Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

International Journal of Forecasting 27 (2011) 887–901

international journal of forecasting

www.elsevier.com/locate/ijforecast

Beyond point forecasting: Evaluation of alternative prediction intervals for tourist arrivals

Jae H. Kim[a,*], Kevin Wong[b], George Athanasopoulos[c], Shen Liu[d]

[a] School of Economics and Finance, La Trobe University, VIC 3086, Australia
[b] School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hong Kong
[c] Department of Econometrics and Business Statistics and Tourism Research Unit, Monash University, Clayton, VIC 3800, Australia
[d] Department of Econometrics and Business Statistics, Monash University, Caulfield East, VIC 3145, Australia

Available online 8 June 2010

Abstract

This paper evaluates the performances of prediction intervals generated from alternative time series models, in the context of tourism forecasting. The forecasting methods considered include the autoregressive (AR) model, the AR model using the bias-corrected bootstrap, seasonal ARIMA models, innovations state space models for exponential smoothing, and Harvey's structural time series models. We use thirteen monthly time series for the number of tourist arrivals to Hong Kong and Australia. The mean coverage rates and widths of the alternative prediction intervals are evaluated in an empirical setting. It is found that all models produce satisfactory prediction intervals, except for the autoregressive model. In particular, those based on the bias-corrected bootstrap perform best in general, providing tight intervals with accurate coverage rates, especially when the forecast horizon is long.

Keywords: Automatic forecasting; Bootstrapping; Interval forecasting

1. Introduction

Tourism forecasting is an area of enormous interest for both academics and practitioners. A large number of studies have compared the forecast accuracies of alternative econometric or time series models for forecasting tourism demand. Li, Song, and Witt (2005), Song and Li (2008) and Witt and Witt (1995) provide comprehensive reviews of this issue. These reviews identify the application of time series methods as a key innovation in this area. According to Song and Li (2008), the main question in past studies on tourism forecasting has been whether or not one could establish a forecasting principle for academics and practitioners; that is, whether one can identify any models or methods that consistently generate more accurate forecasts than others in practice. However, the results from past studies are rather mixed and often conflicting. Li

[*] Corresponding author. Tel.: +61 3 94796616; fax: +61 3 94791654.
E-mail address: J.Kim@latrobe.edu.au (J.H. Kim).

0169-2070/$ - see front matter © 2010 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2010.02.014

et al. (2005) state that no single forecasting method outperforms the alternatives in small samples, while Song and Li (2008) conclude that there has not been a panacea for tourism demand forecasting.

An important point to note from past studies is their preoccupation with point forecasting. To the best of our knowledge, all of the studies discussed in the above-mentioned reviews restrict their attention exclusively to point forecasting. A point forecast is a single number which is an estimate of the unknown true future value. Although it is the most likely realization of the possible future values implied by the estimated model, it provides no information as to the degree of uncertainty associated with the forecast. For this reason, one may justifiably argue that the comparison of alternative point forecasts is of limited use, since it completely neglects the variability associated with forecasting. For an improved and more meaningful comparison of the performances of forecasting models, the degree of uncertainty associated with producing the forecasts should be taken into account explicitly.

Our focus in this paper is on interval forecasting for tourism demand. An interval forecast (or prediction interval) indicates a range of possible future outcomes with a prescribed level of confidence.[1] As Chatfield (1993) and Christoffersen (1998) point out, interval forecasts are of greater value to decision-makers than point forecasts, and should be used more widely in practical applications, as they allow for a thorough evaluation of the future uncertainty, and for contingency planning. We identify tourism forecasting as an area where interval forecasting can add high marginal utility. This is because practitioners and government agencies in the tourism industry actively use the forecasts from time series models as an input to their decision-making, in relation to planning, marketing, and the provision of infrastructure (see, for example, the web-based tourism demand forecasting system detailed by Song, Witt, & Zhang, 2008). The provision of prediction intervals will benefit government bodies, destination marketing authorities and business investors in tourism, who will be able to resort to more *flexible* tourism infrastructure planning and improved *fine tuning* of resources and tourism policies, given that interval forecasts offer a "*range of possible values* that one is quite certain will include the actual value produced by time" (Frechtling, 2001). This is reinforced by the fact that "forecasts cannot be expected to be perfect, and intervals emphasize this" (Makridakis, Wheelwright, & Hyndman, 1998). Interval forecasts also allow tourism planners and managers to check for hypothetical inventory holdings under different probable market and sales outcomes.

In this paper, we conduct an extensive comparison of the accuracy of prediction intervals in the context of tourism forecasting. Our aim is to provide the first empirical evidence within the tourism forecasting literature as to whether popular time series methods are useful for generating accurate prediction intervals, and which method should be preferred in practice. We employ a set of monthly time series for the number of tourist arrivals to Hong Kong and Australia. For the former, we consider the number of tourist arrivals from four individual source markets (Australia, China, the UK, and the US) and three aggregated markets (Asia, Europe, and the total) from 1985 to 2008. For the latter, we consider those from four individual source markets (Germany, New Zealand, the UK, and the US) and two aggregated markets (Europe and the total) from 1980 to 2007.

We consider two types of prediction intervals based on the autoregressive (AR) model: the conventional interval using a normal approximation, and a bias-corrected bootstrap version proposed by Kim (2004). In addition, we consider innovations state space models for exponential smoothing, as presented by Hyndman, Koehler, Ord, and Snyder (2008); Harvey's (1989) structural time series models; and the seasonal ARIMA (SARIMA) models of Box, Jenkins, and Reinsel (1994). These are popular univariate time series frameworks, with models capable of generating prediction intervals. Their model specifications are flexible, and suitable for time series with trend and seasonality. We adopt fully automated and purely data dependent methods for model selection and estimation within each framework. All computational resources are readily available and fully accessible to both academics and practitioners.

The main finding of the paper is that all of the models considered provide prediction intervals with reasonably good properties in terms of coverage and

[1] In this paper, we use the terms "interval forecast" and "prediction interval" interchangeably.

width,[2] except for the AR model, which provides prediction intervals that grossly underestimate the future uncertainty. This finding is interesting, given that it is contrary to the general belief that the prediction intervals generated from time series models are too narrow. Overall, we have found that the bias-corrected bootstrap prediction intervals perform most desirably, especially when the forecast horizon is long.

The paper is organized as follows. In the next section, we provide a brief review of the literature on prediction intervals for time series models, and a discussion of the models used in our analysis. Section 3 gives details of the data and computations; and Section 4 provides the empirical results. Our conclusions are drawn in Section 5.

2. Prediction intervals

2.1. A brief literature review

Since the work of Chatfield (1993), the provision of prediction intervals has attracted particular attention in time series forecasting. Traditionally, prediction intervals have been constructed based on the assumption that forecast errors follow a normal distribution. However, as Chatfield (2001, p. 479) notes, the validity of this normal approximation is doubtful, since the assumption of normality of the forecast error distribution is often not justified in practice. In addition, it is well known that conventional prediction intervals totally ignore the sampling variability associated with parameter estimation, as De Gooijer and Hyndman (2006, p. 460) point out. Largely for these reasons, it is widely believed that prediction intervals are too narrow, under-estimating the degree of future uncertainty (see for example, Chatfield, 2001, p. 487; and Makridakis et al., 1998, p. 470).

