ISSN 1440-771X

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

The tourism forecasting competition

George Athanasopoulos, Rob J Hyndman,

Haiyan Song, Doris C Wu

March 2010

Working Paper 10/08

Revised

The tourism forecasting competition

George Athanasopoulos
Department of Econometrics and Business Statistics
and Tourism Research Unit,
Monash University, VIC 3800, Australia.
Email: George.Athanasopoulos@buseco.monash.edu.au

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800, Australia.
Email: Rob.Hyndman@buseco. monash.edu.au

Haiyan Song
The School of Hotel and Tourism Management,
The Hong Kong Polytechnic University, Hong Kong.
Email: Haiyan.Song@inet.polyu.edu.hk

Doris C Wu
The School of Hotel and Tourism Management,
The Hong Kong Polytechnic University, Hong Kong.
Email: Doris.Wu@inet.polyu.edu.hk

*****Corresponding author

23 March 2010

JEL classification: C22,C52,C53

The tourism forecasting competition

Abstract

We evaluate the performance of various methods for forecasting tourism data. The data used include 366 monthly series, 427 quarterly series and 518 annual series, all supplied to us by tourism bodies or by academics who had used them in previous tourism forecasting studies. The forecasting methods implemented in the competition are univariate and multivariate time series approaches, and econometric models. This forecasting competition differs from previous competitions in several ways: (i) we concentrate only on tourism data; (ii) we include approaches with explanatory variables; (iii) we evaluate the forecast interval coverage as well as point forecast accuracy; (iv) we observe the effect of temporal aggregation on forecasting accuracy; and (v) we consider the mean absolute scaled error as an alternative forecasting accuracy measure. We find that pure time series approaches provide more accurate forecasts for tourism data than models with explanatory variables. For seasonal data we implement three fully automated pure time series algorithms that generate accurate point forecasts and two of these also produce forecast coverage probabilities which are satisfactorily close to the nominal rates. For annual data we find that Naïve forecasts are hard to beat.

Keywords: ARIMA, exponential smoothing, state space model, time varying parameter model, dynamic regression, autoregressive distributed lag model, vector autoregression.

1 Introduction

Over the past four decades, tourism has developed into one of the most rapidly growing global industries. The World Tourism Organization (2008) reports that international tourist arrivals worldwide grew at a rate of 6% in 2007, reaching nearly 900 million, compared to 800 million two years earlier. Both academic interest and the tourism literature have grown, parallel to this growth in the industry, producing many articles that model and forecast tourism flows between various countries. These articles vary immensely in scope, modelling and forecasting techniques, and data types, lengths and frequencies. The three major literature review articles that attempt to summarise these are Witt and Witt (1995), Li et al. (2005) and Song and Li (2008). Despite the authors' best efforts, the diversity of the studies has not led to a consensus about the relative forecasting performance of commonly used methods when they are applied to tourism data. In this paper we apply forecasting methods to a very broad collection of series within the field of tourism. This allows us to draw conclusions for methods within this field, and also to contribute some general observations and conclusions relevant to the broader field of forecasting.

Since the last of the M series of forecasting competitions was published (the M3 competition, see Makridakis and Hibon, 2000), there have been no more major contributions to this line of research. In this paper we take on the challenge of creating a forecasting competition that overcomes some of the limitations highlighted by the commentators of the M3 competition (see Ord, 2001). For example, in the M3 competition, 3003 series from various business and economic sectors were used. This was larger than the number of series used in any previous study. However, the question of whether "bigger means better" was raised by several commentators. Suggestions were made calling for mini-competitions along the lines suggested by Fildes and Ord (2004), who proposed the use of a more homogeneous set of series. In an effort to use a homogeneous set of series which are representative of some population, we use 366 monthly, 427 quarterly and 518 yearly series, all from the field of tourism. We present the data in Section 4. From such a set of series we intend to draw some general inferences on the modelling and forecasting of tourism demand. We will also examine whether the results from previous competitions carry over to this well-defined set of series.

According to the M3 competition results, one of the most accurate forecasting methods was Forecast Pro (see Goodrich, 2000). This method was arguably the most accurate for seasonal data and was second only to the Theta method (Assimakopoulos and Nikolopoulos, 2000) for non-seasonal data. The overall disappointing performance of ARIMA-based methods led the authors to conclude that statistically sophisticated methods do not necessarily produce more accurate forecasts than simpler ones. Like some of those who commented on the M3, we challenge this conclusion in the context of forecasting tourism demand. We evaluate forecasts from both Forecast Pro and the Theta method, as well as forecasts from two newly proposed fully automated forecasting algorithms, one which identifies and estimates ARIMA models and another which identifies and estimates state space models that underly exponential smoothing methods. We present these methods in Section 2 and evaluate their forecasting performance in Section 5.1.

One of the issues we address is the effect of temporal aggregation on forecast accuracy. In Section 5.2, the monthly series are aggregated to be quarterly, and the quarterly series are aggregated to be yearly. Therefore, we can directly compare the accuracy of the forecasts made before and after aggregation.

The importance of producing statements of uncertainty when forecasting has long been undervalued in the empirical forecasting literature. There is a gap in the literature between the methodological developments in density forecasting (see, for example, Tay and Wallis, 2002; West, 2006) and the applied forecasting papers that do not produce statements of uncertainty to accompany their point forecasts. For instance, a significant limitation of all forecasting competitions to date has been the lack of assessment of whether the forecasting methods can produce reliable forecast intervals. This is highlighted in the commentaries on the M3 competition (see Armstrong, 2001; Goodrich, 2001; Koehler, 2001; Tashman, 2001). Not observing the uncertainty associated with a point forecast can lead to a false sense of accuracy. Economic planning based on forecast intervals can be very different to that based on mean or median forecasts. In Section 5.3 we attempt to fill this gap in the literature by also evaluating the forecast coverage probabilities of the forecasting methods.

Using models with exogenous variables for policy analysis and forecasting is common in both the tourism literature and the tourism industry. These are usually labelled as "causal" approaches, and take on various functional forms (see Song and Witt, 2000, for a detailed exposition on their use within the tourism literature). A major challenge in forecasting with this type of model is that once a satisfactory relationship has been specified, the user needs to produce forecasts of the exogenous variables to be able to forecast the variable of interest. For the first time in a large forecasting competition, we consider the forecasting performance of such models. There are 98 quarterly and 129 yearly cases for which we have available the explanatory variables that are typically used in the literature.

An alternative approach is to treat all of the variables as endogenous (Sims, 1980). This leads us to also consider multivariate models in our forecasting competition, as was suggested by Granger (2001) and Ord (2001). We present the econometric approaches and the methodologies we implement in Section 3. The forecast performance of these models is evaluated in Section 5.4.

1.1 Literature review, rules and objectives

As we mentioned in the introduction, attempting to draw general conclusions from the existing tourism literature through survey articles is very difficult. Papers vary in scope, data frequencies, modelling frameworks and estimation techniques, forecast evaluation rules (ex post versus ex ante for causal models), and forecast horizons, and also in results and conclusions. In this section we establish various rules that will be followed in order to achieve a degree of uniformity and fairness across the application and evaluation of the forecasting methods. We also set some objectives and list the questions we ask in this research.

Outline of rules:

- We aim to apply general modelling frameworks with objective and clearly stated decision rules.
- We aim to replicate, and hence evaluate, some of the typical modelling procedures both within and outside the tourism literature.
- All out-of-sample values are used only for evaluating the forecasts generated by the competing methods, never in the modelling stages.
- No intervention is applied to the pure time series methods in terms of one-off events causing possible structural breaks. However, all such exogenous information is included in the causal approaches.
- All models are estimated once only, and only one set of forecasts is produced for each series.
- All forecasts of exogenous variables are ex ante unless otherwise specified.

We should note that all forecasts from the causal approaches were generated by authors Song and Wu (with some assistance of Athanasopoulos), while all forecasts from the time series methods were produced by authors Athanasopoulos and Hyndman. (See the acknowledgement in Section 7 for the Theta method forecasts.)

Summary of objectives:

- The rapid increase in the capacity of computers to store information has generated an abundance of data across all types of industries. For example, Athanasopoulos et al. (2009) generated forecasts for 117 tourism demand series (and that only includes Australian domestic tourism), disaggregated only by selected geographical areas. In total, Tourism Australia generates forecasts for thousands of series every quarter when considering inbound, outbound and domestic travel, as well as numerous levels of disaggregation such as geographical regions, purpose of travel and so on. Hence, accurate automatic forecasting procedures have become a necessity in order to take advantage of such a plethora of information. We evaluate the performance of three fully automated algorithms (with no intervention).

