Bagging in Tourism Demand Modeling and Forecasting

George Athanasopoulos

Department of Econometrics and Business Statistics

Monash University

Australia

Haiyan Song

School of Hotel and Tourism Management

The Hong Kong Polytechnic University

Hong Kong SAR

Jonathan A. Sun

Department of Econometrics and Business Statistics

Monash University

Australia

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Song and Athanasopoulos acknowledge the financial support of the Hong Kong Research Grant Council (Grant No. PolyU 5969/13H). Athanasopoulos acknowledges the financial support of Australian Research Council Discovery Project DP 1413220.

Bagging in Tourism Demand Modeling and Forecasting

George Athanasopoulos, Haiyan Song and Jonathan A. Sun

Abstract

This study introduces bootstrap aggregation (bagging) in modelling and forecasting tourism demand. The aim is to improve the forecast accuracy of predictive regressions while considering fully automated variable selection processes which are particularly useful in industry applications. The procedures considered for variable selection is the general-to-specific (GETS) approach based on statistical inference and stepwise search procedures based on a measure of predictive accuracy (MPA). The evidence based on tourist arrivals from six source markets to Australia overwhelmingly suggests that bagging is effective for improving the forecasting accuracy of the models considered.

Keywords: Australia, bootstrap aggregation, model selection, predictive regression.

1 Introduction

Forecasts of tourism demand form the foundation of policy making, strategic planning and operations management for both public and private tourism stakeholders. The accuracy of such forecasts is imperative in minimising the risks of incorrect decisions and maintaining the sustainable development of tourism in a destination. Over the past few decades using regression class models which incorporate exogenous variables has become common practice. These are generally labelled as “causal econometric models”. The advantage of such models is that they can be used to provide vitally important analysis of the relationships between tourism demand and economic variables such as prices, interest rates and output, amongst others (see for example Crouch, 1992; Li, Song, and Witt, 2005; Song and Li, 2008). A major challenge however with the usual econometric causal models when forecasting is that one has to first forecast the economic variables themselves before forecasting tourism demand, the variable of interest. This indeed is not an easy feat in the incredibly uncertain and volatile times we live in. Although some individual empirical studies have shown a satisfactory forecasting performance of causal models (see for example Li, Wong, Song and Witt, 2006; and Li and Song, 2007) the comprehensive tourism forecasting competition of

Athanasopoulos et al. (2011) showed that the usual causal models using the commonly identified economic predictors had inferior forecasting performance compared to pure time series alternatives. In this paper we pick up from this result and explore a novel to the tourism literature modelling framework in an effort to build causal models with improved forecasting performance. In particular we build what we refer to as "predictive regressions" which use as predictors lags of the commonly used economic variables and therefore forego the challenge of these needing to be forecast. We complement these models with bootstrap aggregation (or bagging) proposed by (Breiman 1996a) and Bühlmann and Yu (2002) in an effort to enhance their forecast performance and we find overwhelming affirmative evidence of this.

The framework we explore is fully automated model selection involving the selection of predictors using several procedures. This is highly relevant to the tourism sector where automated algorithms such as the Hong Kong Tourism Demand Forecasting System (see Song et al. 2013) are regularly used. The first procedure we investigate in building predictive regressions is the general-to-specific (GETS) approach using statistical inference and in particular individual t-tests to eliminate predictors from the model (Hendry and Krolzig, 2005). This process has been successfully implanted in tourism forecasting (see for example Song and Witt, 2003, Narayan, 2004, Katircioglu, 2009, Wang, 2009, Song and Lin, 2010). Despite GETS' popularity, it suffers from two major issues. First, predictors and their lags are used and these are highly correlated. With multicollinearity present, t-tests are unreliable. Secondly, the decision rule in the GETS process is said to be unstable as this is highly dependent on sequential testing. A decision rule is unstable if a small change of the data set would lead to a predictor being dropped or included. This leads to an increased variance of the forecast and reduced accuracy (Breiman, 1996b). To overcome the instability of the decision rule in the GETS model reduction process, we complement GETS with bagging proposed by Breiman (1996a) and Bühlmann and Yu (2002). Bagging is a machine learning algorithm designed specifically to reduce instability of algorithms from generating new learning sets. In our setting this aims to reduce out-of-sample forecast error. As Breiman (1996a, p. 124) claimed, "The evidence, both experimental and theoretical, is that bagging can push a good but unstable procedure a significant step towards optimality." Inoue & Kilian (2008) and Rapach & Strauss (2010 and 2012) demonstrated that the GETS-bagging procedure reduced forecasting errors by large margins in predicting US inflation, and US national and regional employment growth, respectively.

As alternatives to the GETS procedure we explore model selection processes which do not rely on statistical inference but use what we refer to a measure of predictive accuracy (MPA). These include model selection criteria such as the AIC, the bias corrected AIC (AICc) and BIC, and also the leave-one-out cross-validation (CV) statistic (see Hyndman and Athanasopoulos, 2014, for further details). As a full search process is not plausible given the high number of predictors we explore stepwise subset search processes (see Hastie et al. 2009).

Over the past century, tourism has developed into one of the most important drivers of global economy. The World Tourism Organization (2014) reports that in 2013 tourism was worth USD1.4 trillion in exports; accounted for 9% of global GDP and created 1 in 11 jobs. International tourist arrivals grew to a record 1,087 million in 2013 after breaking the 1 billion mark in 2012. In the long term, international tourist arrivals are expected to reach 1.8 billion by 2030. Likewise, the tourism industry makes an important contribution to Australia's economy which is our case study in this paper. Tourism Research Australia (2014) reports that tourism generated AUD 90.7 billion of GDP which was 6% of total GDP in 2013. In the same year, tourism also created employment for 929,000 individuals, which was 8% of the total number of employed persons. Further, the output multiplier of Australia (1.87) showed that for every dollar that tourism earned directly for the Australian economy, an additional 87 cents was generated for other parts of the economy. This ratio was higher than retail trade (1.77), mining (1.70) and education and training (1.41) among others for 2012. In short, tourism has become a critical component of both global and Australian economies.

In this paper we implement the methods introduced above to build predictive regression models for Australian international tourist arrivals. We conclude that the common economic factors used in the literature such as, prices, exchange rates, output are reliable predictors for international tourism demand in Australia up to 2008Q3 signifying the Lehman Brothers Bankruptcy (LBB) and the beginning of the Global Financial Crisis (GFC). These relationships need to be rethought about when forecasting the post-LBB period. We also find overwhelming evidence that bagging does improve the forecasting performance for the predictive regression models in almost all settings however again with the exception of the post-LBB period.

The paper is structured as follows: Section 2 provides a concise literature review related to the topic understudy. Section 3 introduces the general framework for building predictive regression models including the GETS approach and model selection using MPAs; Section 4 presents forecasting with bagging; Section 5 presents that data and Section 6 includes the empirical analysis and the results. Section 7 concludes the study.

2. Literature Review

Given the importance of tourism demand modeling and forecasting in tourism practice, extensive research has been carried out over the past half of a century. Broadly the research in this field is divided into two paths in terms of the nature of forecasting techniques: non-causal time series models and the causal econometric approaches (Song and Turner, 2006). One of the major advantages of the econometric approaches over the time series models lies in their ability to analyse the causal relationships between tourism demand and various economic factors, i.e., a demand elasticity analysis. Recent econometric studies of tourism demand have identified the following key variables influencing international tourism demand: tourists' income, population of the country of origin (or the total income level of the origin country divided by its population, i.e., income per capita), tourism prices in a destination relative to that in the origin country, tourism prices in competing destinations (i.e., substitute prices), and exchange rates (Crouch, 1992; Li, Song, and Witt, 2005; Song and Li, 2008). Demand elasticity analysis has important policy implications, in terms of interpreting the change of tourism demand from an economic perspective, proving policy recommendations as well as evaluating the effectiveness of the existing tourism policies. In addition, many empirical studies suggested that econometric forecasting approaches outperformed their time-series counterparts in tourism demand forecasting competitions (e.g., Li and Song, 2007; Li, Wong, Song and Witt, 2006; Witt, Song and Louvieris, 2003). Given the dual benefit of econometric approaches in terms of both economic analysis power and forecasting capability, this paper focuses on the further methodological development in this direction.

Song, Witt and Zhang (2008) and Song, Gao and Lin (2012) developed a web-based truism demand forecasting system to predict the tourism demand variables such as tourist arrivals, tourists' expenditure on different product categories and the demand for hotel rooms. This system has been used as points of reference for many industry practitioners including theme parks, hotels and government agencies. This forecasting system along with other tourism demand systems normally involve three stages. It starts with a pre-modeling data analysis, followed by statistical modeling and forecasting, and then judgmental forecasting adjustments. One goal of these forecasting is to automate the statistical forecasting procedure of Stage Two without any loss of forecast accuracy (Song, Gao and Lin, 2012). Given the direct relevance of the demand model specification to the statistical forecasting accuracy, selecting a well-defined demand model following a scientific modeling procedure is an essential component of this project (in Stage Two), as with any empirical research in economics, because "there is no *a priori* theory to pre-define a complete and correct specification" (Hendry and Krolzig, 2005, p. C32).