Recently, the bootstrap method (Efron & Tibshirani, 1993) has been proposed as a means of producing a prediction interval which is robust to possible non-normality, and which also takes into account the sampling variability associated with parameter estimation. Notable examples include Thombs and Schucany (1990) for the AR model, and Pascual, Romo, and Ruiz (2004) for the ARIMA model. However, the Monte Carlo results reported in these studies reveal that the bootstrap intervals are still too narrow. Chatfield (2001, p. 487) also states that "bootstrapping does not always work", citing Meade and Islam (1995) as an example. A possible reason for this is that bootstrapping is conducted using parameter estimators which are biased in small samples. Following Clements and Taylor (2001), Kilian (1998) and Kim (2001, 2004) propose the use of the bias-corrected bootstrap for the AR model, where bias-correction is conducted in two stages of the bootstrap. They find that the bias-corrected bootstrap prediction intervals are much wider, with accurate coverage properties. In particular, the bias-corrected bootstrap is found to be highly effective when the time series possess a near unit root or the sample size is small, which are situations frequently encountered in practice. This demonstrates the importance of bias-correction for improving the performance of bootstrap prediction intervals for time series models.

Exponential smoothing is an area that has recently witnessed substantial improvements in interval forecasting. Notable earlier work on prediction intervals for Holt-Winters' method includes that of Chatfield and Yar (1991) and Yar and Chatfield (1990), while Taylor and Bunn (1999) propose an approach based on the quantile regression method of Koenker and Bassett (1978). Building on earlier work (see for example Hyndman, Koehler, Snyder, & Grose, 2002; Ord, Koehler, & Snyder, 1997; Taylor, 2003), Hyndman and Khandakar (2008) present a comprehensive statistical framework for building state space models for exponential smoothing methods, in which prediction intervals can be constructed. They consider fifteen exponential smoothing methods, and for each method they derive two innovations state space models, one with additive errors and one with multiplicative errors, resulting in a total of thirty different models. The models are estimated by maximum likelihood, and the associated prediction intervals are obtained using analytical formulae or (parametric or non-parametric) bootstrapping.[3]

[2] In this paper, a prediction interval is said to have "reasonably good" properties when it has an empirical coverage rate which is statistically no different from the nominal coverage, without being excessively wide.

[3] For forecasting high frequency data (e.g., minute-by-minute), Taylor (2008) proposes a Holt-Winters' exponential smoothing method with a double seasonal component, where prediction intervals could be constructed in the framework of the innovations state space models of Hyndman and Khandakar (2008).

2.2. Models used in this study

In this section, we provide a brief review of the time series models we adopt in this paper. As we demonstrate in the next section, time series of tourist arrivals exhibit a strong linear trend and seasonality. In particular, they often possess strong seasonal variations which include both deterministic and stochastic components (see for example Kim & Moosa, 2001, 2005). Based on this observation, we use seasonal dummy variables to capture the deterministic seasonality in each series. We also take the natural logarithms of the data before applying the time series models to each series. The technical details of all models considered are given in the Appendix.

The AR model is widely used in practice, due mainly to its simplicity. In the tourism forecasting literature, the model is referred to as the regression-based model (see for example Kim & Moosa, 2001, 2005). A long order AR model is fitted in order to capture cycles and stochastic seasonality, along with a linear trend term. The lag length of the AR model is chosen by the AIC. Point forecasts are generated recursively from the estimated model, and prediction intervals are constructed based on a normal approximation. Given its simplicity and restrictiveness, this model may be considered as a “crude benchmark”.

A natural alternative to the AR model is the SARIMA model of Box et al. (1994). It allows for both differencing and moving average components, at both seasonal and non-seasonal frequencies. We implement the fully automated procedure of Hyndman and Khandakar (2008) for selecting the models, and obtain prediction intervals based on a normal approximation, similarly to the pure AR case.

We generate the bootstrap prediction intervals using the bias-corrected bootstrap of Kim (2004), based on the AR model. Similarly to the AR case, a long order AR is fitted to capture cycles and stochastic seasonality, with a linear trend term included. As we have mentioned previously, bootstrap prediction intervals with no bias-correction have been found to be too narrow. For this reason, we do not consider bootstrap prediction intervals for the SARIMA model. Although in principle a bias-corrected bootstrap method for the SARIMA model could be developed, its properties have not been thoroughly examined in the literature to date. In addition, a bias-correction method for the parameter estimators of the SARIMA model is yet to be developed.

The basic structural time series model of Harvey (1989) decomposes an observed time series into different unobserved components. These components can be forecast individually then combined to produce a forecast for the observed series. The point forecasts are generated by running the Kalman filter for the model expressed in state space form. Prediction intervals are obtained under the assumption of normality, using the prediction error variance given by Harvey (1989, p. 222).[4]

For exponential smoothing, we use the statistical framework for innovations state space models, as presented by Hyndman and Khandakar (2008). A detailed classification of the different exponential smoothing methods and the corresponding innovations state space models is given in their Chapter 2. We consider two sets of models within this fully automated framework. For the first set, the model with the lowest AIC of all models with either additive or multiplicative errors is selected; while for the second set the model with the lowest AIC from all models with additive errors only is chosen. This is because models with additive errors are more realistic when the time series are transformed to natural logarithms. Note that the innovations state space models with additive errors have the same general form as Harvey's models. However, Harvey's formulations are more restrictive (see Hyndman & Khandakar, 2008, Chapter 13). We generate prediction intervals for all models using both the analytical formulae and the non-parametric bootstrap.

3. Data and computational details

We use the number of monthly tourist arrivals to Hong Kong, from January 1985 to May 2008 (281 observations), from four individual source markets (Australia, China, the UK, and the US) and three aggregated markets (Asia, Europe, and the total). We also use the number of monthly tourist arrivals

[4] Rodriguez and Ruiz (2009) and Stoffer and Wall (1991) proposed bootstrap prediction intervals for general state space models, based on which bootstrap prediction intervals for Harvey's model can be generated. Since their applicability to Harvey's model with a seasonal component is not fully known at this stage, we leave this method as a subject for future research.