- ARIMA models have not proven as accurate as other forecasting methods, whether the model identification is automated or implemented manually (refer to Fildes and Ord, 2004, and references therein). In fact, Armstrong (2006) lists these as "tested areas with little gain in accuracy". We re-evaluate the forecasting performance of ARIMA models using a recently proposed algorithm that has shown promising performance in smaller scale evaluations.
- Fildes and Ord (2004) and Armstrong (2006) highlight the dominance of the damped trend method in previous forecasting competitions. We evaluate what is gained or lost by considering aggregate model selection procedures (where a model is selected from various possible candidate models) instead of implementing a specific method for all series, such as obtaining all forecasts from the damped trend method.
- One of the findings of Witt and Witt (1995) is that, for annual data, the Naïve method seems to produce the most accurate forecasts (especially for one year ahead). We revisit this result and also examine whether we can aid the performance of forecasting methods by using higher frequency data.
- Previous studies have found the forecast intervals obtained to be too narrow; hence, the actual values fall outside the empirical forecast intervals more often than they should (see for example Makridakis and Winkler, 1989; Chatfield, 2001; Hyndman et al., 2002). Makridakis et al. (1987) conclude that the lower the frequency of the data, the more the coverage probability of the forecast intervals is over-estimated. We re-examine this conclusion by evaluating the forecast interval coverages for the three automated forecasting algorithms for monthly, quarterly and yearly data.
- Allen and Fildes (2001) found that models with exogenous variables forecast better than extrapolating methods when ex post forecasts are used for the regressors. A surprising result from their study is that the forecasting performance of causal models seems to improve when using ex ante, rather than ex post, forecasts. In the tourism literature, Song et al. (2003a) found that econometric models perform better than the no-change, ARIMA and VAR models, using ex post forecasts. In contrast, Witt and Witt (1995) concluded that causal models are outperformed by the no-change model, regardless of whether ex ante or ex post forecasts are used. We evaluate the performance of models with explanatory variables, as implemented in the tourism literature, in a forecasting competition setting in which all out-of-sample values are ex ante forecasts. We then revisit this result with ex post forecasts for the regressors in order to evaluate the best possible result from using causal models in a scenario-based forecasting framework.

- Despite strong warnings about its limitations (see Hyndman and Koehler, 2006), the MAPE remains the most commonly used forecast error measure among both academics and practitioners (see Fildes and Goodwin, 2007), and the tourism forecasting literature is no exception (see Li et al., 2005; Song and Li, 2008). Here we investigate in detail the forecasting results based on the MAPE and the effect that some of its limitations may have on the results. We also consider the MASE (mean absolute scaled error), which was proposed by Hyndman and Koehler (2006) in order to overcome some of the limitations of the MAPE.

2 Pure time series approaches

In this section we present the details of the three fully automated forecasting procedures and the two method-specific procedures that we have implemented in this paper.

2.1 ARIMA forecasting

A non-seasonal ARIMA(p,d,q) process is given by

$$\phi(B)(1-B^d)y_t = c + \theta(B)\varepsilon_t,$$

where $\{\varepsilon_t\}$ is a white noise process with mean zero and variance σ^2, B is the backshift operator, and $\phi(z)$ and $\theta(z)$ are polynomials of orders p and q respectively. To ensure causality and invertibility, it is assumed that $\phi(z)$ and $\theta(z)$ have no roots for $|z| < 1$ (Brockwell and Davis, 1991). If $c \neq 0$, there is an implied polynomial of order d in the forecast function.

The seasonal ARIMA$(p,d,q)(P,D,Q)_m$ process is given by

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = c + \Theta(B^m)\theta(B)\varepsilon_t,$$

where $\Phi(z)$ and $\Theta(z)$ are polynomials of orders P and Q respectively, each containing no roots inside the unit circle. If $c \neq 0$, there is an implied polynomial of order $d + D$ in the forecast function (Box et al., 2008, pp. 381–382).

The main task in automatic ARIMA forecasting is selecting an appropriate model order, that is, the values of p, q, P, Q, D and d. We use the automatic model selection algorithm that was proposed by Hyndman and Khandakar (2008), which is summarised below.

Diebold and Kilian (2000) find strong evidence that unit root pretesting for selecting the level of differencing d improves the forecasting accuracy. For non-seasonal data (we treat yearly data as non-seasonal) we consider ARIMA(p,d,q) models, where d is selected based on successive KPSS unit-root tests (Kwiatkowski et al., 1992). That is, we test the data for a unit root; if the null hypothesis of no unit root is rejected (at the 5% significance level), we test the differenced data for a unit root; and so on. We stop this procedure the first time we fail to reject the null hypothesis.

For seasonal data we consider ARIMA$(p,d,q)(P,D,Q)_m$ models, where m is the seasonal frequency. Unlike Hyndman and Khandakar (2008), we set $D=1$ for all seasonal data (we treat all monthly and quarterly data as seasonal), as we find that their suggested seasonal unit root test does not help in selecting the appropriate order of differencing (Osborn et al., 1999, reached a similar conclusion for seasonal unit root testing). We then choose d by applying successive KPSS unit root tests to the seasonally differenced data. Once the value of d has been selected, we proceed to select the values of p, q, P and Q by minimising the AIC. We allow $c \neq 0$ for models where $d+D<2$.

Once d and D are known, we select the orders p, q, P and Q via Akaike's Information Criterion:

$$\text{AIC} = -2\log(L) + 2(p+q+P+Q+k),$$

where $k=2$ if $c \neq 0$ and 1 otherwise (the other parameter being σ^2), and L is the maximised likelihood of the model fitted to the *differenced* data $(1-B^m)^D(1-B)^d y_t$. The likelihood of the full model for y_t is not actually defined, and so the values of the AIC for different levels of differencing are not comparable.

There are a large number of potential ARIMA models—too many to allow us to estimate every possible combination of p, q, P and Q. Instead, we need a way to efficiently traverse the space of models in order to arrive at the model with the lowest AIC value. Hyndman and Khandakar (2008) proposed the following step-wise algorithm.

Step 1: Try four possible models to start with:

- ARIMA$(2,d,2)$ if $m=1$ and ARIMA$(2,d,2)(1,D,1)$ if $m>1$.
- ARIMA$(0,d,0)$ if $m=1$ and ARIMA$(0,d,0)(0,D,0)$ if $m>1$.
- ARIMA$(1,d,0)$ if $m=1$ and ARIMA$(1,d,0)(1,D,0)$ if $m>1$.
- ARIMA$(0,d,1)$ if $m=1$ and ARIMA$(0,d,1)(0,D,1)$ if $m>1$.

If $d+D \leq 1$, these models are fitted with $c \neq 0$. Otherwise, set $c=0$. Of these four models, select the one with the smallest AIC value. This is called the "current" model, and is denoted by ARIMA(p,d,q) if $m=1$ or ARIMA$(p,d,q)(P,D,Q)_m$ if $m>1$.

Step 2: Consider up to thirteen variations on the current model:

- where one of p, q, P and Q is allowed to vary from the current model by ± 1;

- where p and q both vary from the current model by ± 1;
- where P and Q both vary from the current model by ± 1;
- where the constant c is included if the current model has $c = 0$ or excluded if the current model has $c \neq 0$.

Whenever a model with a lower AIC is found, it becomes the new "current" model and the procedure is repeated. This process finishes when we cannot find a model close to the current model with a lower AIC.

There are several constraints on the fitted models in order to avoid problems with convergence or near unit roots (see Hyndman and Khandakar, 2008, for details). The algorithm is guaranteed to return a valid model because the model space is finite and at least one of the starting models will be accepted (the model with no AR or MA parameters). The selected model is then used to produce forecasts and forecast intervals.

2.2 Innovations state space models for exponential smoothing

Ord et al. (1997), Hyndman et al. (2002) and Hyndman et al. (2005) (amongst others) have developed a statistical framework for the exponential smoothing methods presented in Table 1. The statistical framework incorporates stochastic models, likelihood calculations, forecast intervals and procedures for model selection. We employ this framework for building innovations state space models. The aforementioned papers have shown that these models generate optimal forecasts for all exponential smoothing methods (including non-linear methods).