Most of the existing systems use GETS to modeling and forecasting in which a general econometric model that contains all possible influencing factors that may affect the demand for tourism in a destination. Typical influencing factors considered by this general model include income level of tourists from the source markets, prices of the tourism products/services in the destination (measured by the consumer price index of the destination relative to that of the source markets adjusted by the exchange rates between the destination currency and the currencies of the source markets), the prices of substitute destinations (adjusted by the relevant exchange rates), tourists' travel preferences, and destination's marketing expenditure, etc. (Song, Witt and Li, 2009, pp.2-7). The GETS specifications of the forecasting models in the system normally started with incorporating all these influencing variables together with their lagged values (lagged by four periods for each variable as the system uses quarterly data for model estimation). The general model is termed autoregressive distributed lag (ADL) model. A typical ADL model, therefore, involves at least more than 20 explanatory variables apart from the one-off event dummies. This general model is then estimated using the ordinary least squares (OLS) method. Insignificant variables are then eliminated in the subsequent estimations following a decision rule such as starting from the least significant one according to t-statistics of the estimated parameters (Song and Witt, 2003). The OLS estimation process is repeated until all variables left in the model are both statistically significant and economically plausible (i.e., the coefficients of the variables have correct signs according to economic theory). For a detailed explanation of the GETS modeling approach see, for example, Song and Witt (2003) and Song, et al. (2009, pp.46-69).

GETS modeling has been proved to be effective in tourism forecasting by a number of researchers such as Katircioglu (2009), Narayan (2004), Song and Witt (2003), Song and Lin (2010) and Wang (2009) due to its ease of specification and robustness in model estimation compared with the specific-to-general modeling approach. However, to some extent, the specification of the final forecasting model based on the GETS methodology still suffers from possible subjective influences and the model reduction procedure can vary from researcher to researcher, as the model reduction process is sensitive to the sequence of removing the insignificant variables or the variables that have incorrect signs. As a result, the "optimal" model may not be reached through the GETS procedure.

Another problem associated with the GETS procedure is that the model reduction process is carried out manually by researchers, which is time-consuming and errors may occur as a result of fatigue and incorrect judgments made by the researchers. In practice the elimination of the variables is determined by the decision rules, which refer to the t-statistics of the parameters according to which the variables are eliminated. Using t-statistics as decision rule to select the variables to be included in the final specific forecasting model can be problematic if the explanatory variables are correlated (Breiman, 1996a), as in the case of the ADL models where both the current and lagged values of the explanatory variables are included. In this case, the decision rules for variable selection are said to be unstable. The unstable decision rules also prevent the forecasting system from being fully automated, which is desired by practitioners.

To overcome the instability of the decision rule in model reduction process, the *bootstrap aggregation* or *bagging* method proposed by Breiman (1996a) and Bühlmann and Yu (2002) could be used to select the variables to be included in the final forecasting model. Bagging is a statistical method designed specifically to reduce the forecasting errors through selecting the predictors when the decision rules are unstable. As Breiman (1996a, p. 124) claimed, "The evidence, both experimental and theoretical, is that bagging can push a good but unstable procedure a significant step towards optimality." Inoue & Kilian (2008) and Rapach & Strauss (2010 and 2012) further demonstrated that the GETS-bagging procedure reduced forecasting errors by large margins in predicting the US inflation, US national and regional employment growths, respectively. It is expected that this approach may also be relevant and useful for the specification of the GETS forecasting models for tourism for the reasons highlighted above. Song, et al. (2012) showed that the ADL models produced relatively accurate forecasts. However, the forecasting errors generated by some of the models related to such volatile markets as China and Taiwan were relatively large (Mean absolute Forecasting Errors were greater than 10%). These large forecasting errors were partially due to the difficulty in reducing the general ADL models to specific models suitable for forecasting. Although Song et al (2012) showed that the judgmental adjustments with input from experts were able to reduce the forecasting errors, an alternative approach to take would be to improve the forecasting performance of these models through GETS bagging. This study represents the first attempt to integrate GETS bagging into tourism demand forecasting. A potential benefit of GETS bagging to tourism forecasting is that the automation of the statistical forecasting process is made possible.

3. Building predictive regressions for tourism demand

The predictive regression model we build for forecasting tourism demand from a source country has the general form

$$y_{t+h} = \alpha + \sum_{j=0}^{P-1} \beta_j y_{t-j} + \sum_{j=0}^{P-1} \sum_{k=1}^{K} \gamma_{k,j} x_{k,t-j} + \sum_{m=1}^{M} \delta_m d_{m,t} + \varepsilon_{t+h} \qquad (1)$$

where y_{t+h} is the demand for tourism (measured by arrivals) by residents from source country i and h is the forecasting horizon; y_{t-j} is the jth lag of the dependent variable, $x_{k,t-j}$ is the jth lag of the kth economic variable considered, $d_{m,t}$ is the mth dummy variable and ε_{t+h} is the h-step ahead error term with zero mean and constant variance; α, βs, γs and δs are parameters to be estimated. As we are working with quarterly data in our case study we set $P = 4$. Also the dependent variable and the K economic predictors for each source country will be appropriately transformed to stationary before they enter the modelling framework. In order to simplify the exposition in what follows we present the model in vector notation such that

$$y_{t+h} = \alpha + \boldsymbol{\beta}'\boldsymbol{l}_t + \boldsymbol{\gamma}'\boldsymbol{x}_t + \boldsymbol{\delta}'\boldsymbol{d}_t + \varepsilon_{t+h} \qquad (2)$$

where $\boldsymbol{l}_t$ is a vector containing lags of the dependent variable, $\boldsymbol{x}_t$ is a vector containing the economic variables, $\boldsymbol{d}_t$ is a vector with the dummy variables.

3.1. A general to specific (GETS) approach

In the GETS procedure model selection is only applied to the elements of $\boldsymbol{x}_t$ and $\boldsymbol{d}_t$, i.e., this includes the economic predictors and their lags and the dummy variables. The procedure starts by pre-determining the dimension of $\boldsymbol{l}_t$ based on minimizing the AIC and this is fixed. This is to done in an effort to ensure that enough dynamics are included in the model so that approximate whiteness of residuals can be achieved. Having approximately well behaved residuals is important in this approach as the p-values of the estimated coefficients will be used to select predictors for the model. In an initial step this is confirmed by checking the ACF of the residuals and is further complemented by using the HAC estimator proposed by Newey and West (1987) to account for any possible heteroskedasticity and autocorrelation left in the residuals when constructing the p-values. Once the dimension of the lagged dependent variable vector is determined, the GETS procedure starts by estimating the model including all other predictors using ordinary least squares (OLS). In an iterative process the predictor with the coefficient that has the largest p-value, greater than the pre-determined critical value (we use 0.05 and 0.01) is eliminated from the model. This iterative process stops when the largest p-value is less than the critical value. When this procedure ends, all the predictors in the model are statistically significant.

3.2. Model selection using a Measure of Predictive Accuracy (MPA)

In this approach model selection applies to all elements in $\boldsymbol{l}_t$, $\boldsymbol{x}_t$ and $\boldsymbol{d}_t$ as whiteness of residuals is not essential. A predictor is included if it improves the model's MPA. Therefore, all predictors can be added or dropped from the model (the only exception being the intercept). As the number of predictors is large it is preventative to perform model selection with all possible combinations, i.e., a full search approach. There are 2^Q where $Q = (P-1)(K+1) + M$ possible model combinations to be considered to explore the full model space. Therefore, we consider stepwise procedures to efficiently traverse through the model space. The four procedures we consider are, the backward, forward and hybrid backward and hybrid forward procedures (see Hastie et al. 2009). From the four procedures the hybrid ones performed best, no matter which MPA was used. In the results that follow we present the results from the hybrid forwards procedure in order to save space. The results using the hybrid backwards procedure were qualitatively similar. All results are available upon request.

The forward stepwise regression starts with a model that includes only the intercept. Predictors are added one at a time and the one that most improves the MPA is retained in the model. The procedure is repeated until no further improvement in the MPA can be achieved. In contrast to the forward procedure the backward stepwise regression starts with a model that includes all predictors. Each predictor is removed from the model one at a time and the one that mostly improves the MPA by being removed is eliminated. The procedure repeats until no further improvements in the MPA can be achieved. In the hybrid version of the forward (backward) stepwise regression each time we add (drop) a predictor we also consider dropping (adding) a predictor. Obviously this is relevant in the forward (backward) procedure once at least three predictors have been added (dropped) to the model. Table 1 summarises the forward stepwise (hybrid) procedure. All programming was implemented in R version 3.2.0 and is available upon request. For more examples of stepwise selection process see James et al. (2013).

Table 1: The forward stepwise algorithm.

The forward stepwise (hybrid) procedure starts with a model that only includes an intercept.

1. Calculate MPA for the model. This forms the ‘current MPA’.
2. Consider models with one additional predictor at a time. Select the model with the ‘best MPA’.
3. If the ‘best MPA’ is not better than the ‘current MPA’ the procedure stops; no predictor can improve the model.
4. If ‘best MPA’ is better than the ‘current MPA’, the associated predictor is added to the model and a new ‘current’ MPA is formed. (Back to step 2 if not hybrid).
5. (Hybrid). Drop each predictor currently in the model one at a time. If MPA can be improved then a new ‘current MPA’ is formed. (Back to step 2).

