*, 89(3), pp 71-90.

Svensson, L.E.O. (1994). Estimating Forward Interest Rates with the Extended Nelson and Siegel Method. *Sveriges Riksbank Quarterly Review* 1995:3, pp 13-26.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica*, 68(5), pp 1097-1126.

Williams, J.B. (1938). The Theory of Investment Value. *Harvard University Press.*