The classification of the exponential smoothing methods in Table 1 originated with Pegels (1969) and was expanded by Gardner (1985), Hyndman et al. (2002) and Taylor (2003). Each of the fifteen methods listed has a trend and a seasonal component. Hence, cell (N,N) describes the simple exponential smoothing method, cell (A,N) Holt's linear method, cell (A,A) Holt-Winters' additive method, and so on.

Table 1: *Classification of exponential smoothing methods.*

		Seasonal component		
		N	A	M
Trend component		(None)	(Additive)	(Multiplicative)
N	(None)	N,N	N,A	N,M
A	(Additive)	A,N	A,A	A,M
A_d	(Additive damped)	A_d,N	A_d,A	A_d,M
M	(Multiplicative)	M,N	M,A	M,M
M_d	(Multiplicative damped)	M_d,N	M_d,A	M_d,M

For each method, there are two possible state space models: one corresponding to a model with additive errors and the other to a model with multiplicative errors. Table 2 presents the fifteen models with additive errors and their forecast functions. The multiplicative error models can be obtained by replacing ε_t with $\mu_t\varepsilon_t$ (for further details see Hyndman et al., 2008). We select models by minimising the AIC amongst all models (both additive and multiplicative). We then compute forecast intervals from the selected models using analytic formulae (Hyndman et al., 2008, Chapter 6), or by simulation if the analytic formulae are not available.

Table 2: *State space equations for each additive error model in the classification. Multiplicative error models are obtained by replacing ε_t with $\mu_t\varepsilon_t$. In each case, ℓ_t denotes the level of the series at time t, b_t denotes the slope at time t, s_t denotes the seasonal component of the series at time t, and m denotes the number of seasons in a year; α, β, γ and ϕ are constants with $0 < \alpha, \gamma, \phi < 1$ and $0 < \beta < \alpha$; $\hat{y}_{t+h|t}$ denotes the h-step-ahead forecast based on all of the data up to time t; $\phi_h = \phi + \phi^2 + \cdots + \phi^h$; $\hat{y}_{t+h|t}$ denotes a forecast of y_{t+h} based on all the data up to time t; and $h_m^+ = \lfloor (h-1) \bmod m \rfloor + 1$.*

Trend component	**Seasonal component** **N** (none)	**A** (additive)	**M** (multiplicative)
N (none)	$\mu_t = \ell_{t-1}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t$	$\mu_t = \ell_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t + s_{t-m+h_m^+}$	$\mu_t = \ell_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t s_{t-m+h_m^+}$
A (additive)	$\mu_t = \ell_{t-1} + b_{t-1}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t + hb_t$	$\mu_t = \ell_{t-1} + b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t-m+h_m^+}$	$\mu_t = (\ell_{t-1} + b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$ $\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t-m+h_m^+}$
$\mathbf{A_d}$ (additive damped)	$\mu_t = \ell_{t-1} + \phi b_{t-1}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$	$\mu_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t-m+h_m^+}$	$\mu_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$ $\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t-m+h_m^+}$
M (multiplicative)	$\mu_t = \ell_{t-1}b_{t-1}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t b_t^h$	$\mu_t = \ell_{t-1}b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t b_t^h + s_{t-m+h_m^+}$	$\mu_t = \ell_{t-1}b_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$ $\hat{y}_{t+h\|t} = \ell_t b_t^h s_{t-m+h_m^+}$
$\mathbf{M_d}$ (multiplicative damped)	$\mu_t = \ell_{t-1}b_{t-1}^\phi$ $\ell_t = \ell_{t-1}b_{t-1}^\phi + \alpha\varepsilon_t$ $b_t = b_{t-1}^\phi + \beta\varepsilon_t/\ell_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h}$	$\mu_t = \ell_{t-1}b_{t-1}^\phi + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^\phi + \alpha\varepsilon_t$ $b_t = b_{t-1}^\phi + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} + s_{t-m+h_m^+}$	$\mu_t = \ell_{t-1}b_{t-1}^\phi s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^\phi + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1}^\phi + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} s_{t-m+h_m^+}$

We label this method ETS in the tables that follow. The three letters are an abbreviation of “ExponenTial Smoothing”, and also specify the three components of the stochastic model: Error, Trend and Seasonality. For example, an ETS(A,A,A) is a Holt-Winters’ additive method with an additive error component.

2.3 Forecast Pro

Forecasts from several commercial software packages were considered in the M3 competition. Forecast Pro was arguably the most accurate commercial package, as well as having the most consistent performance across all data. In this paper we evaluate forecasts from the Forecast Pro Extended Edition, Version 4. The forecasting method choice is set to “expert selection”. The software evaluates the forecasting performance of several methods and selects amongst them. Given the nature of the data we consider in this paper, the methods considered by the Forecast Pro algorithm were exponential smoothing, ARIMA models and simple moving averages.

Although the finer details of the model identification, estimation and selection are not revealed, Goodrich (2000) presents some details. Exponential smoothing methods are fitted by minimising the in-sample sum of squared errors. The final method is selected by minimising the BIC, supplemented by some logical rules. With ARIMA models, a general-to-specific approach is followed. First a non-parsimonious state space model is estimated, and is used in turn to obtain approximate parameter estimates for a large number of potential ARIMA models. The final model is selected by the BIC (again supplemented by some logical rules), and then re-estimated using unconditional least squares.

This method is labelled ForePro in the tables that follow. For further details, refer to Goodrich (2000) or to `www.forecastpro.com`.

2.4 Theta method

One method that performed extremely well in the M3 competition (Makridakis and Hibon, 2000) was the Theta method (Assimakopoulos and Nikolopoulos, 2000), which was further analysed and described by Hyndman and Billah (2003). For a given value of θ, a time series y_t is transformed to $x_{t,\theta}$ (dubbed a “theta line”) through

$$x_{t,\theta} = a_\theta + b_\theta(t-1) + \theta y_t, \qquad t = 1, \dots, n.$$

Estimates of a_θ and b_θ are obtained by minimising $\sum_{t=1}^{n}[y_t - x_{t,\theta}]^2$. As in the M3 competition, forecasts are obtained by averaging two theta lines using $\theta = 0$ (which gives a regression time trend) and $\theta = 2$. The theta line for $\theta = 2$ has been extrapolated using simple exponential smoothing for which the smoothing parameter has been chosen by minimising the in-sample one-step-ahead mean squared error, with the starting value for the initial level set equal to y_1.

Hyndman and Billah (2003) show that in this case the forecasts obtained by the Theta method are equivalent to those generated by simple exponential smoothing with an added trend and a constant, where the slope of the trend is half that of a fitted trend line through the original time series y_t.

All monthly and quarterly data are first seasonally adjusted by extracting a seasonal component using classical multiplicative decomposition. The seasonal component is then added to the forecasts generated by the Theta method. The method is implemented using Delphi 7.0 for Windows XP. The forecasting software is TIFIS CM3, which is a non-commercial Forecasting Support System.

2.5 Damped trend

As was highlighted in the introduction, the damped trend method has been singled out from previous forecasting competitions as performing very well. In this paper we estimate the additive damped trend model ETS(A,A$_d$,A) for monthly and quarterly data and ETS(A,A$_d$,N) for yearly data, as presented in Table 2.

2.6 Naïve approaches

We produce forecasts from two naïve approaches which form natural benchmarks. For yearly data, we use $\hat{y}_{t+h|t} = y_t$. Hence, all forecasts are equal to the most recent observation. This method is labelled Naïve in the tables that follow. For monthly and quarterly data, we use $\hat{y}_{t+h|t} = y_{t-m+h_m}$, where $h_m = [(h-1) \mod m] + 1$, with $m = 4$ for quarterly data and $m = 12$ for monthly data. Hence, all forecasts for seasonal data are equal to the most recent observation of the corresponding season. This method is labelled SNaïve, standing for "Seasonal Naïve", in the tables that follow.

3 Models with explanatory variables

The general tourism demand function in the tourism modelling and forecasting literature (e.g., Song and Witt, 2000) takes the form:

$$y_t^i = f(g_t^i, p_t^i, p_t^{is}, dummy\ variables, \varepsilon_t), \tag{1}$$

where y_t^i is the demand variable measured by tourist arrivals from origin country i to (or expenditure in) the destination country; g_t^i is the income level of origin country i in real terms; p_t^i represents the relative cost of living in the destination country for tourists from origin country i, measured as the relative CPI of the destination country to that of the origin country in constant prices, adjusted by the relevant exchange rates; p_t^{is} represents tourism prices in substitute destinations, and is measured by a weighted average price index of a set of alternative destinations to the destination country. For a detailed exposition on the price variables, refer to Wong et al. (2007) for a case study with Hong Kong as the destination country. All variables are transformed to their natural logarithms. The dummy variables include seasonal dummies and one-off events such as terrorist attacks, epidemics, or other events that relate to particular regions, and ε_t is a random error term.