The MPAs we consider are the usual information criteria, i.e., the AIC defined as

$$\mathrm{AIC} = N \log\left(\frac{SSE}{N}\right) + 2(Q + 2)$$

the bias corrected AIC defined as

$$\mathrm{AICc} = AIC + \frac{2(Q + 2)(Q + 3)}{N - Q - 3}$$

and BIC is defined as

$$\mathrm{BIC} = N \log\left(\frac{SSE}{N}\right) + (Q + 2)\log(N)$$

where *SSE* is the sum of squared errors, *N* is the number of observation used for estimation and *Q* is the number of predictors in the model (excluding the intercept). The properties of the AIC and the BIC are well known. BIC will choose the same or fewer predictors compared to the AIC as it penalises additional predictors more heavily than the AIC. Of the three the one that has never been applied in the tourism literature is the AICc. The AICc is particularly useful in small samples with many predictors where the AIC is biased towards selecting too many predictors. The final MPA we consider is the cross-validation (CV) statistic by implementing leave-one-out cross-validation which has also never been used in the tourism literature. When the number of observations is large, minimising AIC is identical to minimizing the CV statistic (Stone, 1977). Calculating the CV can be a time consuming process in other situations. Fortunately, for regression the CV statistic can be calculated very effectively by the following equation

$$\mathrm{CV} = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{e_j}{(1-h_j)}\right]^2$$

where e_j is the residual from fitting the model under consideration to all N observations and h_j are the diagonal elements from the hat-matrix defined as $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ where $\boldsymbol{X}$ is a matrix of all predictors. For more on the model selection procedures applied here please refer to Hyndman and Athanasopoulos (2014).

4. Bagging forecasts

In this section we follow the exposition of Inoue and Killian (2008) to demonstrate the application of bagging with correlated regressors. The starting point is the predictive regression model in the form of equation (2). Forecasting with bagging involves generating a large number, $B,$ of pseudo samples which we refer to as bootstrap samples.

Let $\boldsymbol{z}_t$ be a vector containing all the predictors at time t such that $\boldsymbol{z}_t = (1, \boldsymbol{l}_t', \boldsymbol{x}_t', \boldsymbol{d}_t')'$. Suppose our sample ends at time T, hence y_T denotes the most recent observation. Arrange the data in a matrix of dimensions $(T-h)\times(1+P(1+K)+M)$ denoted by

$$\boldsymbol{A} = \begin{bmatrix} y_{1+h} & \boldsymbol{z}_1' \\ \vdots & \vdots \\ y_T & \boldsymbol{z}_{T-h}' \end{bmatrix}.$$

Generate a bootstrap sample b by drawing with replacement from matrix $\boldsymbol{A}$ blocks of m rows so that the dependence in the error term is captured. We denote this by

$$\boldsymbol{A}^{(b)} = \begin{bmatrix} y_{1+h}^{(b)} & \boldsymbol{z}_1'^{(b)} \\ \vdots & \vdots \\ y_T^{(b)} & \boldsymbol{z}_{T-h}'^{(b)} \end{bmatrix}.$$

For each bootstrap sample implement model selection and estimate the model from $\boldsymbol{A}^{(b)}$. Fit the model back to the most recent observations in $\boldsymbol{A}$ and generate forecast $\hat{y}_{T+h}^{(b)}$. Repeat the process for $b = 1, \ldots, B$. In theory $B = \infty$. In practice Inoue and Kilian (2008) suggest that $B = 100$ provides a reasonable approximation. Also the block size *m* is chosen to capture the dependence in the error term. If the forecast model is correctly specified in that $E(\varepsilon_{t+h}|I_t) = 0$, where I_t is the information set at time *t*, then $m = h$ is sufficient (see e.g. Gonçalves and Killian, 2004). For further details on these parameters the please refer to Inoue and Killian (2008) and references therein.

The final forecast is then given by

$$\hat{y}_{T+h}^{(bag)} = \frac{1}{B}\sum_{b=1}^{B} y_{T+h}^{(b)}.$$

In the empirical evaluation that follows we find that instead of using the averaging across the B bootstrap forecasts taking the median generated slightly better results and therefore we present these. A similar result was found for combining forecasts from an ensemble of neural networks in Kourentzes et al. (2014).

5. Data

Our case study aims to build predictive models to forecast Australian international tourism demand.

5.1 Dependent variable

We consider quarterly tourist arrivals to Australia over the period 1981:Q1 to 2012:Q3 from six origin countries: Canada, Germany, Japan, New Zealand, UK and the US. The incoming tourist data are obtained from *Tourism Research Australia: International Visitor Survey.* Figure 1 provides time series plots of the arrivals data for each source country over the entire sample. The solid vertical line indicates 2008Q3 as the date of the Lehman Brother Bankruptcy (LBB) and the dashed line shows that beginning of the hold-out sample to be used for the forecast evaluation. The first four plots show a clear upward trend of tourist arrivals from Canada, US, Germany and UK. This increase seems to have been affected by the LBB and the global financial crisis that followed. Tourist arrivals from Japan are very different to all other source countries showing a downward trend since the mid-nineties. Tourist arrivals from New Zealand also seem different from the other four countries recovering quickly after the LBB and showing an upward trend quickly after that.

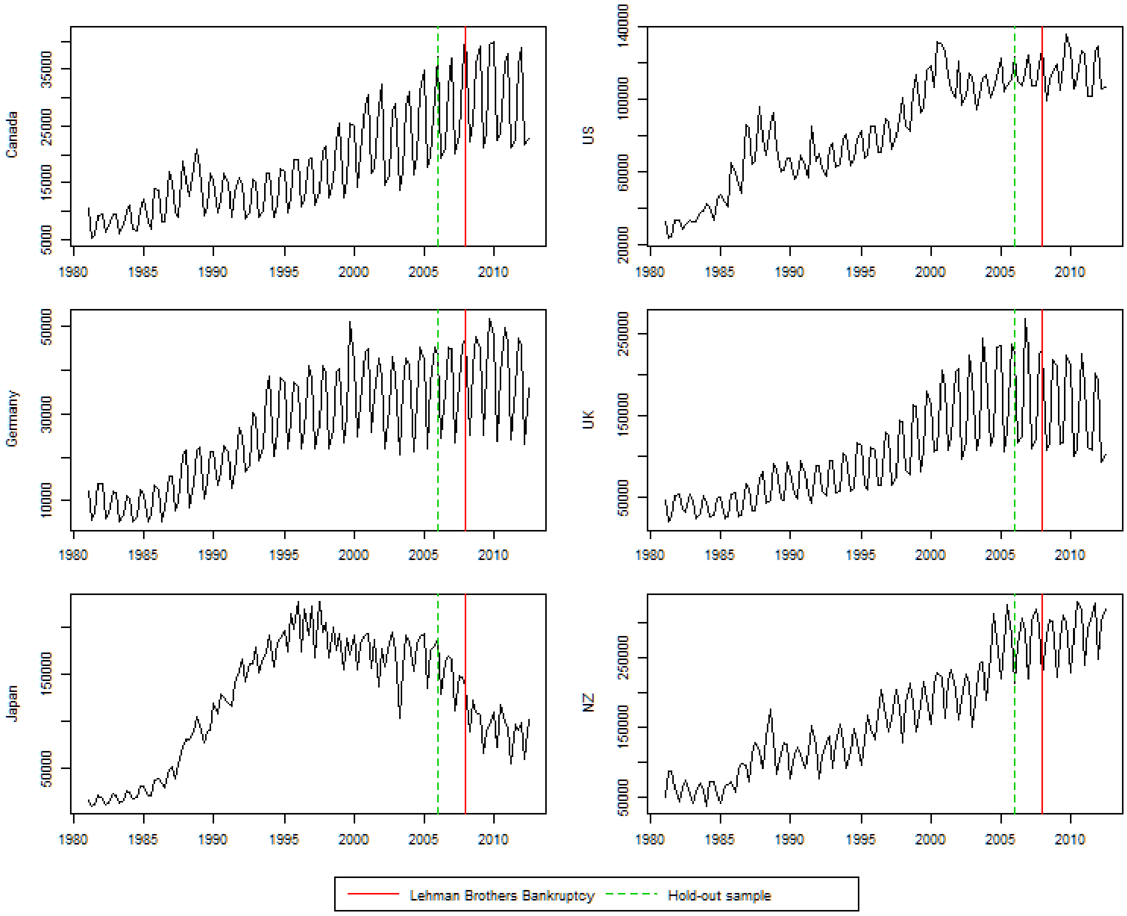

Figure 1: Natural logarithms of tourist arrivals to Australia.

Figure 2 is a seasonal plot for the tourist arrivals time series providing some visualisation. Observing these it becomes immediately obvious that seasonality between the first four source countries shows a consistent pattern with the January (which is the summer period for Australia) and the October (including the beginning of summer and the Christmas and New Year's holiday period) quarters being the peaks and the April and July quarters being the troughs. The one source country that is very different is New Zealand. Peak arrivals from New Zealand occur during the July quarter followed by the April and October quarters. Unlike all other source countries the trough clearly occurs during the January (summer) quarter. The seasonal plots are also useful revealing anomalies or one off events. For example in the US plot the peak arrivals for all July quarters occurs in 2000 during the Sydney Olympic games.