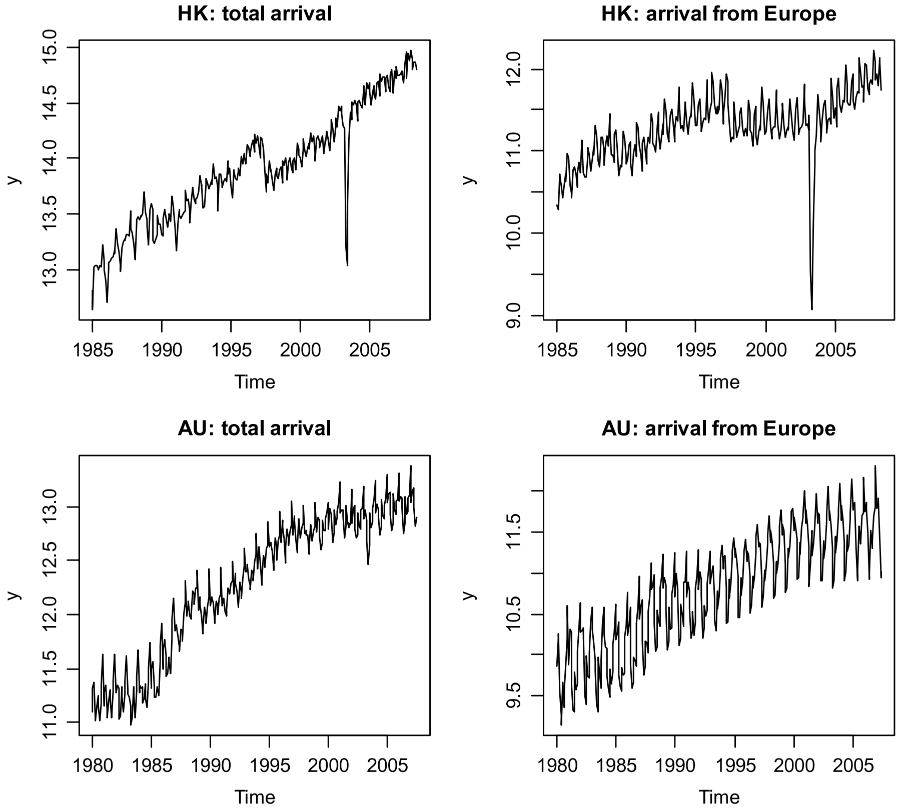

Fig. 1. Time plots of selected time series (in natural logarithms).

to Australia, from January 1980 to June 2007 (330 observations), from four individual source markets (Germany, New Zealand, the UK, and the US) and two aggregated markets (Europe and the total). These series represent different sections of the market which are of interest to academics and practitioners. The Hong Kong tourist arrivals data are obtained from the Hong Kong Tourism Board, while the Australian data are provided by Tourism Research Australia.

All time series are transformed to natural logarithms for model estimation and forecasting. Fig. 1 presents time plots for four selected time series: the total number of arrivals and the number of arrivals from Europe, to Hong Kong and Australia respectively. All of the time series show a strong upward linear trend and seasonality with mild cycles, which are both typical features of time series of tourist arrivals. The number of tourist arrivals to Hong Kong was also significantly affected by the SARS outbreak in 2003 (from April to July). The effects of this unexpected event were smoothed out using dummy variables. This is justifiable, since we are evaluating the performance of prediction intervals under normal economic conditions.[5] In Fig. 2 we report the sample autocorrelation functions of these time series in first differences, after the deterministic seasonality has been filtered out using monthly dummy variables. It is evident that, for all four cases, statistically significant stochastic seasonality is still present. Based on this, as was stated in the previous section, monthly seasonal dummy variables are used to capture the deterministic seasonality for all time series before model fitting is undertaken.

We evaluate the performance of alternative prediction intervals in a purely empirical setting. We use a rolling window of 120 observations (10 years) for

[5] We conducted the forecasting exercise without smoothing out the impact of this unexpected event, at an early stage of this study. As expected, the accuracy of the forecasts deteriorates. When the market is subject to extreme panic and instability, judgemental forecasts may be preferable to time series forecasts.

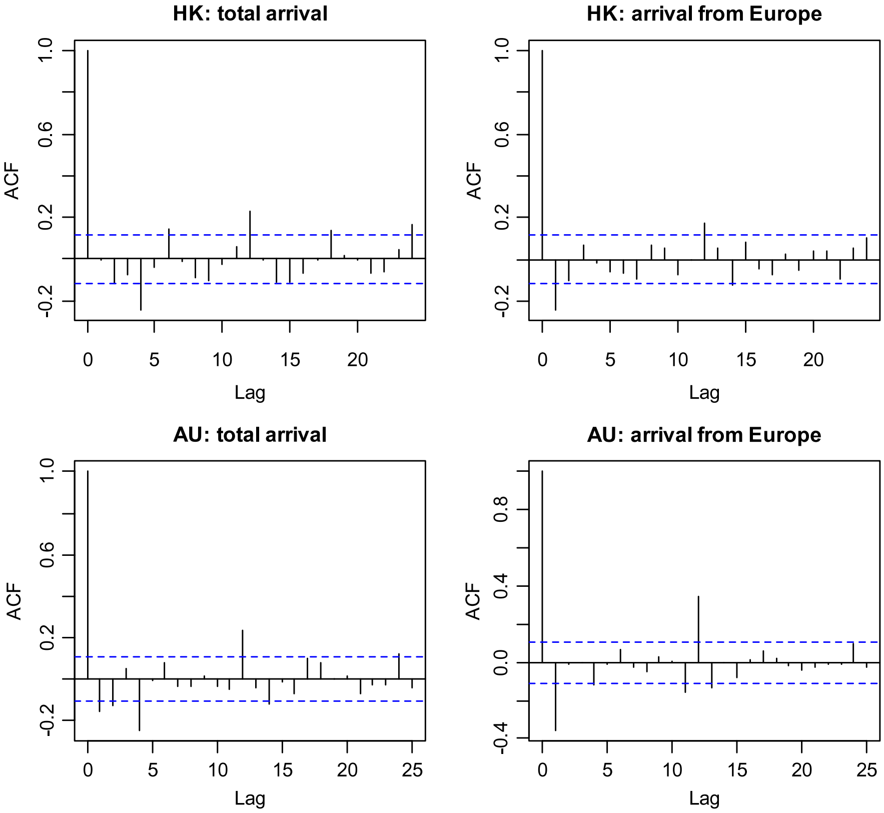

Fig. 2. Sample autocorrelation functions of selected time series: the first differences of the time series, filtered with seasonal dummy variables.

estimation, and generate 1- to 12-step-ahead out-of-sample prediction intervals for each window. That is, we take the first 120 observations for estimation, and generate prediction intervals for each of the next 12 months. We then take the next 120 observations (from 2 to 121) and generate prediction intervals. For each sub-sample, we update the model and parameter estimates, going through the automatic model selection and estimation procedures. This process continues until we reach the end of the data set. In total, we have obtained 150 and 199 prediction intervals respectively for Hong Kong and Australia, for each forecast horizon of 1 to 12. The use of 120 sample observations is based on the consideration that the sample size is large enough to justify the asymptotic theories involved in the model selection and estimation. On the other hand, this is also a moderate sample length that has commonly been adopted in empirical applications in tourism forecasting. We use a forecast horizon of 12 months because the models adopted in this paper are mainly for short-term forecasting. This exercise may be likened to a situation where a forecaster is generating prediction intervals over a period of more than twenty years, using time series data from the past 10 years and adopting an automatic forecasting method, with a forecast horizon of 12 months. All computations are conducted using the programming language R (R Development Core Team, 2008), which is a free and open-source language, and making use of the R packages *BootPR* (Kim, 2008) and *forecast* (Hyndman, 2008). The R code used in this study can be provided on request.