The models we consider are special cases of

$$y_t = \beta_0 + \sum_{j=1}^{k}\sum_{i=0}^{p_j}\beta_{j,i}x_{j,t-i} + \sum_{i=1}^{p}\phi_i y_{t-i} + \varepsilon_t, \tag{2}$$

where y_t is the tourism demand variable (now dropping the superscripts i and s for brevity); $x_{j,t}$ are the exogenous variables included in our model for $j = 1,\dots,k$; p is the maximum number of lags of the regressand and p_j the maximum number of lags of the regressors; and ε_t is the error process, which is assumed to be white noise.

3.1 Autoregressive distributed lag model (ADLM)

In the initial specification of the general ADLM given by equation (2), all possible variables are included. For quarterly data, $p_j = p = 4$, and for annual data, $p_j = p = 1$. This unrestricted specification of the ADLM is an error correction model which has been considered in many previous tourism studies (see Song and Witt, 2003, for further details). In the tables that follow, this specification is labelled as ADLM.

The model is refined by implementing the dynamic econometric modelling technique known as the general-to-specific approach (as advocated by Hendry, 1986). The least significant regressor (i.e., the one with the largest p-value) is deleted from the model, then the simplified model is re-estimated. This process is repeated until the coefficients of all of the remaining regressors are statistically significant at the 5% significance level (one-tailed). The final model should be simple in structure and display no autocorrelation or heteroscedasticity, and preferably no non-normality either. In the tables that follow, this specification is labelled as ADLM$_R$.

Setting $p = p_j = 0$, equation (2) gives a static regression specification which has been considered in many previous tourism forecasting studies (refer to the literature reviews by Li et al., 2005; Song and Li, 2008). We label this specification as SR in the tables that follow. As alternatives to this static specification, we also consider a static regression fitted to the differenced data, which we label ΔSR. In addition, for the quarterly data we consider a static regression fitted to the seasonally differenced data, which we label as Δ^mSR.

In our attempt to be as objective and general as possible within a static regression framework, one might argue that we are missing features of the data that could be captured and exploited through a more rigorous modelling framework, resulting in more accurate forecasts. In an effort to achieve this, we have also implemented the following regression framework, which we label DR (Dynamic Regression) in the tables that follow.

Starting from a static regression in each case, we examine whether any unexploited dynamics remain by observing the estimated autocorrelation and partial autocorrelation functions of the residuals. We first account for unit roots and seasonal unit roots by taking first and/or seasonal differences of all of the variables, guided by the magnitude of the first order and seasonal autocorrelations. Then, any dynamics left over in the residuals are modelled via ARMA specifications, guided by the autocorrelation and partial autocorrelation functions. If the choice between AR and MA components becomes arbitrary (as either one or a combination of them completely captures all dynamics), then this choice is made by minimising the AIC.

3.2 Time varying parameter (TVP) model

The TVP model is used in order to allow the coefficient of the explanatory variables to change over time. This method is more adaptable when the assumption of constant coefficients is not valid, and structural changes in econometric models need tackling. The TVP approach uses a recursive estimation process in which the most recent information is weighted more heavily than the information obtained in the distant past. With the restriction $p = 0$ being imposed on the coefficients in equation (2), the TVP model can be expressed in state space form as:

$$y_t = \boldsymbol{\beta}_t' \boldsymbol{x}_t + \varepsilon_t \tag{3}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\Phi} \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \tag{4}$$

where $\boldsymbol{x}_t$ is a $(k+1)$-dimensional vector of the explanatory variables (including a column of ones for the intercept); $\boldsymbol{\beta}_t$ is a $(k+1)$-dimensional vector of parameters and is known as the *state* *vector*; $\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$ refers to the temporary disturbance; and $\boldsymbol{\omega}_t \sim \text{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\omega)$ is the permanent disturbance. The coefficient matrix Φ is initially assumed to be known.

Equation (3) is called the *measurement* or *system* equation, while equation (4) is known as the *transition* or *state* equation, which is used to simulate the way in which the parameters in the system equation evolve over time. If $\boldsymbol{\Phi}$ is an identity matrix, the transition equation becomes a random walk process:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t. \tag{5}$$

In most cases, the random walk process is adequate for capturing the parameter changes in various economic models (see, for example, Bohara and Sauer, 1992; Kim, 1993; Greenslade and Hall, 1996; Song and Witt, 2000). We adopt this approach for both quarterly and annual data. For quarterly data, seasonal dummies are also included to capture the seasonal component in the data. Equations (3) and (4) are estimated using the Kalman filter (for details of the estimation procedure, see Harvey, 1989). In the tables that follow, we label this specification as TVP.

In order to forecast tourism demand using frameworks that incorporate exogenous variables, we must first produce forecasts for each of the exogenous variables. As is typical in the tourism literature (see Song et al., 2003b; Song and Wong, 2003), we forecast these using exponential smoothing methods. For the quarterly data, we apply Holt-Winters' additive method. For the yearly data, we have either aggregated the quarterly forecasts (where possible) or we apply Holt's linear method.

3.3 Vector autoregressive (VAR) model

In contrast to the previous modelling frameworks, we now treat all of the variables in equation (1) as endogenous:

$$\begin{pmatrix} y_t \\ \boldsymbol{x}_t \end{pmatrix} = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_1 \begin{pmatrix} y_{t-1} \\ \boldsymbol{x}_{t-1} \end{pmatrix} + \cdots + \boldsymbol{\Phi}_p \begin{pmatrix} y_{t-p} \\ \boldsymbol{x}_{t-p} \end{pmatrix} + \boldsymbol{\varepsilon}_t, \tag{6}$$

where $\boldsymbol{\varepsilon}_t \sim \text{NID}(0, \Sigma_\varepsilon)$. This framework was popularized by Sims (1980), and has been the main workhorse in multivariate economic modelling and forecasting ever since. It is also popular in the tourism literature with unrestricted models and also models incorporating long-run restrictions (see, for example, Kulendran and King, 1997; Witt et al., 2003; Song and Witt, 2006; Bonham et al., 2009). A great advantage of this approach in terms of the effort required for forecasting is that the system generates forecasts for all of the variables. Hence, we do not need to generate forecasts separately for the $\boldsymbol{x}$ variables.

We consider three sets of forecasts from the following general approaches.

1. VAR models for all variables in levels. We choose the lag length of the model by minimising the AIC. We label these as VAR(AIC) in the tables that follow. The maximum lag lengths we consider are 4 for quarterly data and 2 for annual data. We also consider the largest of these models: VAR(4) for quarterly data and VAR(2) for annual data. This ensures that at least for quarterly data, the VAR models contain up to and including the seasonal lag.

2. Reduced VAR models for the growth rates of all variables. We first calculate the growth rates of all variables by considering the first differences of the natural logarithms. The growth rate of variable z_t is calculated as $100\ln(z_t/z_{t-1})$. For quarterly data, we select the lag length by minimising the AIC with a maximum lag order of 4. For yearly data we set the lag length to 1, due to the short samples. All insignificant coefficients (at the 5% significance level) are then restricted to zero. We impose the restrictions one at a time by eliminating the parameter with the lowest t-statistic at each step. The restricted VARs are estimated using the seemingly unrelated regression estimation method (Zellner, 1963), as not all equations include the same regressors. Athanasopoulos and Vahid (2008) found this type of VAR model to be successful in forecasting macroeconomic variables. We label them as ΔVAR_R(AIC) in the tables that follow.

4 Data and forecast error measures

The data we use include 366 monthly series, 427 quarterly series and 518 yearly series. They were supplied by both tourism bodies (such as Tourism Australia, the Hong Kong Tourism Board and Tourism New Zealand) and various academics, who had used them in previous tourism forecasting studies (please refer to Section 7 for acknowledgements and details of the data sources and availability). Descriptive statistics for the data are shown in Table 3.