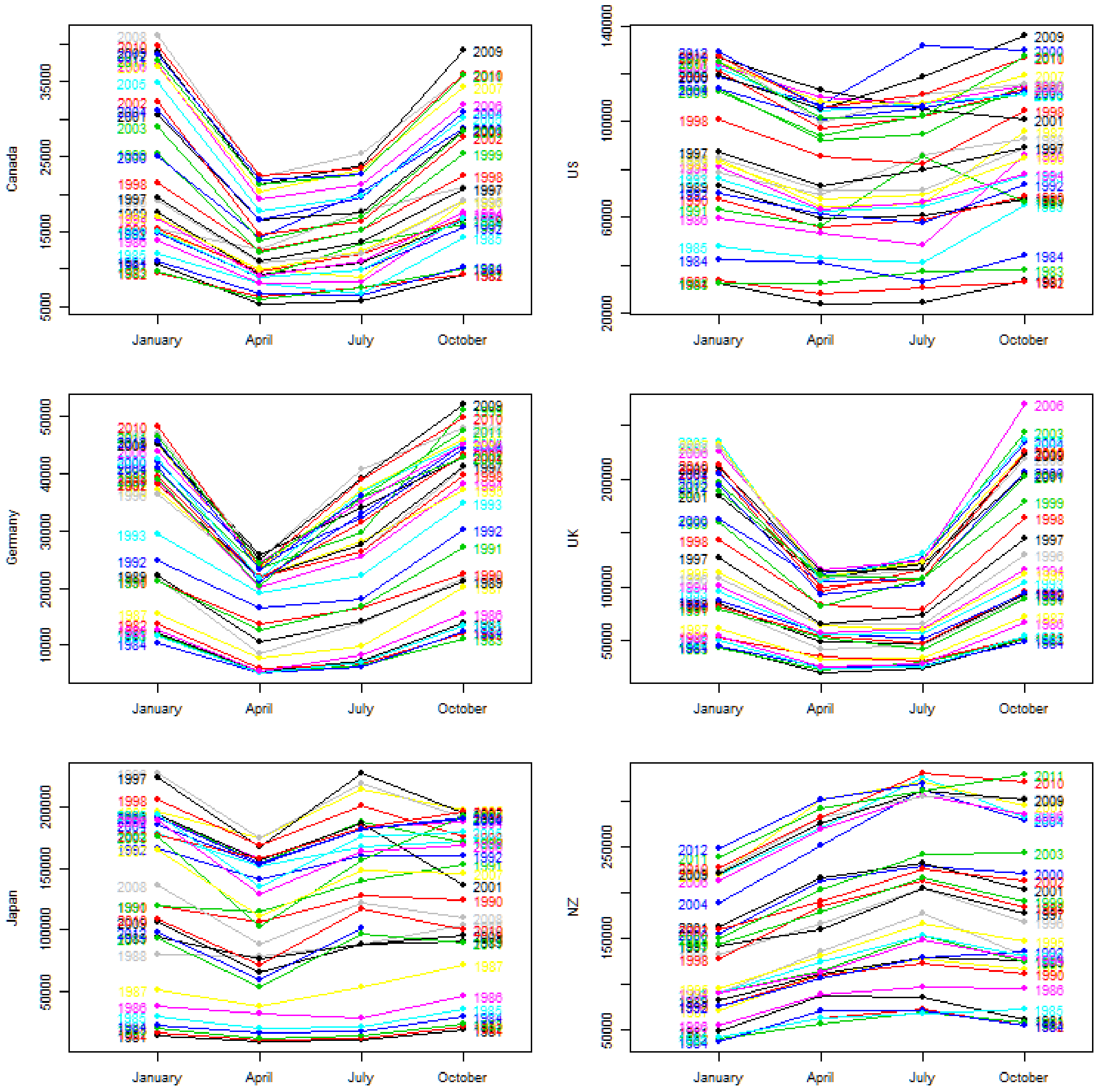

Figure 2: Seasonal plots of tourist arrivals to Australia.

5.2 Economic predictors

Law of demand states that, the demand for a good or service is inversely related to its price, ceteris paribus. This is measured by the *own price* variable defined as the ratio between CPIs and standardized by exchange rate

$$P_{i,t} = \frac{CPI_{AUD,t}/EX_{i,t}^{AUS}}{CPI_{i,t}}$$

where $i = 1, \dots, 6$ represents the six source counties, $CPI_{i,t}$ represents CPI of source country i at time t, and $EX_{i,t}^{AUS}$ is the exchange rate between Australian dollar and the currency of the origin country i. In addition, the demand of a good is also affected by the price of substitute and/or complementary goods. In the case of Australia, New Zealand seems to be a reasonable choice and we define the *substitute price* as

$$S_{i,t} = \frac{CPI_{NZ,t}}{EX_{i,t}^{NZD}}.$$

This predictor will obviously not be considered when New Zealand is the source country. Seasonally adjusted GDP in constant 2007 prices for the source countries using the expenditure approach and the unit currency of the source country is used as the *income variable*

$$GDP_{i,t}.$$

The final economic variable considered is the interest rate spread defined as

$TERMS_{i,t} = LIR_{i,t} \ - SIR_{i,t}$

where $LIR_{i,t}$ is the long term 10 year government bond and $SIR_{i,t}$ is the short term 90 days government bill of the source country. This variable has been widely used in the economic literature as a reliable predictor of future real economic activity and as a leading indicator to the business cycle (see Stock and Watson, 2003, and Anderson et al., 2007). We should note that we also considered other economic variables such as consumer confidence, oil prices, and share price, among others. However these were not regularly selected in the model building procedures and therefore we have not included them in any further analysis. All economic predictors were obtained from the OECD database.

5.3 Other variables

Beside the four economic predictors we also include seasonal dummies and two one-off dummy variables: one for the Sydney Olympics and one for the events of 9/11 which take the value of one for 2000Q3 and 2001Q3 respectively and zero otherwise. Dummies for other one off events such as the Bali bombings were not found to be selected.

4.4 Data transformations

In the modelling procedures presented in Section 2 both dependent variables and predictors need to be suitably transformed to stationary before entering the modelling framework. The time plots of the dependent variables clearly indicate a multiplicative and heteroscedastic seasonal pattern and therefore all variables are firstly log transformed using natural logs. All dependent variables are deemed to require seasonal and first order differencing with the exception of the US arrivals which only requires seasonal differencing. These decisions are reached following a sequence of seasonal (OCSB) and non-seasonal (ACF and KPSS) unit root tests and also after observing the ACF of the transformed series according to the tests. Table 2 summarises the decisions after the hypothesis tests and the final transformations implemented on each of the dependant variables in order to achieve stationarity.

After considering both unit root tests (ADF and KPSS), the economic predictors are transformed as follows: $\Delta\log(P_{i,t})$, $\Delta\log(S_{i,t})$, $\Delta\log(GDP_{i,t})$, and $\Delta TERMS_{i,t}$. In other words, we consider the growth rates for each predictor with the exception for interest rate spread that is already in percentages.

Table 2: variable transformations to achieve stationarity.

		OCSB	$\boldsymbol{\Delta^4}$	**ADF**	**KPSS**	$\boldsymbol{\Delta}$	**Transformation**
Canada	log	1	1	1	0	1	$\Delta^4\Delta log$
US	log	0	1	0	0	0	$\Delta^4 log$
Germany	log	1	1	1	1	1	$\Delta^4\Delta log$
UK	log	1	1	1	0	1	$\Delta^4\Delta log$
Japan	log	1	1	1	1	1	$\Delta^4\Delta log$
NZ	log	0	1	1	0	1	$\Delta^4\Delta log$

Note: the first column indicates that each series was logged. An entry of 1 (0) under the test columns indicates that a seasonal or non-seasonal unit root was (not) found. An entry of 1 (0) under the differencing columns indicates the action taken after also observing the ACF of the differenced series. The final column indicates the final transformation implemented.

5.5 Forecast evaluation procedure

Forecast evaluation is implemented in an expanding window setup. The holdout sample begins in 2006Q1. Therefore after transformations there are $T_{in} = 92$ observations (93 for the US) and $T_{out} = 27$ observations for estimation and test sets respectively. Introducing index i to matrix $\boldsymbol{A}$ we have

$$\boldsymbol{A} = \begin{bmatrix} y_{1+h} & \boldsymbol{z}_1' \\ \vdots & \vdots \\ y_{T_{in}+i-1} & \boldsymbol{z}_{T_{in}+i-1-h}' \end{bmatrix}.$$

A competing model is estimated and forecast $\hat{y}_{T_{in}+h+i-1|T_{in}+i-1}$ is generated. The window is expanded until the end of the sample, i.e., $i = 1, \dots, T_{out} - h + 1$. Therefore there are $T_{out} - h + 1$ forecasts for each h available for forecast evaluation. We should note that forecasts are back-transformed to levels before the forecast error measures that follow are calculated.

5.6 Measures of forecast accuracy

Let y_{t+h} and $\hat{y}_{t+h|t}$ be the ($t+h$)th observation and the h-step ahead forecast respectively. The ($t+h$)th forecast error is $e_{t+h} = y_{t+h} - \hat{y}_{t+h|t}$. We consider two measures for evaluating forecast accuracy the Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE),

$$MAPE_h = \frac{1}{H}\sum\nolimits_{H} \left|\frac{e_{t+h}}{y_{t+h}}\right| \times 100$$

$$RMSE_h = \sqrt{\frac{1}{H}\sum\nolimits_{H} e_{t+h}^2}$$

where $H = T_{out} - h + 1$ is the size of the hold-out sample. Note that RMSE is scale dependent while MAPE is scale independent.

5.7 Benchmarks

The first benchmark we consider is an AR model which also includes the full set of dummy variables defined as

$$y_{t+h} = \alpha + \boldsymbol{\beta}'\boldsymbol{l}_t + \boldsymbol{\delta}'\boldsymbol{d}_t + \varepsilon_{t+h}.$$

The dummy variables are always included in the AR model and hence only lagged dependent variables will be selected. The maximum number of lags is 4 and the number of lags selected is determined by minimizing the AIC. As the predictive models using the GETS approach has the exact same lagged dependent variables as the AR model, any forecasting accuracy improvement over the AR model can be contributed to the predictors. This forms one of our natural benchmarks.

The other benchmark we use is the seasonal random walk model represented by a $ARIMA(0,0,0)(0,1,0)_s$ where $s = 4$ is the seasonal period. A h-step-ahead forecast for this model is also defined as a seasonal naïve forecast and is equal to the final year's observation in the same seasonal period. As we will observe in the empirical results in many cases this is a challenging benchmark to beat. In what follows we refer to this benchmark as SNaïve.