We generate prediction intervals of the nominal coverage rate of 95%. This means that in repeated sampling, a prediction interval is expected to cover the true future value with a probability of 0.95. We calculate the mean coverage rate for the forecast horizon h as

$$C(h) = \frac{\#\{L_h \le Y_h \le U_h\}}{N},$$

where Y_h is the true future value, L_h and U_h are the lower and upper bounds of a prediction interval respectively, N is the total number of prediction intervals for forecast horizon h, and # indicates the frequency at which the condition inside the bracket is satisfied. If a prediction interval provides an accurate assessment of future uncertainty, the value of $C(h)$ should be close to 0.95. To test whether $C(h)$ is statistically different to the nominal coverage of 0.95, we use the 95% confidence interval based on a normal approximation to a binomial distribution.[6] That is,

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{N}}, p + 1.96\sqrt{\frac{p(1-p)}{N}}\right],$$

where $p = 0.95$. If the calculated value of $C(h)$ belongs to this interval, we cannot reject the hypothesis that the true coverage is equal to the nominal coverage 0.95, at the 5% significance level. In addition to $C(h)$, we also use the mean width for each horizon h, defined as the mean value of $(U_h - L_h)$ over N prediction intervals. A higher value of the mean width indicates that there is more uncertainty associated with the forecasting, and thus the prediction intervals are less informative. We prefer a forecasting model that generates prediction intervals whose $C(h)$ belongs to the above confidence interval. If two or more models generate such prediction intervals, we prefer the one with the smaller value of the mean width.

4. Empirical results

In this section, we compare the point and interval forecasts generated from the alternative models. These include forecasts from: (i) the AR model (AR); (ii) the AR model using the bias-corrected bootstrap (BOOT)[7]; (iii) the SARIMA model; (iv) Harvey's structural model (ST); and the two sets of innovations state space models for exponential smoothing, (v) ETS1, where the model is selected from all models with either multiplicative or additive errors, and (vi) ETS2, where the model is selected from those with additive errors only.

4.1. Point forecasting

Although the focus of this paper is on interval forecasting, it is also instructive to compare the accuracy of point forecasts. We calculate the MSFE (mean squared forecast error) values of the point forecasts from each model in the same way as we did for interval forecasting. That is, we use the rolling window of 120 observations and generate 12-step-ahead point forecasts. For all models, the point forecasts are found to be fairly accurate, with small MSFE values. To evaluate the statistical significance of the differences in MSFE values, we have conducted the Diebold and Mariano (1995) test, with the BOOT forecasts as the benchmark. The null hypothesis is MSFE(BOOT) = MSFE(i), tested against the two-tailed alternative, where $i \in \{\text{AR, SARIMA, ST, ETS1, ETS2}\}$. Note that MSFE(BOOT) is used as the benchmark, and a negative Diebold-Mariano statistic indicates that BOOT has a lower MSFE value than its competitor. Table 1 reports the outcomes of the test for each forecast horizon. For each cell, the entry represents (b, a, c), where a is the number of cases of not rejecting the null; b the number of cases of rejecting with a negative statistic; and c the number of cases of rejecting with a positive statistic; all at the 5% significance level. As an example, for the AR model and for $h = 2$, the entry (2, 11, 0) indicates that the null hypothesis that the MSFE of the BOOT forecasts is equal to that of the AR forecasts is not rejected for eleven time series; however, for two of the time series, the MSFE of BOOT is found to be statistically lower than that of AR. For the other cases, it is evident that the null hypothesis fails to be rejected for nearly all time series and forecast horizons, except for the AR model. Hence, we conclude that, overall, all models except for the AR provide point forecasts of an equal level of accuracy to the BOOT forecasts.

4.2. Interval forecasting

Figs. 3 and 4 plot the values of the mean coverage rates and widths of alternative prediction intervals for tourist arrivals to Hong Kong and Australia, respectively. For simplicity of exposition, we have chosen

[6] The validity of this approximation depends on the independence of successive trials. Although the details are not reported, we have applied Christoffersen's (1998) independence test for the prediction intervals for $h = 1$, and find that the independence is satisfied overall for nearly all cases. However, we should note that the independence assumption will not usually be satisfied for $h > 1$.

[7] See Kim (2003) for details of bias-corrected point forecasting based on the bootstrap.

Table 1
Results of the Diebold-Mariano test for the equality of the MSFEs of point forecasts.

h	AR	SARIMA	ST	ETS1	ETS2
1	(0, 13, 0)	(1, 11, 1)	(0, 13, 0)	(0, 12, 1)	(0, 13, 0)
2	(2, 11, 0)	(1, 10, 2)	(0, 12, 1)	(0, 12, 1)	(1, 12, 0)
3	(5, 8, 0)	(1, 10, 2)	(0, 12, 1)	(1, 11, 1)	(1, 11, 1)
4	(8, 5, 0)	(1, 10, 2)	(1, 11, 1)	(2, 10, 1)	(3, 9, 1)
5	(8, 5, 0)	(1, 11, 1)	(1, 11, 1)	(2, 10, 1)	(2, 10, 1)
6	(8, 5, 0)	(1, 12, 0)	(1, 10, 2)	(2, 10, 1)	(3, 9, 1)
7	(8, 5, 0)	(1, 12, 0)	(1, 11, 1)	(2, 10, 1)	(2, 10, 1)
8	(7, 6, 0)	(2, 11, 0)	(1, 11, 1)	(2, 10, 1)	(2, 10, 1)
9	(7, 6, 0)	(1, 12, 0)	(1, 11, 1)	(1, 11, 1)	(2, 10, 1)
10	(5, 8, 0)	(0, 12, 1)	(1, 10, 2)	(1, 10, 2)	(1, 11, 1)
11	(4, 8, 1)	(1, 11, 1)	(1, 10, 2)	(2, 9, 2)	(0, 11, 2)
12	(4, 7, 2)	(1, 11, 1)	(1, 10, 2)	(1, 9, 3)	(1, 10, 2)

The null hypothesis, MSFE(BOOT) $=$ MSFE(i), is tested against the two-tailed alternative, where $i \in$ {AR, SARIMA, ST, ETS1, ETS2}, while MSE(BOOT) is used as a benchmark.
AR: autoregressive model.
BOOT: forecasting with bootstrap bias-corrected AR parameters.
ST: Harvey's structural time series models.
SARIMA: seasonal ARIMA models.
ETS1: state space exponential smoothing models, the best model is chosen from among both multiplicative and additive models.
ETS2: state space exponential smoothing models, the best model is chosen from additive models only.
For each cell, the entry represents (b, a, c), where a is the number of cases of not rejecting the null; b the number of cases of rejecting with a negative statistic; and c the number of cases of rejecting with a positive statistic; all at the 5% significance level. Note that $a + b + c = 13$ represents the total number of time series considered.

five representative source markets for the two destinations.[8] For both figures, the graphs in the first column show the mean coverage rates from different models over the forecast horizon h. The solid horizontal lines represent the 95% confidence bands around the nominal coverage rate of 0.95. For ETS1, we report the properties of prediction intervals generated using the non-parametric bootstrap, while for ETS2 we only report the results based on the analytical formulae. This is only for the sake of simplicity, as the results for the other cases are qualitatively similar. Unless otherwise stated, ETS refers to both ETS1 and ETS2 for the rest of the paper.