A subset of these series was used for evaluating the forecasting performance of methods that use explanatory variables. There were 93 quarterly series and 129 yearly series for which we had explanatory variables available. With the exception of 34 yearly series (which represented tourism numbers by purpose of travel at a national level) all other series represented total tourism numbers at a country level of aggregation.

For each series we split the data into the estimation sample and a hold-out sample which was hidden from all of the co-authors. For each monthly series, the hold-out sample consisted of the 24 most recent observations; for quarterly data, it was the last 8 observations; and for yearly data it consisted of the final 4 observations. Each method was implemented (or trained) on the estimation sample,

Table 3: *Descriptive statistics of the data*

	Monthly	Quarterly	Yearly
Total no. of series	366	427	518
mean length	298	99	24
median length	330	110	27
min length	91	30	11
max length	333	130	47
No. of series (No. of observations)	74 (≤ 200)	125 (≤ 100)	112 (≤ 20)
	18 (201–300)	302 (> 100)	375 (21–30)
	264 (>300)		31 (>30)

and forecasts were produced for the whole of the hold-out sample for each series. The forecasts were then compared to the actual withheld observations.

For each forecast horizon h, we first consider the percentage better (PB) measure (as in Makridakis and Hibon, 2000). The PB shows the percentage of times that each method produces more accurate forecasts than SNaïve for monthly and quarterly data, and than Naïve for yearly data.

We also consider three alternative forecast error measures: the mean absolute percentage error measure (MAPE)

$$\text{MAPE}_h = \frac{1}{S}\sum_{s=1}^{S}\left|\frac{y_h^s - \hat{y}_h^s}{y_h^s}\right|,$$

and two scaled error measures suggested by Hyndman and Koehler (2006): the mean absolute scaled error (MASE)

$$\text{MASE}_h = \frac{1}{S}\sum_{s=1}^{S}\text{ASE}_h^s,$$

and the median absolute scaled error (MdASE)

$$\text{MdASE}_h = \text{median}(\text{ASE}_h^s),$$

where y_h^s is the hth observation of the hold-out sample of series s and $\hat{y}_h^s$ is the forecasted value of this, S is the number of series,

$$\text{ASE}_h^s = \left|\frac{y_h^s - \hat{y}_h^s}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i^s - y_{i-1}^s|}\right|$$

and n is the number of observations in the estimation sample of series s.

5 Results

5.1 Time series forecasting results

The tables that follow present the PB, MAPE, MASE and MdASE results. The columns labelled 'Average $1-h$' show the average forecast error measure over the forecast horizons 1 to h. The last column of each table, labelled 'Average rank,' shows the average ranking for each forecasting method over all forecast horizons in the hold-out sample.

To assist in evaluating the forecasting performance of the methods, we consider three aspects of the results: (i) the average rankings of the methods over all forecast horizons, i.e., the last column in each table; (ii) the rankings of the methods for the average error measures over the subsets $1-h$ considered in each case; and (iii) the performance of the methods for $h=1$ step ahead. These aspects are considered in no particular order of importance, and lead us to some general conclusions.

Monthly data

The results for monthly data are presented in Table 4, and are summarised as follows:

- Forecast Pro, ARIMA and ETS consistently forecast more accurately than SNaïve.
- The improvements in forecast accuracy over SNaïve are in places quite considerable. For example, for short-term forecast horizons ($h=1$ to 6) Forecast Pro, ARIMA and ETS achieve improvements of over 10% both for MAPE and MASE.
- When considering the MAPE, Forecast Pro produces the most accurate forecasts, but when considering the MASE, the ARIMA methodology is more accurate.
- The Theta method and the Damped trend method seem to be inferior to the other methods for forecasting monthly data.
- However, the Damped trend method is much more accurate than the Theta method and SNaïve for one-step-ahead forecasting.

Table 4: *Forecast accuracy measures for monthly data*

Method	Forecast horizon (h) 1	2	3	6	12	18	24	Average 1–3	1–12	1–24	Average rank
	PB relative to SNaïve										
ARIMA	**61.48**	**63.11**	58.20	55.46	53.01	53.01	**58.20**	**60.93**	**57.10**	**56.96**	2.00
ForePro	57.92	62.30	**59.02**	53.28	52.73	52.19	55.46	59.74	56.17	56.01	2.33
ETS	53.55	59.02	58.74	**56.28**	**53.28**	50.27	51.09	57.10	54.94	54.63	3.00
Theta	49.18	53.83	49.45	52.19	51.64	**54.92**	55.46	50.82	51.68	53.19	3.67
Damped	58.20	61.20	55.46	48.09	50.27	43.99	54.10	58.29	52.80	53.20	3.79
	MAPE										
ForePro	**16.75**	**16.22**	**17.17**	**17.32**	20.54	**17.11**	23.27	**16.71**	**18.38**	**19.91**	1.46
ETS	17.86	17.30	18.30	20.89	**20.44**	19.74	23.65	17.82	19.67	21.15	2.92
ARIMA	17.38	17.65	18.45	19.13	21.09	18.02	24.29	17.83	19.37	21.13	3.17
Theta	19.29	20.11	20.30	20.20	21.02	18.50	22.51	19.90	21.02	22.11	4.21
SNaïve	*19.89*	*21.56*	*20.64*	*20.94*	*21.09*	*19.97*	***22.30***	*20.70*	*21.38*	*22.56*	*4.54*
Damped	17.90	19.03	22.25	26.53	20.70	24.19	22.35	19.73	22.30	23.47	4.71
	MASE										
ARIMA	**1.00**	**1.16**	1.28	**1.29**	**1.07**	**1.68**	**1.45**	1.15	**1.21**	**1.38**	1.67
ForePro	1.02	1.17	**1.25**	1.30	1.12	1.69	1.54	**1.14**	1.22	1.40	2.25
ETS	1.19	1.26	1.32	1.40	1.14	1.88	1.61	1.26	1.30	1.49	3.75
SNaïve	*1.23*	*1.43*	*1.40*	*1.47*	*1.09*	*1.78*	*1.48*	*1.35*	*1.37*	*1.54*	*4.33*
Theta	1.35	1.50	1.71	1.43	1.15	1.73	1.48	1.52	1.42	1.55	4.33
Damped	1.08	1.36	1.67	1.71	1.08	2.19	1.47	1.37	1.47	1.66	4.67
	MdASE										
ARIMA	**0.78**	0.89	1.02	1.01	0.77	1.13	1.09	0.90	0.90	1.02	2.71
ForePro	0.82	**0.86**	**0.83**	0.95	0.85	1.14	1.12	**0.84**	**0.89**	**1.01**	2.88
ETS	0.89	0.87	0.91	0.96	0.82	1.23	1.19	0.89	0.91	1.03	3.13
Theta	0.95	0.94	1.01	**0.93**	0.85	**1.12**	1.17	0.97	0.96	1.06	3.50
Damped	**0.78**	0.93	0.98	1.19	**0.81**	1.51	**1.08**	0.90	0.98	1.11	4.00
SNaïve	*1.01*	*1.18*	*1.05*	*1.05*	*0.85*	*1.21*	*1.13*	*1.08*	*1.02*	*1.14*	*4.79*

Quarterly data

A summary of the results for quarterly data presented in Table 5:

- Forecast Pro and ARIMA consistently forecast better than SNaïve.
- ETS forecasts more accurately than SNaïve for the first four quarters.
- For $h = 1$ and 2, the improvements of these methods over SNaïve are of the order of 10% or more.
- The Damped trend method forecasts quarterly data extremely well. It is consistently in the top two methods, regardless of the forecast error measure considered.
- When it is applied to quarterly data, the Theta method is still generally outperformed by SNaïve.

- SNaïve seems to produce forecasts which are more accurate than those of any of the other methods for the seasonal horizons (i.e., $h = 4$ and $h = 8$). Even when considering the MdASE, SNaïve forecasts more accurately than ETS and ARIMA (though only marginally).