6. Empirical results

Table 3 summarizes the three classes of competing models in the forecast evaluations that follow. The first class of models are the predictive regressions build using statistical inference in the GETS algorithm. The second class of models are predictive regression models built by using MPAs and the third class of models are our benchmarks. We apply bagging to both the first two classes of models in an effort to improve their forecast accuracy.

Table 3: A summary of the competing models

	No bagging	**With bagging**
GETS	p-value (0.05) p-value (0.01)	p-value (0.05) p-value (0.01)
MPA	AIC AICc BIC CV	AIC AICc BIC CV
Benchmarks	AR(p) SNaïve	

6.1 Does bagging improve accuracy in forecasting tourism demand?

The first question we ask from our forecast results is whether implementing bagging on the predictive regression models improves their out-of-sample forecast accuracy. The evidence from this study is overwhelmingly affirmative. Table 4 shows the percentage change in forecast accuracy measures before and after bagging. A negative (positive) entry indicates a percentage decrease (increase) in the forecast accuracy measure. Hence for bagging to be effective negative entries are required in Table 4.

In the majority of the cases across all six source countries bagging has improved forecasting accuracy of GETS considering both MAPE and RMSE. The improvements are larger in size for MAPE than RMSE. On average across all the counties, bagging improved the forecasting performance of GETS for each forecast horizon. Furthermore, bagging is especially effective for 1 and 2 steps ahead forecasts. For 1 and 2 steps ahead bagging improved the MAPE of GETS forecasts for 22 out of 24 cases (92%). For the RMSE, bagging improved forecast accuracy for 20 out of 24 cases (83%). In general, there were more improvements for GETS 0.05 compared to GETS 0.01. This is to be expected as the GETS algorithm using a p-value of 0.01 only choses the most significant predictors and therefore will not allow bagging to generate diverse forecasts that we average over that provides the advantage of bagging.

In general, similar to the results for GETS, bagging also improves forecast accuracy for the predictive models selected by MPAs. Bagging improved forecast accuracy the least for BIC which is not surprising. As a consistent and the most parsimonious criterion with a heavy penalty component compared to other MPAs and similar to the GETS using p-value of 0.01, the reduction in forecast diversity impairs the improvements that can be achieved by bagging. Also similarly to the GETS results, bagging is most effective for 1 to 2 steps ahead forecasts. For 1 step ahead for both MAPE and RMSE, all MPA models improved forecasting accuracy with the exception of the models for Canada.

6.2 Forecast evaluations of competing models

The four panels in Figure 3 show the MAPE and RMSE respectively across all six countries. These are complemented by the results presented in Table 5. A negative (positive) entry in Table 5 shows the percentage decrease (increase) in MAPE or RMSE over the SNaïve benchmark. A bold entry indicates that the predictive regression model is more accurate than the AR benchmark.

On immediately obvious observation from Figure 3 is that all the dashed lines are squeezed downwards when moving from the left hand side (LHS) panels where no bagging is implemented to the right hand side (RHS) panels where bagging is implemented. As expected the variance of the forecasts across the models is damped. The band of forecast error measure across the bagged alternatives is now much tighter than when bagging is not applied. All predictive regression models were more accurate than the SNaïve benchmark for all $h = 1$ to 4-steps-ahead even without bagging, when considering the MAPE. When considering the RMSE, other than for $h = 1$-step-ahead, is it only after bagging that predictive models become more accurate than the SNaïve benchmark (with some exceptions for $h = 3$). Furthermore it is only after bagging that the predictive regression models become more accurate than the AR benchmark for $h = 1$ and $h = 2$ when considering MAPE and $h = 1$ when considering the RMSE.

Table 4: The percentage difference in forecast accuracy when bagging is implemented to the predictive regression models compared to no bagging.

	MAPE						RMSE					
	GETS		MPA				GETS		MPA			
h	0.05	0.01	CV	AIC	AICc	BIC	0.05	0.01	CV	AIC	AICc	BIC
Averages across all six countries												
1	-13.0	-9.2	-11.8	-8.2	-10.9	-7.1	-11.8	-5.8	-11.3	-11.6	-10.7	-8.0
2	-10.8	-2.5	-2.6	-2.1	-2.9	-7.9	-8.4	-0.1	-4.8	-3.5	-4.6	-2.5
3	-10.0	3.6	-0.6	15.1	9.3	6.8	-6.6	-3.2	-3.0	4.4	2.3	5.7
4	-12.3	0.7	-2.4	14.1	11.7	4.8	-11.3	-1.5	-2.6	9.4	10.5	8.1
Av	-11.5	-1.9	-4.3	4.7	1.8	-0.8	-9.5	-2.7	-5.4	-0.3	-0.6	0.8
Canada												
1	-10.0	-16.0	1.6	1.1	1.0	18.9	-11.9	-16.4	6.1	6.8	4.5	11.4
2	-0.2	2.2	-12.1	-4.9	-4.8	1.0	10.6	3.0	-0.8	15.6	14.5	9.5
3	-13.5	-1.2	-16.9	8.2	1.2	12.7	12.6	10.1	9.3	24.7	16.3	17.7
4	25.0	29.8	0.8	29.7	25.5	27.2	8.5	22.1	4.5	20.0	17.1	26.0
Av	0.3	3.7	-6.6	8.5	5.7	15.0	4.9	4.7	4.8	16.8	13.1	16.1
Germany												
1	-17.9	-9.3	-8.7	-13.6	-17.8	-12.5	-19.9	-5.1	-10.7	-12.2	-13.4	-5.3
2	-13.9	-13.0	4.9	9.0	8.5	6.9	-12.0	2.4	2.8	6.0	4.3	1.3
3	0.7	2.0	-6.2	21.9	18.3	21.0	0.3	-2.7	-3.0	17.7	17.2	12.8
4	-5.2	-4.1	-9.5	28.0	28.3	21.2	-4.7	-2.9	-7.9	16.1	21.5	14.5
Av	-9.1	-6.1	-4.9	11.3	9.3	9.2	-9.1	-2.1	-4.7	6.9	7.4	5.8
Japan												
1	-9.0	-5.7	-15.6	-4.7	-6.2	0.9	-7.0	-4.3	-12.1	-4.3	-6.1	-1.8
2	-4.7	-18.0	2.4	-9.5	-11.3	-15.5	-8.0	-13.8	1.6	-6.2	-7.7	-8.1
3	-8.7	-9.2	-8.3	13.4	15.6	-1.5	-7.8	-5.8	-0.2	10.0	11.1	1.1
4	-7.8	-2.8	-10.9	3.2	6.0	4.4	-11.2	-4.2	-8.2	5.0	9.2	4.9
Av	-7.5	-8.9	-8.1	0.6	1.0	-2.9	-8.5	-7.0	-4.7	1.1	1.6	-1.0
NZ												
1	-9.9	-3.6	-18.9	-20.9	-19.7	-14.9	-14.8	0.0	-16.0	-17.0	-15.1	-13.9
2	-9.6	10.7	-8.8	-8.6	-7.1	4.3	-5.6	22.6	-3.8	-2.8	-2.4	6.0
3	-8.4	-13.5	-15.6	-11.1	-8.6	13.2	-9.2	-7.3	-6.2	-4.6	-1.4	21.6
4	-17.3	-18.6	-13.8	13.6	13.0	14.7	-6.0	-9.6	-3.0	21.7	22.5	16.5
Av	-11.3	-6.2	-14.3	-6.7	-5.6	4.3	-8.9	1.4	-7.3	-0.7	0.9	7.6
UK												
1	-6.4	-9.2	-6.8	-7.6	-6.2	-8.2	-4.0	-7.2	-4.3	-6.7	-3.9	-7.0
2	-7.6	-3.7	0.3	-4.6	-6.0	-5.7	-4.8	-1.8	2.1	-2.1	-3.1	-2.7
3	0.2	-1.7	7.1	10.6	6.0	0.2	1.6	0.1	3.5	6.2	3.1	0.9
4	-15.2	-4.3	4.4	10.1	5.0	-3.7	-13.1	-0.2	5.4	5.0	3.2	-1.7
Av	-7.2	-4.7	1.3	2.1	-0.3	-4.3	-5.1	-2.3	1.7	0.6	-0.2	-2.6
US												
1	-27.5	-22.1	-17.1	-27.7	-21.7	-24.5	-20.9	-14.5	-16.4	-21.7	-21.0	-11.0
2	-24.2	-12.4	-26.5	-20.9	-24.6	-20.8	-24.7	-12.8	-30.7	-31.5	-33.4	-21.3
3	-13.3	-13.3	-18.3	-8.6	-15.6	-14.9	-20.7	-14.1	-21.5	-27.5	-32.4	-19.6
4	-29.2	-18.5	-8.5	-15.9	-16.1	-17.4	-25.9	-14.3	-6.1	-11.5	-10.1	-11.8
Av	-23.5	-16.6	-17.6	-18.3	-19.5	-19.4	-23.0	-13.9	-18.7	-23.1	-24.2	-15.9

Note: A negative (positive) entry indicates the percentage (%) decrease (increase) in MAPE or RMSE when bagging is applied to each model.