4.2.1. Tourist arrivals to Hong Kong

For tourism arrivals from Australia, only the ETS and ST models have mean coverage rates within the 95% confidence band for all values of h. The BOOT intervals have mean coverage rates within the band in most cases, although they under-cover when $h \geq 10$. The SARIMA prediction intervals also tend to under-cover the true values, while the AR intervals grossly under-cover the true future values for all values of h. These features are also evident from the mean width properties, where the AR and SARIMA prediction intervals are much narrower than the others. The prediction intervals from the ETS and ST models become much wider than the BOOT prediction intervals for $h > 5$. Overall, the ETS and ST models provide accurate prediction intervals, but the BOOT prediction intervals also perform well for $h \leq 9$.

For arrivals from China, all prediction intervals have mean coverage rate values within the 95% confidence band, except those from the AR, which seriously under-cover the true values for $h \geq 5$. This is reflected in the prediction intervals from the AR model being too narrow. The BOOT and SARIMA prediction intervals are much tighter than the ETS and ST prediction intervals for $h \geq 6$. Hence, in the case of China, the BOOT and SARIMA prediction intervals should be preferred to the others. For the UK, only

[8] The complete set of figures and related discussions for all source markets are given in an earlier version of the paper, available at http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2008/wp11-08.pdf. Note that the results of the unreported source markets are qualitatively similar to those reported.

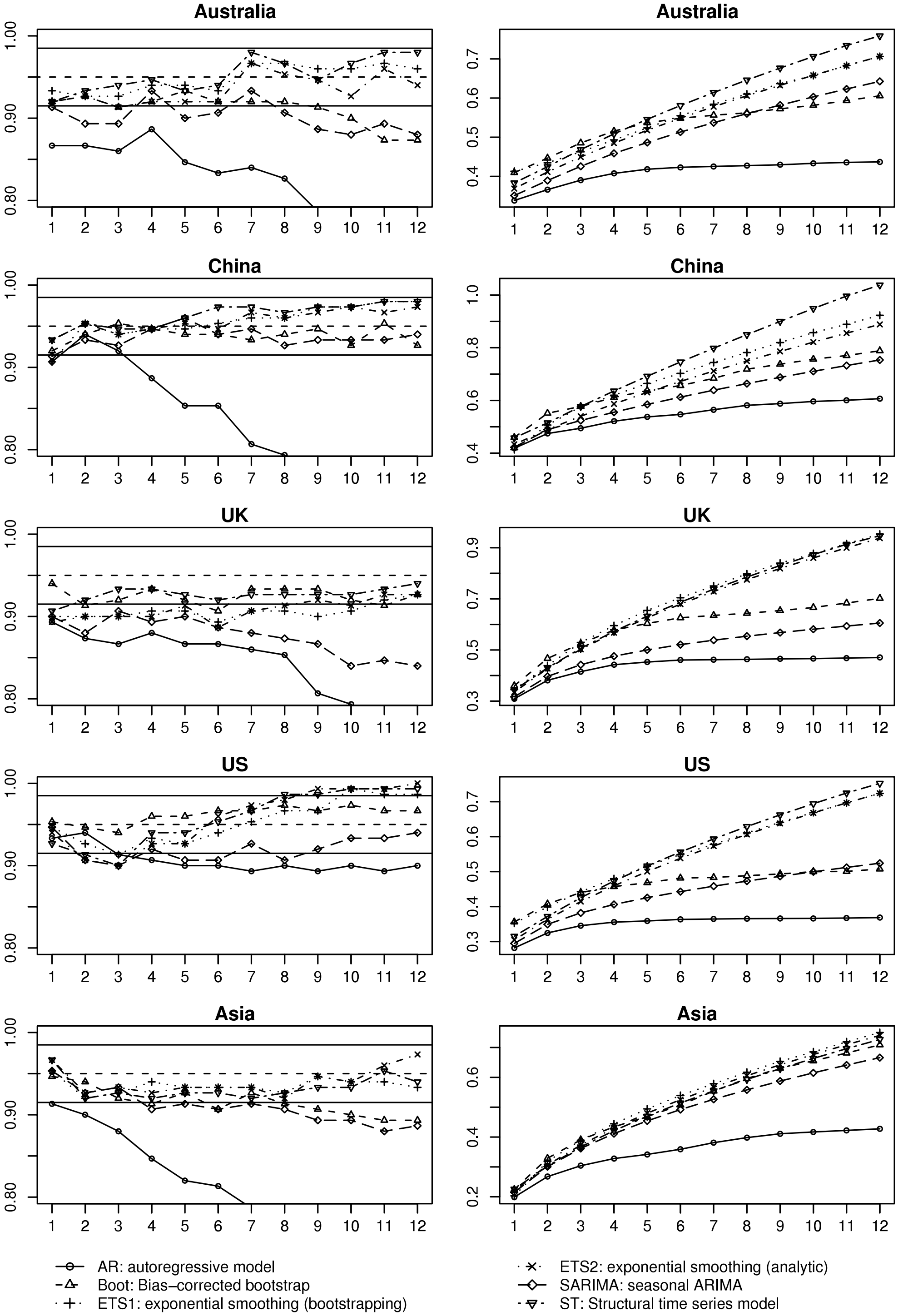

Fig. 3. Mean coverage rates and widths of alternative prediction intervals: tourist arrivals to Hong Kong. The first column compares the mean coverage rates of prediction intervals, while the second compares the mean width of the prediction intervals, over forecast horizons 1–12. The solid horizontal lines in the coverage rate graphs indicate 95% confidence intervals around the nominal coverage rate of 0.95.