Table 5: *Forecast accuracy measures for quarterly data*

Method	Forecast horizon (h)						Average		Average
	1	2	3	4	6	8	1–4	1–8	rank
	PB relative to SNaïve								
Damped	62.06	**60.66**	**54.80**	**52.69**	**67.68**	**54.57**	**57.55**	**58.02**	1.50
ForePro	**62.76**	55.74	52.69	52.22	60.42	52.69	55.85	55.91	2.75
ARIMA	62.30	58.08	52.22	49.18	57.38	52.46	55.44	56.15	2.75
Theta	53.40	51.52	49.65	48.48	59.72	54.33	50.76	53.45	3.88
ETS	60.89	52.46	53.40	47.78	57.38	50.82	53.63	53.81	4.00
	MAPE								
ForePro	11.78	12.38	13.99	14.21	15.05	22.90	**13.09**	15.72	2.63
Damped	11.91	**11.68**	14.85	14.21	**13.83**	22.28	13.16	**15.56**	3.13
ETS	**11.73**	12.59	**13.70**	14.78	15.88	24.01	13.20	16.05	3.25
SNaïve	13.95	14.79	14.41	**13.61**	18.02	**21.15**	14.19	16.46	3.75
Theta	13.89	13.90	14.47	15.07	15.24	21.71	14.33	16.15	3.88
ARIMA	12.81	12.72	14.67	14.79	16.13	22.21	13.75	16.23	4.38
	MASE								
ARIMA	**1.10**	1.30	1.18	1.24	1.80	1.80	1.21	1.47	2.63
Damped	1.11	**1.18**	1.21	1.23	**1.60**	1.81	**1.18**	**1.43**	2.63
ForePro	1.13	1.29	**1.16**	1.22	1.79	1.86	1.20	1.48	2.88
SNaïve	1.34	1.45	1.22	**1.18**	2.08	**1.79**	1.30	1.59	4.00
Theta	1.47	1.43	1.22	1.28	1.74	**1.79**	1.35	1.56	4.13
ETS	1.19	1.36	1.17	1.30	1.96	1.99	1.26	1.58	4.75
	MdASE								
Damped	0.92	**0.93**	0.90	**0.89**	**1.12**	1.45	0.91	**1.08**	2.75
Theta	1.04	0.99	0.88	0.92	1.26	**1.34**	0.96	1.11	2.88
ARIMA	0.87	1.01	0.86	0.93	1.31	1.45	0.92	1.11	3.00
ForePro	**0.83**	1.00	0.85	0.91	1.39	1.52	**0.90**	1.14	3.63
ETS	0.90	1.01	**0.81**	0.94	1.34	1.50	0.92	1.14	4.25
SNaïve	1.15	1.08	0.90	0.92	1.57	1.39	1.01	1.21	4.50

Yearly data

A summary of the results for yearly data which are presented in Table 6:

- When considering the PB measure over all forecast horizons, all time series approaches forecast more accurately more times than Naïve.
- When considering average error measures, the Theta method is the only method that is competitive to Naïve. The Theta method forecasts more accurately than Naïve when considering MASE.

Table 6: *Forecast accuracy measures for yearly data*

Method	Forecast horizon (h)				Average		Average
	1	2	3	4	1–2	1–4	rank
	PB relative to Naïve						
Theta	**50.58**	**60.62**	**70.85**	**71.04**	**55.60**	**63.27**	1.00
ForePro	47.63	57.94	59.89	58.50	52.79	55.99	2.25
Damped	44.40	54.83	59.46	60.42	49.61	54.78	3.75
ETS	47.30	55.02	58.11	57.53	51.16	54.49	4.00
ARIMA	44.80	55.20	58.20	57.51	50.00	53.93	4.00
	MAPE						
Naïve	**21.47**	**20.80**	24.12	28.05	**21.14**	23.61	1.50
Theta	23.06	21.17	**22.94**	**26.61**	22.12	**23.45**	1.50
ForePro	23.71	22.49	27.28	31.96	23.10	26.36	3.25
ETS	23.57	23.26	28.56	35.35	23.41	27.68	4.50
ARIMA	25.06	25.32	28.06	33.69	25.19	28.03	5.00
Damped	24.71	24.41	29.43	34.05	24.56	28.15	5.25
	MASE						
Theta	**1.32**	**1.96**	**2.63**	**3.20**	**1.64**	**2.28**	1.00
Naïve	**1.32**	2.08	2.95	3.64	1.70	2.50	2.00
ForePro	1.49	2.18	3.10	3.85	1.83	2.65	3.50
ARIMA	1.56	2.20	3.05	3.70	1.88	2.63	4.00
ETS	1.50	2.22	3.13	4.01	1.86	2.71	5.00
Damped	1.55	2.29	3.23	3.92	1.92	2.75	5.50
	MdASE						
ForePro	**1.06**	**1.49**	2.28	2.88	**1.28**	1.93	1.50
Theta	1.10	1.56	**2.21**	**2.69**	1.33	**1.89**	1.75
ETS	1.09	1.62	2.29	3.01	1.35	2.00	3.75
Damped	1.13	1.61	2.34	2.91	1.37	2.00	3.75
ARIMA	1.14	1.70	2.35	2.92	1.42	2.02	5.25
Naïve	1.10	1.62	2.43	3.16	1.36	2.08	5.00

5.2 Does temporal aggregation improve the forecast accuracy?

The analysis so far has shown that Forecast Pro, ETS and ARIMA produce the most accurate forecasts for seasonal data. On average these methods produce more accurate forecasts than SNaïve. However, for yearly data none of the three methods forecast more accurately than Naïve in our evaluation.

In this section we investigate whether temporally aggregating the forecasts generated from these methods for higher frequency data, can produce more accurate forecasts for yearly data. We use all 366 monthly series and forecast $h = 1$ and $h = 2$ years ahead. The results from this investigation are presented in Table 7.

Table 7: *Comparing forecast errors from forecasting yearly data directly and temporally aggregating the forecasts produced for monthly and quarterly data*

	ETS		ARIMA		ForePro	
	$h=1$	$h=2$	$h=1$	$h=2$	$h=1$	$h=2$
	MAPE					
Yearly	11.79	16.49	10.99	14.59	11.44	15.36
Quarterly to Yearly	10.32	14.32	9.94	13.98	9.95	14.48
Monthly to Yearly	10.29	14.29	9.93	13.96	9.92	14.46
Yearly from Naïve	*10.70*	*15.01*	*10.70*	*15.01*	*10.70*	*15.01*
	MASE					
Yearly	1.50	2.25	1.43	2.01	1.49	2.15
Quarterly to Yearly	1.37	2.09	1.28	1.89	1.29	2.05
Monthly to Yearly	1.36	2.08	1.28	1.89	1.29	2.04
Yearly from Naïve	*1.43*	*2.17*	*1.43*	*2.17*	*1.43*	*2.17*
	MdASE					
Yearly	1.21	1.85	1.16	1.72	1.30	1.87
Quarterly to Yearly	1.09	1.72	1.06	1.65	1.11	1.78
Monthly to Yearly	1.08	1.71	1.05	1.63	1.11	1.78
Yearly from Naïve	*1.27*	*1.93*	*1.27*	*1.93*	*1.27*	*1.93*

The rows labelled 'Yearly' in Table 7 show the forecast error measures for $h = 1$ and $h = 2$ when forecasting the yearly data directly using each of the three methods. The rows labelled 'Monthly to Yearly' and 'Quarterly to Yearly' show the forecast error measures when forecasting the yearly data; however, the forecasts now come from aggregating the forecasts generated for the monthly and quarterly series, respectively. These results show that, in all cases, the aggregated forecasts (whether they are produced from the monthly data or the quarterly data) are more accurate than the forecasts produced from the yearly data directly. The rows labelled '*Yearly from Naïve*' show the Naïve forecast errors from forecasting the yearly data directly. In each case, the aggregated forecast error measures are smaller than the Naïve ones.

5.3 Forecast interval coverage

Producing estimates of uncertainty is an important aspect of forecasting which is often ignored in academic empirical applications, and is even more neglected in business practice. In this section we evaluate the performance of forecasting methods in producing forecast intervals that provide coverages which are close to the nominal rates. Tables 8, 9 and 10 show the percentage of times that the nominal 95% and 80% forecast intervals contain the true observations for monthly, quarterly and yearly data respectively.