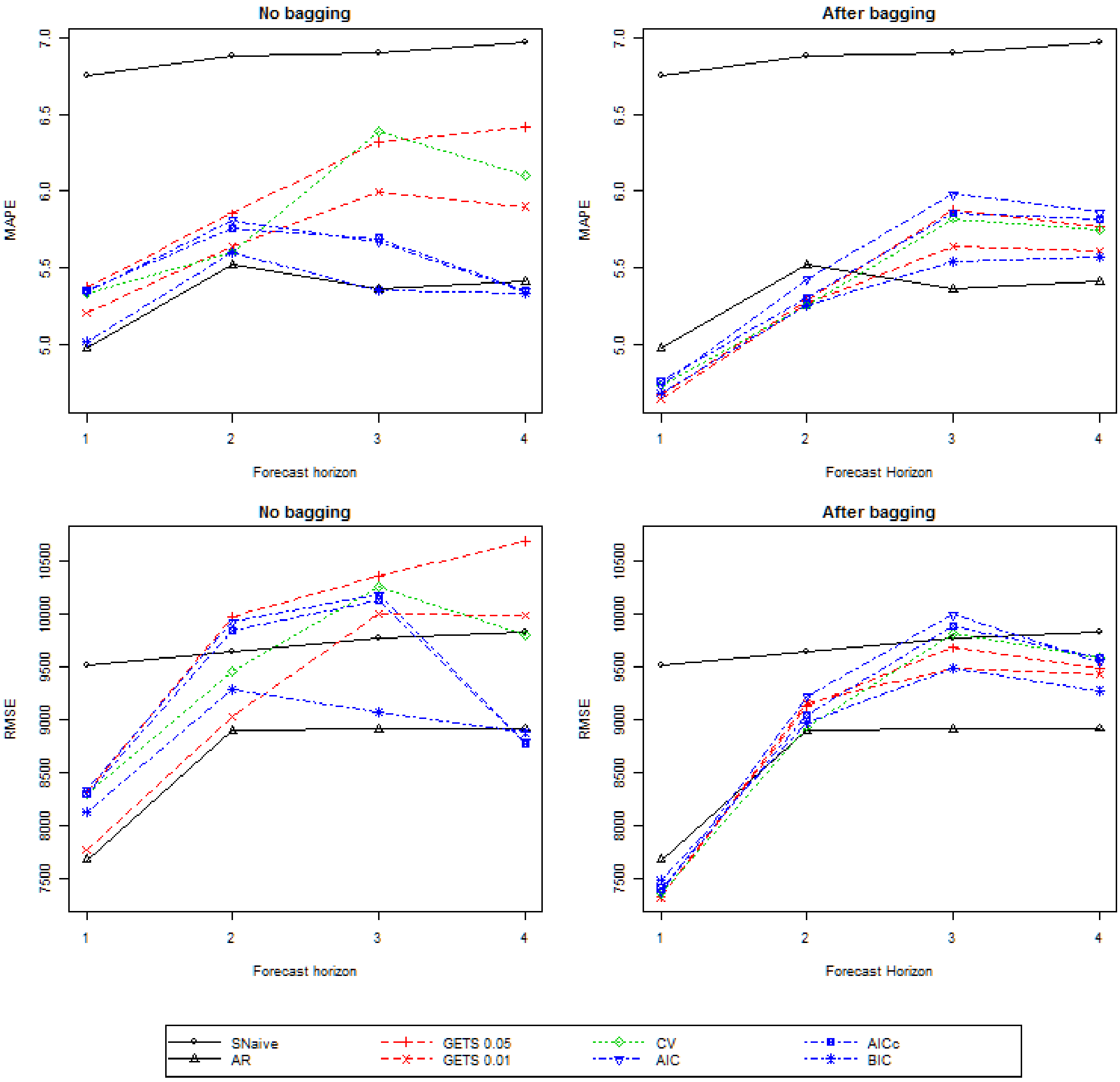

Figure 3: MAPE and RMSE for $h = 1$ to 4-steps-ahead

Table 5: Percentage (%) difference in MAPE and RMSE between the seasonal Naïve benchmark and the predictive regression models.

		GETS		MPA				GETS		MPA			
h	*AR*	0.05	0.01	CV	AIC	AICc	BIC	0.05	0.01	CV	AIC	AICc	BIC
	MAPE (no bagging)							MAPE (after bagging)					
1	-26.3	-20.4	-23.0	-21.1	-20.9	-20.8	-25.8	**-30.8**	**-31.2**	**-30.0**	**-30.0**	**-29.6**	**-30.7**
2	-19.8	-15.0	-18.0	-18.6	-15.6	-16.4	-18.7	**-23.2**	**-23.6**	**-23.7**	**-21.2**	**-22.9**	**-23.7**
3	-22.4	-8.5	-13.2	-7.5	-17.9	-17.6	-22.4	-15.0	-18.4	-15.7	-13.4	-15.3	-19.8
4	-22.4	-7.9	-15.3	-12.4	**-23.3**	**-23.2**	**-23.4**	-17.3	-19.6	-17.5	-15.9	-16.6	-20.1
Av	-22.7	-13.0	-17.4	-14.9	-19.4	-19.5	-22.6	-21.6	**-23.2**	-21.7	-20.1	-21.1	**-23.6**
	RMSE (no bagging)							RMSE (after bagging)					
1	-19.3	-12.7	-18.3	-12.7	-12.4	-12.8	-14.5	**-23.0**	**-22.0**	**-30.0**	**-22.7**	**-20.7**	**-21.0**
2	-7.7	3.4	-6.3	-2.0	2.9	2.0	-3.6	-6.0	-2.2	**-23.7**	-3.7	-0.4	-1.0
3	-8.8	6.0	2.4	5.0	4.2	3.7	-7.1	-4.9	2.2	-15.7	4.6	7.5	3.1
4	-9.3	8.7	2.4	-0.3	**-10.6**	**-10.7**	**-9.7**	-4.8	-3.6	-17.5	-2.9	-1.3	-1.6
Av	-11.3	1.4	-5.2	-2.5	-4.0	-4.4	-8.7	-9.7	-6.4	-21.7	-6.2	-3.7	-5.1

A negative (positive) entry indicates a decrease (increase) in the error measure compared the seasonal Naïve benchmark. Bold entries indicate that the predictive regression model has performed better than the AR benchmark.

Taking a closer look at the forecast results on a country by country basis we identify two extreme cases. These are the cases of Japan and New Zealand. Tourist arrivals from Japan show a general downwards strong trend since 1995. This makes the SNaïve forecasts inappropriate and the predictive regression models perform much better than this benchmark for Japan. On the other hand there is New Zealand where the SNaïve convincingly beats the predictive regression models. We should note that the individual country results are not presented here to save space but they are available upon request. New Zealand is a special case for Australia. Although it is considered international travel the economic and social ties between the two countries makes it arguably domestic in nature. This is initially identified by the seasonal plots shown in Figure 2. Furthermore, there have been discussions in the political arena since 2009 contemplating changing the actual status of the travel between the two countries to domestic (see for example among others David Stone's article in The New Zealand Herald, Stone, 2009,).Therefore it may be the case that the economic variables used in the demand model may not be the best predictors for tourist arrivals from New Zealand as these variables may have a much smaller effect on domestic type travel. We should note that no matter what we did with New Zealand and for which period we looked at, the predictive regression models could not forecast any more accurately than the SNaïve benchmark. Given these results we continue our analysis by removing these two countries from the results in order to remove the biases these two countries introduce.

Figure 4 and Table 6 show the MAPE and RMSE across the four countries (we have now removed Japan and New Zealand). We notice now that the SNaïve benchmark has become a much more relevant and not easy to beat benchmark. For both MAPE and RMSE the predictive regression models forecast more accurately than the SNaïve benchmark only for $h = 1$-step-ahead. These improvements over the SNaïve benchmark become substantial after bagging as percentage improvements range between 15% and 19% as shown in Table 6. All predictive regression models are more accurate than the AR benchmark for $h = 1$ and $h = 2$ for MAPE. The best performing models for MAPE is the GETS(0.01) and the model selected by BIC which after bagging both are more accurate than the AR benchmarks for all forecast horizons. This is also the case for these two models when considering the RMSE. In this case the model selected by CV is also more accurate than the AR benchmark for all forecast horizons.

Table 6: Percentage (%) difference in MAPE and RMSE between the seasonal Naïve benchmark and the predictive regression models excluding NZ and Japan.

		GETS		MPA				GETS		MPA			
h	*AR*	0.05	0.01	CV	AIC	AICc	BIC	0.05	0.01	CV	AIC	AICc	BIC
	MAPE (no bagging)							MAPE (after bagging)					
1	-9.9	-0.8	-5.4	-6.4	-2.1	-3.5	**-10.4**	**-15.6**	**-18.5**	**-13.0**	**-13.0**	**-13.7**	**-16.9**
2	8.3	15.9	**7.2**	11.5	13.0	12.0	**6.0**	**2.7**	**-0.1**	**2.3**	**6.3**	**3.7**	**0.4**
3	7.1	22.3	11.3	23.1	10.0	11.9	**3.8**	15.0	7.3	13.7	18.1	14.0	**6.8**
4	7.9	19.3	8.3	11.9	**2.1**	**2.4**	**2.9**	**7.9**	**5.3**	9.1	13.3	10.8	**5.3**
Av	3.3	14.2	5.3	10.0	5.7	5.7	**0.6**	**2.5**	**-1.5**	**3.0**	6.2	3.7	**-1.1**
	RMSE (no bagging)							RMSE (after bagging)					
1	-9.4	-6.0	**-9.6**	-8.7	-5.8	-6.8	-8.7	**-17.1**	**-18.7**	**-15.6**	**-15.8**	**-15.6**	**-15.3**
2	15.4	20.5	**12.8**	16.5	23.3	22.9	16.5	**8.2**	**8.2**	**6.5**	**10.8**	**8.6**	**8.7**
3	15.9	22.7	17.2	20.6	23.4	25.9	20.6	17.1	**13.1**	**15.3**	19.2	17.1	**12.9**
4	15.6	26.0	**14.0**	**7.8**	**7.2**	**7.6**	**7.8**	**7.7**	**10.2**	**8.9**	**9.9**	**9.8**	**9.5**
Av	9.4	15.8	**8.6**	**9.1**	12.0	12.4	**9.1**	**4.0**	**3.2**	**3.8**	**6.0**	**5.0**	**3.9**

Note: a negative (positive) entry indicates a decrease (increase) in the error measure compared the seasonal Naïve benchmark. Bold entries indicate that the predictive regression model has performed better than the AR benchmark.