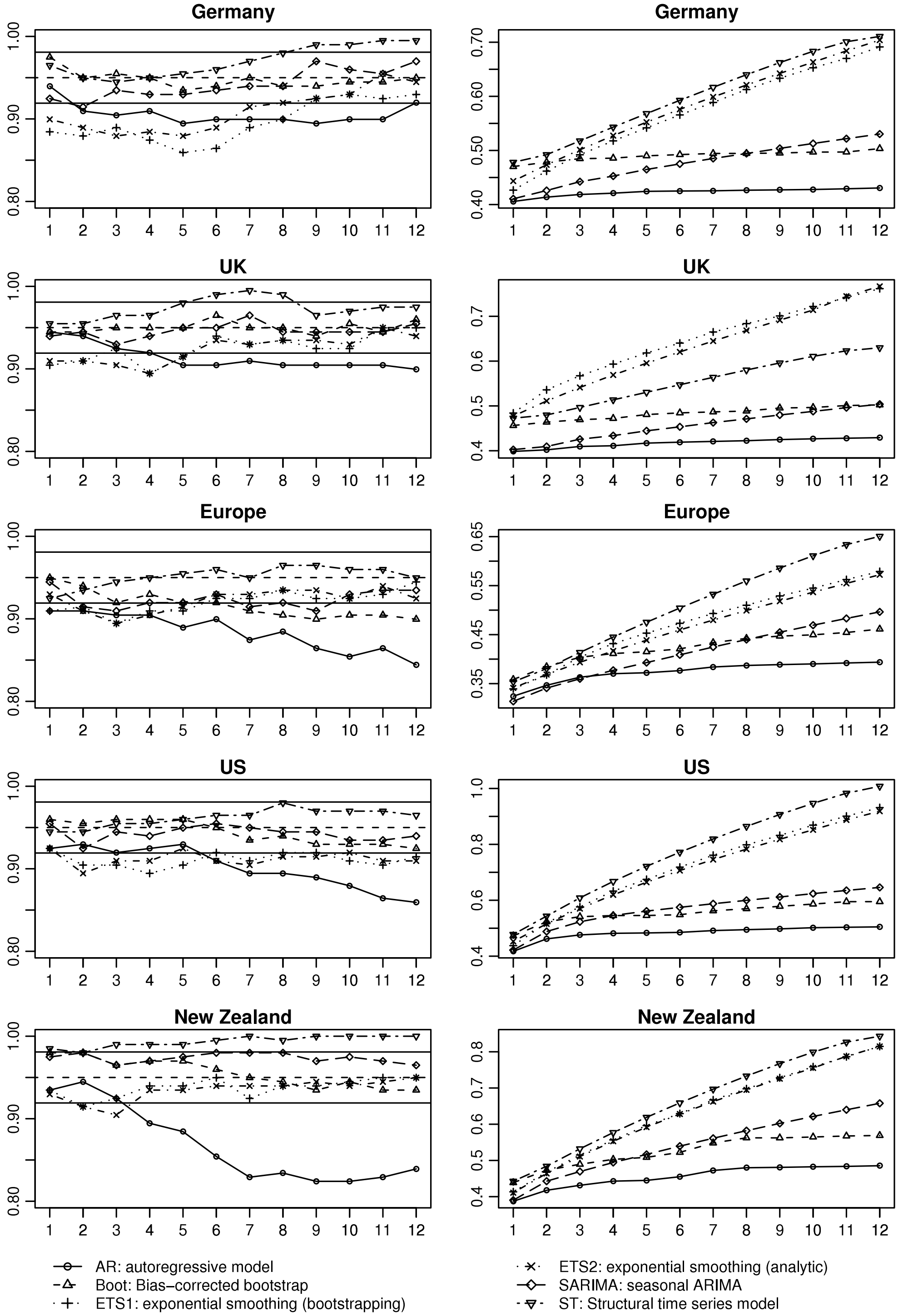

Fig. 4. Mean coverage rates and widths of alternative prediction intervals: tourist arrivals to Australia. The first column compares the mean coverage rates of prediction intervals, while the second compares the mean width of the prediction intervals, over forecast horizons 1–12. The solid horizontal lines in the coverage rate graphs indicate 95% confidence intervals around the nominal coverage rate of 0.95.

the BOOT and ST prediction intervals have all values of the mean coverage rate within the 95% confidence band, while the others under-cover the true values. The AR and SARIMA prediction intervals grossly under-cover the true values, while the ETS prediction intervals, though still under-covering, perform much better than these two. Looking at the widths of the intervals, the BOOT prediction intervals are much tighter than the ETS and ST prediction intervals for $h \geq 6$. Hence, the BOOT prediction intervals should be preferred in the case of the UK.

In the case of the US, the BOOT prediction intervals have the best coverage properties: all coverage rates are within the 95% confidence band and are very close to the nominal coverage rate 0.95. However, the other prediction intervals also show good coverage properties. The SARIMA prediction intervals have all mean coverage rates inside the 95% confidence band for all forecast horizons, while the ST and ETS intervals do for most of the forecast horizons. The only exception to this good performance is the AR model, which again provides prediction intervals that under-cover the true values in many cases. Looking at the width properties, the BOOT and SARIMA prediction intervals are again tighter than the ETS and ST intervals. Hence, similarly to the case of China, the BOOT and SARIMA prediction intervals should be preferred. For Asia, only the ETS and ST prediction intervals have desirable coverage rates for all values of h, while the BOOT and SARIMA prediction intervals under-cover the true values for $h \geq 9$. The AR prediction intervals substantially under-cover the true values for nearly all forecast horizons. All of the prediction intervals (except for AR) have similar mean width values. Hence, in this case, the ETS and ST prediction intervals perform most desirably, but the BOOT and SARIMA prediction intervals also perform reasonably well.

4.2.2. Tourist arrivals to Australia

Fig. 4 presents the mean coverage rates and widths of alternative prediction intervals for tourist arrivals to Australia. As in the case of Hong Kong, the AR models provide prediction intervals which are much inferior to the others. Therefore, for the sake of simplicity, the AR prediction intervals will not be discussed any further.

For Germany, only the BOOT prediction intervals have all of their mean coverage values inside the 95% confidence intervals. The SARIMA intervals also perform well, with the value of the mean coverage being outside the 95% confidence band only when $h = 2$. The BOOT and SARIMA intervals are also very tight, with their mean width values being much smaller than those of the ETS and ST intervals for nearly all values of h. The latter become increasingly wide as the forecast horizon increases. For the UK, the BOOT and SARIMA intervals have all of their mean coverage values inside the 95% confidence band, while the ETS and ST intervals either under- or over-cover the true values for a few forecast horizons. The former are much tighter than the latter, again especially for longer forecast horizons. For Europe, the ST is the only model for which all of the prediction intervals are inside the 95% confidence band. The others tend to under-cover the true future values, the ETS intervals for shorter forecast horizons and the BOOT and SARIMA intervals for longer horizons. Looking at the widths of the prediction intervals, the ST intervals are much wider, suggesting that the other intervals underestimate the future uncertainty in this case. For the US, the BOOT, SARIMA and ST prediction intervals have mean coverage values inside the 95% confidence band for all values of h. The BOOT prediction intervals are the tightest for nearly all forecast horizons.

4.2.3. Discussion

All of the models considered generate prediction intervals with reasonable performances, except for the AR model, which often grossly underestimates the future uncertainty. This is particularly the case for short forecast horizons ($h \leq 4$), where, in most cases, all models provide prediction intervals with correct coverage rates and similar width properties. Overall, we have found a strong tendency for the BOOT method to outperform its competitors. In general, the BOOT prediction intervals have the most desirable coverage properties, providing an informative assessment of the future uncertainty. The SARIMA, ETS and ST prediction intervals also perform reasonably well; however, the SARIMA intervals can sometimes be too narrow, and hence underestimate the future uncertainty, while the ETS and ST intervals tend to become wider, relative to the BOOT intervals, as the forecast horizon increases. In fact, for shorter forecast horizons ($h \leq 4$), the ETS and ST models often provide tighter prediction intervals than the BOOT method. In general, the BOOT prediction intervals are slightly wider than the others for shorter horizons, but become much tighter as the forecast horizon increases.

As was stated in Section 2, it is widely accepted that the prediction intervals generated from time series models are too narrow. This is because the conventional intervals (1) assume normality, which may not hold; (2) ignore the sampling variability associated with parameter estimation; and (3) do not make adjustments for the small sample biases of parameter estimators, where applicable. In this paper, it is found that, consistent with the general belief, the conventional prediction intervals from the AR model are far too narrow. However, the other prediction intervals (those from the bias-corrected bootstrap version of the AR, the SARIMA model, Harvey's structural time series model, and the state space models for exponential smoothing) are found to be much wider, producing satisfactory coverage probabilities.