As was discussed in Section 1.1, forecasting methods often tend to overestimate the coverage probabilities of the forecast intervals they generate. This is the case with the forecast intervals

Table 8: *Forecast interval coverage for monthly data*

Nominal	Forecast horizon (h)											Average	
coverage	1	2	3	4	5	6	9	12	15	18	24	1–12	1–24
	Forecast Pro												
95%	95	95	93	91	90	94	94	94	93	92	92	93	93
80%	85	83	82	84	80	82	84	84	84	84	85	83	83
	ETS												
95%	95	93	92	93	92	93	94	95	95	94	95	94	94
80%	83	82	84	86	83	84	87	86	84	85	84	85	85
	ARIMA												
95%	89	85	83	87	78	85	87	89	83	83	84	86	85
80%	77	70	66	72	62	69	73	74	67	69	67	71	70

Table 9: *Forecast interval coverage for quarterly data*

Nominal	Forecast horizon (h)								Average	
coverage	1	2	3	4	5	6	7	8	1–4	1–8
	Forecast Pro									
95%	94	95	94	94	93	91	92	90	94	93
80%	84	85	87	84	83	83	78	77	85	83
	ETS									
95%	94	97	96	95	95	93	92	92	95	94
80%	86	87	87	85	84	87	84	81	86	85
	ARIMA									
95%	82	82	86	83	82	78	82	78	83	82
80%	66	66	74	67	67	63	67	63	68	67

Table 10: *Forecast interval coverage for yearly data*

Nominal coverage	Forecast horizon (h) 1	2	3	4	Average 1–4
	ForePro				
95%	85	82	76	74	79
80%	71	68	62	57	65
	ETS				
95%	85	80	76	75	79
80%	69	68	63	62	66
	ARIMA				
95%	72	71	67	65	69
80%	55	53	49	47	51
	Theta				
95%	78	73	68	64	71
80%	61	57	51	50	55

produced by the ARIMA methodology for all frequencies. However, a significant new finding here is that Forecast Pro and ETS produce coverage probabilities that are very close to the nominal rates for monthly and quarterly data. In fact, these methods slightly *underestimate* the coverage probabilities for the nominal 80% forecast intervals. As Makridakis et al. (1987) found, as we move to lower frequency data there is an increase in the tendency of methods to overestimate coverage probabilities. This is the case with all methods here.

Koehler (2001) stated that it would be interesting to see whether it is possible to find statistically based forecast intervals for the Theta method. These were subsequently derived by Hyndman and Billah (2003). We have applied that result and produced forecast intervals from the Theta method for the annual data (as there is no seasonal component in the model), which are presented in the last two rows of Table 10. As with the other methods, the forecast intervals produced by the Theta method are badly undersized. We should note that there are alternative methods for constructing prediction intervals for each of the forecasting methods. We leave such comparisons for a separate study.

5.4 Forecast evaluation results: cases with explanatory variables

Quarterly data

Table 11 shows the PB results, and Tables 12–13 show the MAPE and MASE results respectively. [1]

[1] We do not present the MdASE results for both quarterly and yearly cases with explanatory variables in order to save space. These are available upon request.

Some general observations from the analysis of these results are as follows.

- The pure time series approaches are consistently the most accurate. The best ranked of these approaches are Damped, Forecast Pro, ARIMA and ETS.
- Of the frameworks that use explanatory variables, the TVP, DR and $\Delta VAR_R(AIC)$ perform best.
- For all forecast horizons $h = 1$ to 8, and for all forecast error measures, one of the pure time series frameworks is always the most accurate.
- It is only for MAPE and for $h = 1$ that models with explanatory variables outperform the pure time series approaches (with the exception of ETS).
- For $h = 4$, not many methods can forecast more accurately than SNaïve.
- The most inaccurate forecasts are those generated by the frameworks that do not perform any differencing. This sends a warning to forecasters who use quarterly variables in levels.
- From the PB results we can compare the forecasting accuracy of the pure time series approaches to that of the most accurate causal model. If for each forecast horizon we compare the PB measure for each pure time series approach (Damped, Forecast Pro, ARIMA and ETS) to the most accurate model with explanatory variables, then the average of these differences across all the forecast horizons still shows an improvement in PB for the pure time series (more than 1%). If in contrast for each forecast horizon we compare the most accurate time series approach to the most accurate model with explanatory variables, the best time series approach forecasts more accurately more than 11% of the time.

Table 11: *PB values for the quarterly cases with explanatory variables*

Method	Forecast horizon (h)					Average		Average rank
	1	2	4	6	8	1–4	1–8	
Damped	64.52	**67.74**	53.76	**73.12**	**58.06**	**60.22**	**61.29**	1.88
ForePro	67.74	51.61	48.39	52.69	47.31	55.91	55.91	3.13
ARIMA	64.52	51.61	46.24	50.54	52.69	55.65	54.84	4.00
ETS	**69.89**	53.76	39.78	51.61	48.39	53.76	53.36	4.88
Theta	55.91	52.69	40.86	56.99	48.39	47.85	50.81	5.75
$\Delta VAR_R(AIC)$	65.59	45.16	41.94	51.61	43.01	51.88	51.75	6.00
TVP	53.76	46.24	44.09	55.91	47.31	46.51	50.00	6.63
$\Delta^m SR$	49.46	50.54	**54.84**	48.39	51.61	52.15	51.08	6.75
DR	62.37	46.24	43.01	45.16	49.46	47.04	47.72	7.63
ΔSR	54.84	40.86	50.54	49.46	45.16	46.77	46.51	9.25
VAR(AIC)	56.99	45.16	35.48	35.48	40.86	44.35	43.15	10.38
VAR(4)	56.99	41.94	30.11	34.41	37.63	41.40	40.73	11.38
ADLM	56.99	38.71	25.81	37.63	32.26	39.52	39.11	12.25
$ADLM_R$	47.31	40.86	34.41	39.78	35.48	39.52	40.19	12.25
SR	40.86	33.33	25.81	32.26	35.48	30.38	32.80	14.63

Table 12: *MAPE values for the quarterly cases with explanatory variables*

Method	Forecast horizon (h)					Average		Average
	1	2	4	6	8	1–4	1–8	rank
Damped	11.10	**8.33**	9.83	**11.79**	29.21	**9.67**	11.90	3.63
ETS	**9.58**	10.00	10.71	15.31	32.65	10.13	**11.81**	4.38
ForePro	11.42	10.62	9.85	14.99	30.00	10.32	12.72	5.13
Theta	12.03	10.75	10.78	14.15	26.52	11.28	12.14	5.25
SNaïve	13.83	11.78	**8.96**	16.14	27.47	11.10	12.62	5.38
TVP	10.35	12.12	9.89	15.37	27.65	11.38	12.58	5.63
DR	10.23	10.55	12.24	17.96	**26.31**	11.46	13.25	6.50
ΔVAR$_R$(AIC)	10.94	11.14	11.25	17.08	38.12	10.91	12.47	7.63
ARIMA	14.25	11.26	11.21	16.62	27.08	11.77	14.13	8.13
ΔSR	10.94	12.46	10.63	17.31	32.73	12.20	13.62	8.38
Δ^mSR	14.90	12.95	9.19	20.76	38.06	12.04	14.08	10.00
VAR(AIC)	10.79	12.25	17.11	22.03	35.40	14.21	17.41	11.25
ADLM$_R$	12.54	14.29	15.83	21.53	33.40	14.56	16.65	11.75
VAR(4)	11.43	13.21	17.62	23.68	37.00	14.59	17.90	13.25
ADLM	11.92	14.06	17.87	22.92	38.70	15.15	18.09	13.75
SR	25.18	27.03	29.67	35.11	48.56	28.13	30.56	16.00

Table 13: *MASE values for the quarterly cases with explanatory variables*

Method	Forecast horizon (h)					Average		Average
	1	2	4	6	8	1–4	1–8	rank
Damped	0.85	**0.99**	1.01	**1.48**	1.65	**0.99**	**1.28**	2.25
ForePro	**0.82**	1.27	0.94	1.84	1.68	1.03	1.34	2.38
ARIMA	0.86	1.35	1.01	1.97	1.66	1.09	1.40	4.50
ETS	0.86	1.22	1.08	1.92	1.94	1.08	1.45	5.25
Theta	1.27	1.33	1.03	1.76	**1.62**	1.28	1.48	5.50
SNaïve	1.22	1.39	**0.92**	2.06	1.69	1.18	1.51	6.38
TVP	1.18	1.48	1.00	1.90	1.75	1.30	1.53	6.75
ΔVAR$_R$(AIC)	0.94	1.34	1.11	2.08	2.21	1.13	1.56	7.25
DR	0.99	1.31	1.19	2.31	2.13	1.24	1.69	8.38
Δ^mSR	1.33	1.60	0.97	2.59	2.02	1.29	1.76	9.50
ΔSR	1.23	1.48	1.09	2.13	2.26	1.36	1.75	10.00
VAR(AIC)	1.10	1.52	1.61	2.70	2.89	1.52	2.12	11.88
VAR(4)	1.04	1.64	1.64	2.93	2.98	1.52	2.19	12.75
ADLM$_R$	1.31	1.75	1.74	2.83	2.97	1.65	2.20	13.25
ADLM	1.11	1.74	1.74	2.97	3.05	1.61	2.23	14.00
SR	3.30	3.39	3.27	4.66	4.64	3.29	3.94	16.00

Yearly data

Table 14 shows the PB results, and Tables 15 and 16 show the MAPE and MASE results respectively.