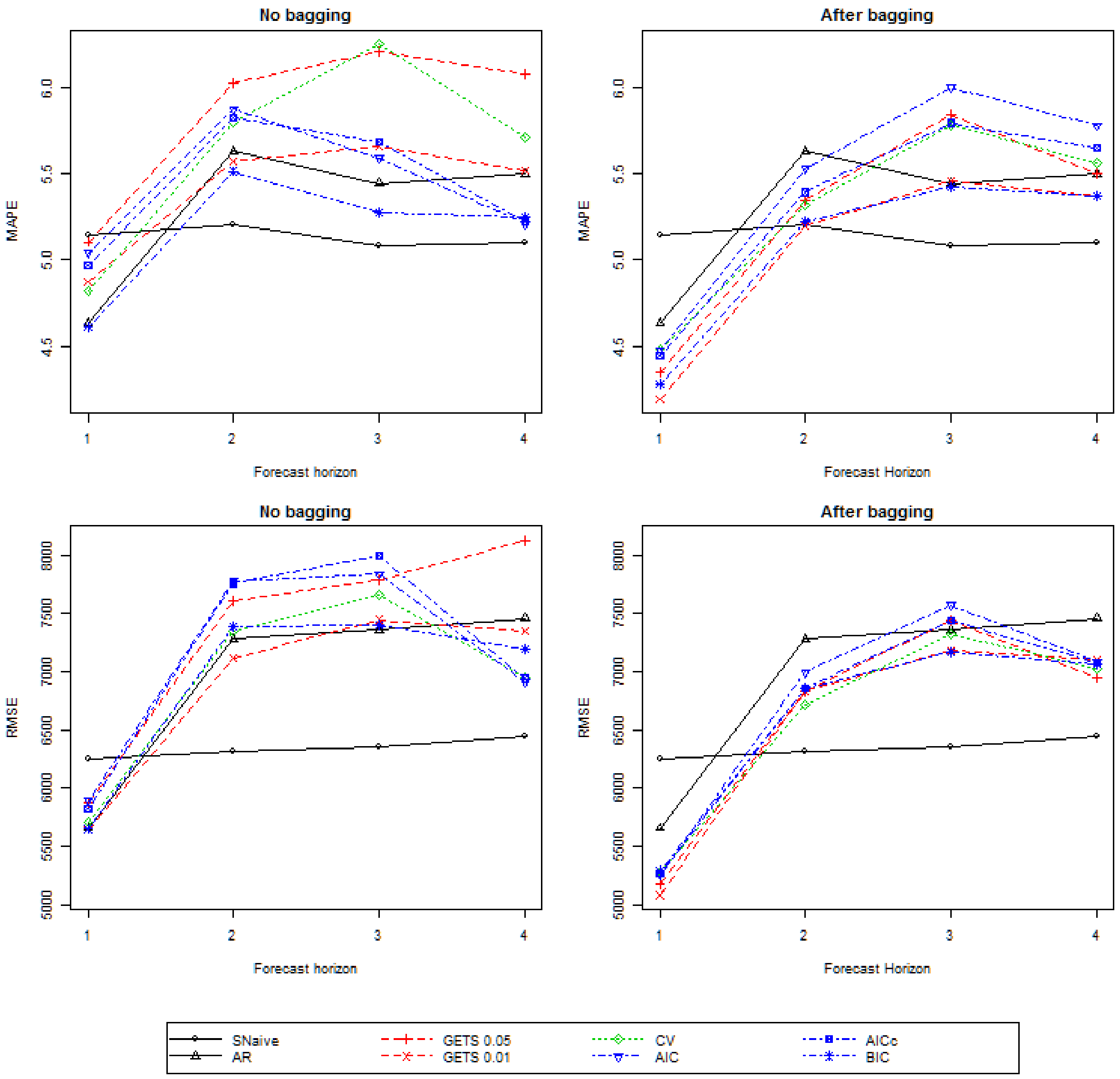

Figure 4: MAPE and RMSE for $h = 1$ to 4-steps-ahead excluding New Zealand and Japan.

We next ask the question of why do the predictive regression models only forecast well for $h = 1$ and at most 2-steps-ahead. Where do they break down? In what follows we repeat the forecasting evaluation of above but now we consider the pre- and post-period of the Lehman Brothers Bankruptcy (LBB) and the beginning of the GFC. The top two panels in Figure 5 show the results for MAPE for the pre-LBB period and the bottom two panels show the MAPE results for the post-LBB period. A quick glance at the plots and they are in stark contrast to each other. The first point of notice is that for the pre-LBB period the AR is the more accurate benchmark of the two. This is reversed for the post-LBB period although for $h = 1$ the two benchmarks are very close to each other. Once again the most parsimonious predictive regression models, i.e., GETS(0.01) and models selected by BIC, are among the most accurate both before and after bagging. Both these models are more accurate than both benchmarks for all forecast horizons in the pre-LBB period after bagging has been implemented. In the case of the models selected by BIC these are more accurate than the benchmarks even before bagging. We should note that the models selected by the BIC show a 12.8% improvement in MAPE over the SNaïve after bagging, compared to 12.4% improvement before bagging. This small improvement is driven by the substantial improvement for $h = 1$-step-ahead, jumping from 20.4% to 27.3%.

The results for the post-LBB period clearly show the failure of the predictive regression models to forecast any more accurately than either benchmark. This possibly highlights a breakdown in the predictive power invested in the economic predictors and their relationships with tourism demand. In this case the results also clearly show that bagging does not in any way improve forecast accuracy. In fact it actually hinders it. For example the model selected by BIC performs closely and if anything better than the AR benchmark but only before bagging.

Table 7: Percentage (%) difference in MAPE between the seasonal Naïve benchmark and the predictive regression models for the pre- and post-LBB periods excluding NZ and Japan.

		GETS		MPA				GETS		MPA			
h	*AR*	0.05	0.01	CV	AIC	AICc	BIC	0.05	0.01	CV	AIC	AICc	BIC
	Pre-LBB (no bagging)							Pre-LBB (after bagging)					
1	-16.8	**-16.9**	**-19.4**	**-18.3**	-15.5	-9.4	**-20.4**	**-28.0**	**-31.4**	**-28.0**	**-25.6**	**-27.8**	**-27.3**
2	-4.6	-0.6	-3.8	**-11.2**	-1.8	-2.0	**-14.6**	**-12.5**	**-13.3**	**-14.1**	**-10.8**	**-13.6**	**-12.7**
3	-1.9	6.4	**-4.8**	3.5	**-2.0**	**-2.8**	**-7.2**	2.4	**-2.1**	1.0	8.6	-0.4	**-3.4**
4	-7.2	-6.2	-8.1	**-8.8**	**-9.1**	**-9.1**	**-8.1**	-5.2	**-8.1**	-4.8	-3.0	-2.8	-7.8
Av	-7.6	-4.4	-9.1	**-8.7**	-7.1	-5.8	**-12.6**	**-10.8**	**-13.7**	**-11.5**	-7.7	**-11.1**	**-12.8**
	Post-LBB (no bagging)							Post-LBB (after bagging)					
1	-1.8	16.1	6.1	10.3	16.1	16.7	-1.4	10.1	1.1	10.9	15.0	13.1	4.5
2	13.5	24.7	**11.9**	27.2	24.5	24.6	13.9	22.0	13.5	24.5	26.4	25.3	14.2
3	8.8	28.0	14.5	34.1	19.1	22.5	**8.2**	35.5	13.4	32.9	37.0	34.0	13.0
4	11.8	25.9	12.4	22.7	**4.9**	**4.9**	**6.0**	19.5	10.9	19.5	25.2	22.1	12.1
Av	8.1	23.7	11.2	23.6	16.2	17.2	**6.7**	21.8	9.7	21.9	25.9	23.6	11.0

Note: a negative (positive) entry indicates a decrease (increase) in MAPE compared the seasonal Naïve benchmark. Bold entries indicate that the predictive regression model has performed better than the AR benchmark.