From the methodological aspect, we have adopted two innovative approaches. One is the use of a rolling horizon methodology for evaluating the performances of prediction intervals over a range of samples; while the other is the use of automatic model selection, which removes the subjectivity from the model selection procedures. Although these represent sound approaches, it should be noted that the findings of this paper are limited to the current data set. We call for more extensive empirical research efforts of this kind in the future, in order to further explore the validity of interval forecasting for tourism forecasting.

5. Concluding remarks

Time series forecasting for tourist arrivals has been an area of extensive empirical research. While a large number of studies published over the years have reported that the application of time series methods has been a great success, their major concern has been the issue of point forecasting. Although interval forecasts can be invaluable for decision-makers in both industry and government, there have been no studies that have looked at the accuracy of prediction intervals in the context of forecasting the number of tourist arrivals.

The purpose of this paper is to evaluate the performance of the prediction intervals generated using alternative time series models. We consider univariate time series models such as the AR model, the bias-corrected bootstrap for the AR model (Kim, 2004), the innovations state space models for exponential smoothing (as presented by Hyndman & Khandakar, 2008), Harvey's (1989) structural time series models, and the seasonal ARIMA model of Box et al. (1994). We employ an automatic forecasting approach, in which the prediction intervals are generated from forecasting models whose specifications are determined automatically using a fully data-dependent procedure. The performances of the prediction intervals are evaluated in a purely empirical setting, calculating the coverage rate and width of the intervals using the rolling window method. For this purpose, we use thirteen monthly time series for the number of tourist arrivals to Hong Kong and Australia.

The main finding of the paper is that, in general, the bias-corrected bootstrap prediction intervals perform most desirably, providing tight intervals with accurate coverage values. The prediction intervals from the exponential smoothing and structural time series models also show satisfactory probability coverage properties, although they tend to become wider at longer forecast horizons relative to the bias-corrected bootstrap intervals. The seasonal ARIMA prediction intervals tend to under-estimate the future uncertainty. The AR model also performs rather poorly, grossly underestimating the future uncertainty. However, this paper demonstrates that most of the popular time series models adopted in this study generate prediction intervals with desirable statistical properties, at least in the context of tourism forecasting.

Acknowledgements

We thank Andrew Maurer and Claude Calero at Tourism Research Australia for providing the Australian data, explanations and comments. Athanasopoulos and Song acknowledge financial support from Tourism Research Australia, the Sustainable Tourism Cooperative Research Centre, Australian Research Council grant DR0984399; and the Hong Kong Polytechnic University (Grant No. G-YF35). We would like to thank Rob Hyndman (editor) and two referees for constructive comments.

Appendix

As was stated in Section 2.2, all time series are filtered using seasonal dummy variables before these models are applied. This is because the time series of tourist arrivals possess both stochastic and deterministic seasonality, as we have seen in Section 3.

A.1. Autoregressive model and bias-corrected bootstrap

We consider an AR(p) model of the form

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \beta t + u_t, \qquad (1)$$

where u_t is *i.i.d.* with zero mean and fixed variance. The AR part of Eq. (1) is stationary, with all of its characteristic roots outside the unit circle. The unknown AR order is chosen by the AIC, with the maximum order being 18. Given the observed time series $\{Y_t\}_{t=1}^{n}$, the unknown parameters are estimated using the least squares (LS) method. The LS estimator for $\alpha = (\alpha_0, \ldots, \alpha_p)$ is denoted by $\hat{\alpha} = (\hat{\alpha}_0, \ldots, \hat{\alpha}_p)$, and the associated residuals by $\{e_t\}_{t=p+1}^{n}$. The point forecasts for Y_{n+h} made at n can be generated in the usual way, conditional on the last p observations of Y, using $\hat{\alpha}$. The $100(1-\theta)\%$ prediction intervals for Y_{n+h} can be constructed based on a normal approximation.

Only a sketchy description of the bias-corrected bootstrap procedure of Kim (2004) is given here. Let the bias of $\hat{\alpha}$ be denoted as $Bias\,(\hat{\alpha})$. This bias can be estimated using the analytical formula of Shaman and Stine (1988), and the bias-corrected estimator can be obtained as

$$\hat{\alpha}^c \equiv (\hat{\alpha}_0^c, \hat{\alpha}_1^c, \ldots, \hat{\alpha}_P^c) = \hat{\alpha} - Bias(\hat{\alpha}). \qquad (2)$$

The residuals associated with $\hat{\alpha}^c$ are denoted by $\left\{e_t^c\right\}_{t=p+1}^{n}$. Generate an artificial data set recursively using the backward AR form, as

$$Y_t^* = \hat{\alpha}_0^c + \hat{\alpha}_1^c Y_{t+1}^* + \cdots + \hat{\alpha}_p^c Y_{t+p}^* + v_t^*, \qquad (3)$$

where the p starting values are set equal to the last p values of the original series, and v_t^* is a random draw from $\left\{e_t^c\right\}_{t=p+1}^{n}$ with replacement. Using the artificial data set $\left\{Y_t^*\right\}_{t=1}^{n}$, the parameters of the forward model (1) are estimated, and the LS estimators are denoted by $\hat{\alpha}^*$. Obtain the bias-corrected estimator $\hat{\alpha}^{*c} = \hat{\alpha}^* - Bias(\hat{\alpha}^*)$, again using the Shaman-Stine bias formula, as in Eq. (2). The bootstrap replicates of the AR forecast for Y_{n+h} made at time period n are generated recursively as

$$\begin{aligned} Y_n^*(h) = \hat{\alpha}_0^{*c} &+ \hat{\alpha}_1^{*c} Y_n^*(h-1) + \cdots \\ &+ \hat{\alpha}_p^{*c} Y_n^*(h-p) + u_{n+h}^*, \end{aligned} \qquad (4)$$

where $Y_n^*(j) = Y_{n+j}^* = Y_{n+j}$ for $j \le 0$, and u_{n+h}^* is a random draw from $\left\{e_t^c\right\}_{t=p+1}^{n}$ with replacement. Repeat Eqs. (3) and (4) many times, say B, to yield the bootstrap distribution for the AR forecast $\left\{Y_n^*(h;i)\right\}_{i=1}^{B}$. The $100(1-\theta)\%$ prediction intervals for Y_{n+h}, based on the percentile method (Efron & Tibshirani, 1993), are calculated as $[Y_n^*(h,\tau), Y_n^*(h,1-\tau)]$, where $Y_n^*(h,\tau)$ is the 100τth percentile of the bootstrap distribution $\left\{Y_n^*(h;i)\right\}_{i=1}^{B}$, and $\tau = 0.5\theta$.

A.2. Seasonal ARIMA models

The general form of the seasonal ARIMA model (Box et al., 1994) with periodicity s ($s = 12$ for monthly data) can be written as

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta_s^D Y_t = \mu + \theta_q(B)\Theta_Q(B^s)u_t, \qquad (5)$$

where u_t is a white noise process with fixed variance and μ is a constant. B is the backward shift operator, and $\Delta^d \equiv (1-B)^d$ and $\Delta_s^D = (1-B^s)^D$ are the operators for the dth order monthly difference and the Dth order annual difference, respectively. $\phi_p(B)$, $\theta_q(B)$, $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are the polynomials in B and B^s, which can be written as

$$\begin{aligned} &\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \\ &\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q, \\ &\Phi_P(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps}, \quad \text{and} \\ &\Theta_Q(B^s) = 1 - \Theta_1 B^s - \cdots - \Theta_Q B^{Qs}, \end{aligned}$$

where the ϕs, θs, Φs and Θs are parameters to be estimated. The model parameters are estimated using the maximum likelihood method. The point forecast $Y_n(h)$ for Y_{n+h} is generated recursively using the estimated coefficients, conditional on the observed time series. The prediction intervals can be constructed in the usual way under the assumption of normality. For model selection, we follow the automatic procedure described by Hyndman and Khandakar (2008), where the numbers of differencing and seasonal differencing

d and D are determined using the Canova and Hansen (1995) seasonal unit root test, and the orders p, q, P_s, and Q_s of model (5) are selected using the AIC, following the step-wise procedure for traversing the model space.

A.3. Harvey's basic structural model

The basic structural time series model of Harvey (1989) decomposes an observed time series into different unobserved components. These components can be forecast individually and combined to produce a forecast for the observed series. The model may be written as

$$Y_t = \mu_t + \gamma_t + \varepsilon_t,$$

where Y_t is the observed time series, μ_t is the trend component, γ_t is the seasonal component and ε_t is the random component. The trend and seasonal components are assumed to be uncorrelated, while ε_t is assumed to be white noise.

The trend, which represents the long-term movement in a series, can be represented by

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t,$$
$$\beta_t = \beta_{t-1} + \zeta_t,$$

where $\eta_t \sim NID(0, \sigma_\eta^2)$ and $\zeta_t \sim NID(0, \sigma_\zeta^2)$. The seasonal component is specified as

$$\gamma_t = -\sum_{j=1}^{s-1} \gamma_{t-j} + w_t,$$

where $w_t \sim NID(0, \sigma_w^2)$. The seasonality is deterministic when $\sigma_w^2 = 0$.

The point forecasts are generated by running the Kalman filter for the above structural model expressed in state space form. The prediction intervals are obtained under the assumption of normality, using the prediction error variance given by Harvey (1989, p. 222).

A.4. State space models for exponential smoothing

The general form of these models can be written as

$$Y_t = w(X_{t-1}) + r(X_{t-1})\varepsilon_t;$$
$$X_t = f(X_{t-1}) + g(X_{t-1})\varepsilon_t,$$

where ε_t is a Gaussian white noise with zero mean and fixed variance, and

$$X_t = (l_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})'$$

is a state vector, while l_t, b_t, and s_t denote the level, slope, and seasonal components at time t, respectively, and m is the length of seasonality. According to Hyndman and Khandakar (2008), the updating formulae of all exponential smoothing methods are special cases of the above general model.

The innovations state space models are estimated using the maximum likelihood method, and the model selection is done automatically using the AIC. Further details of the point forecasting, interval forecasting, and automatic model selection procedures are given by Hyndman and Khandakar (2008).

References

Box, G. E. P., Jenkins, G. M., & Reinsel, C. (1994). *Time series analysis: forecasting and control*. Englewood Cliffs: Prentice Hall.

Canova, F., & Hansen, B. (1995). Are seasonal patterns constant over time? A test for seasonal stability. *Journal of Business and Economic Statistics*, *13*, 237–252.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, *11*(2), 121–135.

Chatfield, C. (2001). Prediction intervals for time-series forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researcher and practitioners* (pp. 475–494). Boston: Kruger Academic Publishers.

Chatfield, C., & Yar, M. (1991). Prediction intervals for multiplicative Holt-Winters. *International Journal of Forecasting*, *7*, 31–37.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*, 841–862.

Clements, M. P., & Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting*, *17*, 247–267.

De Gooijer, J., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*, 443–473.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman Hall.

Frechtling, D. C. (2001). *Forecasting tourism demand: methods and strategies*. Oxford: Butterworth-Heinemann.

Harvey, A. C. (1989). *Forecasting structural time series models and the Kalman Filter*. Cambridge: Cambridge University Press.

Hyndman, R. J. (2008). *Forecast: forecasting functions for time series*. R package version 1.14 http://www.robjhyndman.com/software/forecast/.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3).

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Berlin: Springer-Verlag.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454.

Kilian, L. (1998). Small sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, *80*, 218–230.

Kim, J. H. (2001). Bootstrap-after-bootstrap prediction intervals for autoregressive models. *Journal of Business and Economic Statistics*, *19*, 117–128.

Kim, J. H. (2003). Forecasting autoregressive time series with bias-corrected parameter estimators. *International Journal of Forecasting*, *19*, 493–502.

Kim, J. H. (2004). Bootstrap prediction intervals for autoregression using asymptotically mean-unbiased estimators. *International Journal of Forecasting*, *20*, 85–97.

Kim, J. H. (2008). *BootPR: bootstrap prediction intervals and bias-corrected forecasting*. R package version 0.56.

Kim, J. H., & Moosa, I. A. (2001). Seasonal behaviour of monthly international tourist flows: specification and implications for forecasting models. *Tourism Economics*, *7*, 381–396.

Kim, J. H., & Moosa, I. A. (2005). Forecasting international tourist flows to Australia: a comparison between the direct and indirect methods. *Tourism Management*, *26*, 69–78.

Koenker, R. W., & Bassett, G. W. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.

Li, G., Song, H., & Witt, S. F. (2005). Recent developments in econometric modelling and forecasting. *Journal of Travel Research*, *44*, 82–99.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: methods and applications*. New York: John Wiley Sons.

Meade, N., & Islam, T. (1995). Prediction intervals for growth curve forecasts. *Journal of Forecasting*, *14*, 413–430.

Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, *92*, 1621–1629.

Pascual, L., Romo, J., & Ruiz, E. (2004). Bootstrap predictive inference for ARMA process. *Journal of Time Series Analysis*, *25*, 449–465.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Rodriguez, A., & Ruiz, E. (2009). Bootstrap prediction intervals in state-space models. *Journal of Time Series Analysis*, *30*, 167–178.

Shaman, P., & Stine, R. A. (1988). The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association*, *83*, 842–848.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—a review of recent research. *Tourism Management*, *29*, 203–220.

Song, H., Witt, S. F., & Zhang, X. (2008). Developing a web-based tourism demand forecasting system. *Tourism Economics*, *14*(3), 445–468.

Stoffer, D. S., & Wall, K. D. (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. *Journal of the American Statistical Association*, *86*, 1024–1033.

Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, *19*, 273–289.

Taylor, J. W. (2008). An evaluation of methods for very short term electricity demand forecasting using minute-by-minute British data. *International Journal of Forecasting*, *24*, 645–658.

Taylor, J. W., & Bunn, D. W. (1999). A quantile regression approach to generating prediction intervals. *Management Science*, *45*(2), 225–237.

Thombs, L. A., & Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, *85*, 486–492.

Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, *11*, 447–475.

Yar, M., & Chatfield, C. (1990). Prediction intervals for the Holt-Winters forecasting procedure. *International Journal of Forecasting*, *6*, 127–137.