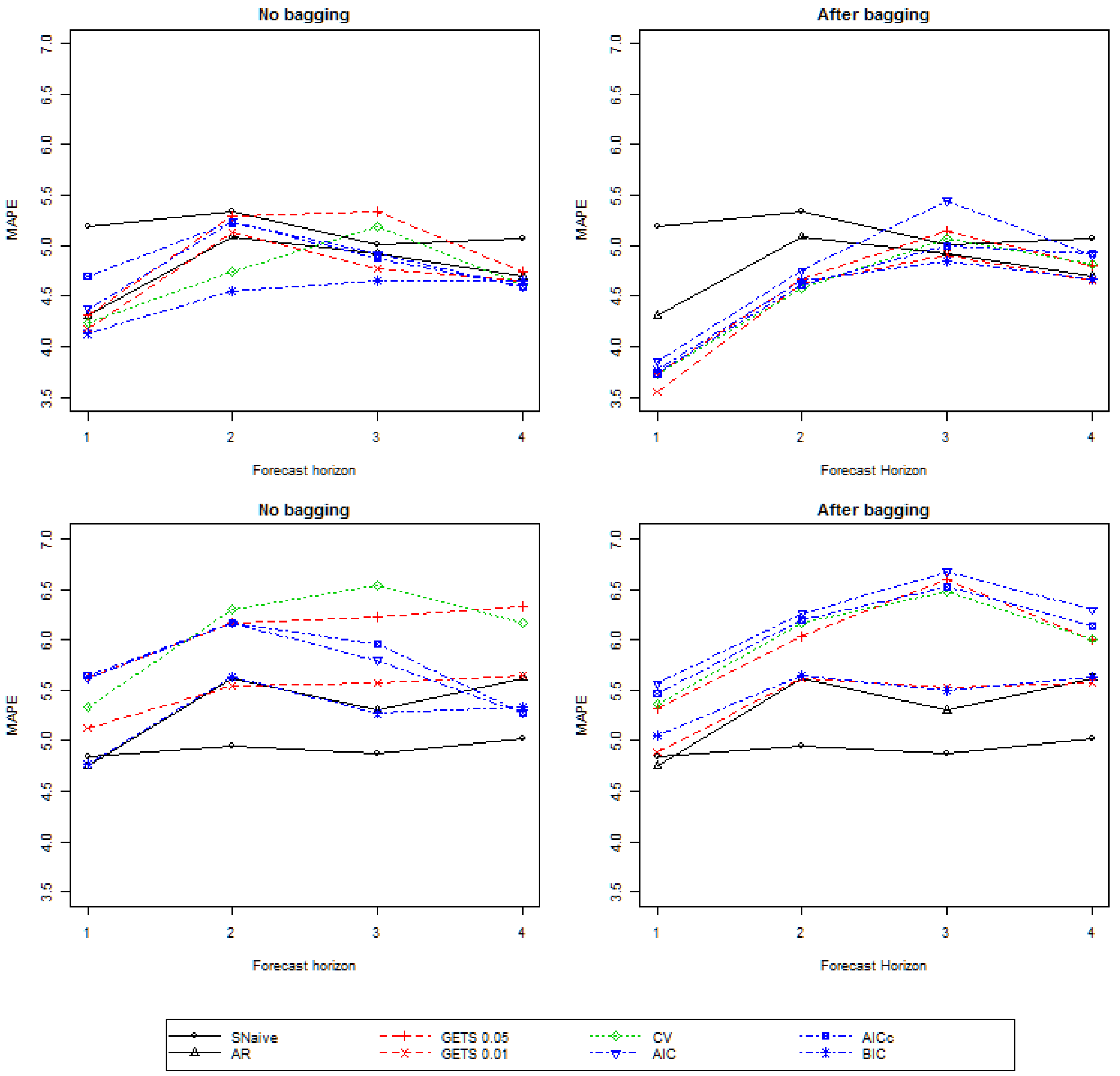

Figure 5: MAPE for $h = 1$ to 4-steps-ahead for the pre- and post- LBB period excluding New Zealand and Japan.

7 Concluding remarks

In this study we examine fully automated model selection procedures in an effort to improve the out-of-sample forecast accuracy of causal econometric models commonly used in both academia and industry to forecast tourism demand. We consider two broad variable selection procedures. In the first procedure known as the general-to-specific approach predictors are selected using individual t-statistics. In the second procedure predictors are selected based on measures of predictive accuracy. The measures we consider are the AIC, the bias corrected AICc, the BIC and the cross-validation statistic. We complement these procedures with bootstrap aggregation a procedure which aims to improve the predictive accuracy of the models based on resampling. We find overwhelming evidence that bagging improves the out-of-sample forecasting accuracy of the predictive regression models for predicting tourism demand, so much so that in many cases it is only after bagging that the predictive regressions become more accurate than the simple benchmarks. We recommend that such machine learning procedures (boosting is another alternative) should be considered in the field of tourism and should be implemented in similar situations.

The empirical results based on tourist arrivals data from six source markets suggested that the common economic factors used in the literature such as, prices, exchange rates, output are mostly reliable predictors for international tourism demand in Australia up to 2008Q3. 2008Q3 marks the Lehman Brothers Bankruptcy (LBB) which triggered the beginning of the Global Financial Crisis (GFC). In general the most accurate were the models built implementing the general-to-specific procedure with a p-value 0.01 and the models built using BIC as the measure of predictive accuracy. These findings are in line with the principle of parsimony which dictates that the more parsimonious models are often best for forecasting. We should note that these models were most accurate and bagging was most effective for one-step and two-steps ahead forecasts during the pre-LBB period.

For the post-LBB period no predictive models were more accurate than the simple benchmarks; an AR model and a seasonal random walk. Furthermore, it was only for the post-LBB period that implementing bagging did not improve the forecast accuracy of the predictive regression models. It seems that the shock of the GFC may have significantly disturbed the economic relationships between tourism demand and the typically employed economic predictors. Obviously this observation is based on Australian data and is limited within the general automated modelling framework employed here and therefore should be treated cautiously. There is scope in future studies to expand this automated framework to other countries. Furthermore, in an effort to improve forecast accuracy of predicate regressions per se there may scope of relaxing somewhat the fully automated nature of the procedure and target source country specific modelling challenges where effects of shocks such as the GFC can be modelled using more flexible functional forms such as non-linear processes or linear models accounting for structural breaks.

References

Anderson, H. M., Athanasopoulos, G., & Vahid, F. (2007). Nonlinear autoregressive leading indicator models of output in G-7 countries. *Journal of Applied Econometrics*, *22*, 63–87.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822–844.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*(6), 2350–2383.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, *30*(4), 927–961.

Crouch, G. (1992). Effect of Income and Price on International Tourism. *Annals of Tourism Research*, 19, 643-664.

Gonçalves, S., & Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, *123*(1), 89–120.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer-Verlag.

Hendry, D. F., & Krolzig, H. M. (2005). The properties of automatic gets modelling. *Economic Journal*, *115*(502), 32–61.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice* (1st ed.). OTexts.

Inoue, A., & Kilian, L. (2008). How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation. *Journal of the American Statistical Association*, *103*(482), 511–522.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Katircioglu, S. T. (2009). Revisiting the Tourism-led-growth Hypothesis for Turkey Using the Bounds Test and Johensen Approach for Cointegration. *Tourism Management*, *30*(1), 17-20.

Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, *41*(9), 4235–4244.

Li, G., K. K. F. Wong, H. Song and S. F. Witt (2006). Tourism Demand Forecasting: A Time Varying Parameter Error Correction Model. *Journal of Travel Research*, *45* (2), 175-185.

Li, G. and H. Song (2007). New Forecasting Models. *Journal of Travel and Tourism Marketing*, 21 (4), 3-13.

Li, G., H. Song and S. F. Witt (2005). Recent Developments in Econometric Modeling and Forecasting. *Journal of Travel Research, 44* (1), 82-99.

Narayan, P. K. (2004). Fiji's Tourism Demand: The ARDL Approach to Cointegration. *Tourism Economics*, *10* (2), 193-206.

Newey, W., & West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.

Rapach, D. E., & Strauss, J. K. (2012). Forecasting US state-level employment growth: An amalgamation approach. *International Journal of Forecasting*, *28*(2), 315–327.

Rapach, D. E., & Strauss, J. K. (2010). Bagging or Combining (or Both)? An Analysis Based on Forecasting U.S. Employment Growth. *Econometric Reviews*, *29*(5-6), 511–533.

Song, H., Gao, B. Z., & Lin, V. S. (2012). Combining statistical and judgmental forecasts via a web-based tourism demand forecasting system. *International Journal of Forecasting*, *29*(2), 295–310.

Song, H. and S. Lin (2010). Impacts of the Financial and Economic Crisis on Tourism in Asia. *Journal of Travel Research, 49* (1), 16-30.

Song, H. and G. Li (2008). Tourism Demand Modelling and Forecasting- A Review of Recent Research. *Tourism Management*. 29 (2), 203-220.

Song, H. and L. Turner (2006). Tourism Demand Forecasting. In L. Dwyer, & P. Forsyth (Eds.), *International Handbook on the Economics of Tourism*. Cheltenham, Edward Elgar.

Song, H. and S. F. Witt (2003), General-to-Specific Modeling to International Tourism Demand Forecasting. *Journal of Travel Research, 42* (1), 65-74.

Song, H., S. F. Witt and G. Li (2009). *The Advanced Econometrics of Tourism Demand*. London, Routledge.

Song, H., S. F. Witt and X. Y. Zhang (2008). A Web-based Tourism Demand Forecasting System. *Tourism Economics, 14* (3), 445-468.

Stock J.H., Watson, M.W. (2003). Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, 41, 788–829.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, *5*(4), 595–645.

Stone, D. (2009). Borderless Tasman still far away, *The New Zealand Herald*, March 13.

Tourism Research Australia (2014), Tourism’s Contribution to the Australian Economy, 1997–98 to 2012–13, Tourism Research Australia, Canberra.

World Tourism Organization. (2014). *UNWTO Tourism Highlights*.

Wang, Y. S. (2009). The Impact of Crisis Events and Macroeconomic Activity on Taiwan’s International Inbound Tourism Demand. *Tourism Management, 30* (1), 75-82.

Witt, S. F., Song, H. and P. Louvieris (2003). Statistical Testing in Forecasting Model Selection. *Journal of Travel Research* Vol. 42, No. 2, pp151-158.