Some general observations from the analysis of these results are as follows.

- No method can forecast more accurately than Naïve when considering the MAPE.
- The pure time series approaches are consistently more accurate than the methods with explanatory variables.
- For the models with explanatory variables, TVP generates the most accurate forecasts.
- The three methods that compete with Naïve are Theta, TVP and Forecast Pro.
- As was the case with the quarterly data, the results send a clear message to practitioners who model these types of variables in levels, as they are consistently the most inaccurate.
- If for each forecast horizon we consider the average improvement across the four times series approaches (Theta, Forecast Pro, ARIMA and ETS) over the most accurate model with explanatory variables the average improvement is greater than 3%. Comparing the best time series approach to the best model with explanatory variables for each forecast horizon the average improvement is greater than 13%.

Table 14: *PB values for the yearly cases with explanatory variables*

Method	Forecast horizon (h)				Average		Average rank
	1	2	3	4	1–2	1–4	
ForePro	**53.73**	**56.72**	64.18	62.69	**55.22**	59.33	1.50
Theta	45.74	55.81	**69.77**	**69.77**	50.78	**60.27**	1.75
Damped	42.64	45.74	53.49	57.36	44.19	49.81	4.25
ETS	42.64	46.51	52.71	56.59	44.57	49.61	5.00
TVP	43.41	42.64	53.49	54.26	43.02	48.45	5.25
ΔVAR$_R$(AIC)	43.12	43.12	53.21	54.13	43.12	48.39	5.75
DR	44.19	43.41	45.74	48.06	43.80	45.35	7.00
ARIMA	41.76	42.86	51.65	47.25	42.31	45.88	8.50
ADLM	42.64	42.64	46.51	46.51	42.64	44.57	8.75
ΔSR	49.61	41.86	42.64	43.41	45.74	44.38	9.25
ADLM$_R$	41.09	40.31	50.39	49.61	40.70	45.35	9.75
SR	27.91	34.11	45.74	50.39	31.01	39.53	11.25
VAR(AIC)	37.21	41.09	37.98	44.96	39.15	40.31	12.00
VAR(2)	32.56	37.21	34.88	38.76	34.88	35.85	13.50

Table 15: *MAPE values for the yearly cases with explanatory variables*

Method	Forecast horizon (h)				Average		Average rank
	1	2	3	4	1–2	1–4	
Naïve	49.13	**32.65**	27.93	**32.97**	**40.89**	**35.67**	1.50
ForePro	51.94	33.23	28.61	33.74	42.58	36.88	3.25
Theta	54.22	35.16	**27.82**	33.05	44.69	37.56	3.50
TVP	51.27	36.16	30.86	38.34	43.71	39.16	5.00
ETS	51.60	34.45	35.01	41.31	43.03	40.59	5.25
Damped	54.59	36.52	32.17	37.39	45.55	40.16	6.50
ΔSR	**49.11**	35.80	35.67	42.84	42.45	40.86	6.50
DR	53.11	35.39	35.02	41.70	44.25	41.31	7.00
ΔVAR_R(AIC)	55.97	35.31	36.20	40.72	45.64	42.05	7.75
ARIMA	55.45	45.03	33.88	44.49	50.24	44.71	9.75
ADLM_R	63.47	42.70	36.71	43.22	53.09	46.52	10.75
ADLM	64.72	40.18	39.18	45.64	52.45	47.43	11.50
VAR(AIC)	65.23	43.48	47.29	54.47	54.35	52.62	12.75
VAR(2)	72.59	47.73	50.73	58.29	60.16	57.33	14.25
SR	99.77	76.70	49.41	64.07	88.23	72.49	14.75

Table 16: *MASE values for the yearly cases with explanatory variables*

Method	Forecast horizon (h)				Average		Average rank
	1	2	3	4	1–2	1–4	
Theta	0.63	**0.80**	**0.90**	**1.17**	0.72	**0.88**	1.50
Naïve	0.60	0.82	1.01	1.32	**0.71**	0.94	2.25
TVP	**0.59**	0.84	1.01	1.35	0.72	0.95	3.00
ForePro	0.64	0.83	1.01	1.32	0.74	0.95	3.50
Damped	0.68	0.88	1.06	1.35	0.78	0.99	5.50
ARIMA	0.69	0.92	1.06	1.34	0.80	1.00	5.75
ETS	0.70	0.94	1.14	1.46	0.82	1.06	7.75
ΔVAR_R(AIC)	0.73	0.98	1.13	1.46	0.85	1.07	8.25
ADLM_R	0.79	0.98	1.12	1.46	0.88	1.09	9.00
ΔSR	0.66	1.01	1.22	1.62	0.84	1.13	9.75
DR	0.74	0.99	1.19	1.56	0.86	1.12	10.00
ADLM	0.87	1.01	1.25	1.68	0.94	1.20	11.75
VAR(AIC)	0.92	1.15	1.44	1.83	1.03	1.33	13.50
SR	1.10	1.23	1.39	1.79	1.17	1.38	13.75
VAR(2)	1.01	1.29	1.57	2.01	1.15	1.47	14.75

5.5 Ex ante versus ex post forecasting for models with exogenous variables

The results from the models that include exogenous variables indicate that these models cannot forecast as accurately as pure time series approaches. In this section we perform ex post forecasting for these models; i.e., we use the observed out-of-sample values for the exogenous variables. This eliminates any uncertainty related to the forecasting of the exogenous variables. Comparing the ex post forecasts with the ex ante forecasts allows us to evaluate whether the forecasting performance of these models improves enough for it to be deemed reasonable to use models with exogenous variables for scenario-based forecasting.

In Table 17 we present the percentage improvements in forecast accuracy generated by the use of ex post rather than ex ante forecasting for models that use explanatory variables. Here, we only present the results for the three approaches we consider to be the most accurate based on the ex ante results. In general, and surprisingly, we find that these methods become even less accurate when actual out-of-sample values are used for the exogenous variables. There are only two exceptions: for quarterly data, when $h = 1$ for TVP and ΔSR for both MAPE and MASE. However, the improvement in these models for $h = 1$ does not change their rankings against the pure time series approaches. We should note that the methods that we have not presented here do not show any improvement when performing ex post forecasting.

The same curious and counterintuitive result is also observed in Allen and Fildes (2001). The authors attribute this result to the fact that not all of the studies they consider include both ex ante and ex post forecasting. They speculate that possibly models in those studies that only report ex ante forecasts are better specified than in those studies that only report ex post forecasting. In our study, however, this is not the case. All ex post and ex ante forecasting is performed by the same model specifications so there should be no discrepancy between the two.

We are genuinely puzzled by this result. Could it be that these models are generally misspecified in this study? Even if these models are misspecified, why should this cause forecasting bias that favours ex ante more than ex post forecasting? If there was structural change in the relationship between the explanatory variables and the forecast variable, we would expect the ex ante forecasts to do at least as poorly as the ex post forecasts.

Table 17: *Percentage improvements of ex post over ex ante forecasting*

Method	Forecast horizon (h)									
	1	2	3	4	Average	1	2	3	4	Average
	MAPE					MASE				
	Quarterly data									
TVP	5.24	−2.89	−19.51	−17.75	−9.07	5.08	−3.18	−19.76	−5.12	−6.44
DR	−4.24	−10.42	−22.54	−15.54	−13.94	−5.23	−8.77	−19.58	−9.35	−11.44
ΔSR	3.76	−2.63	−10.79	−19.38	−7.39	3.60	−4.36	−12.77	−5.50	−5.30
	Yearly data									
TVP	−15.73	−46.56	−63.45	−77.44	−47.36	−18.20	−33.95	−43.26	−46.74	−38.50
DR	−20.66	−51.58	−59.09	−74.85	−50.03	−23.10	−36.95	−44.82	−48.51	−41.20
ΔSR	−19.07	−47.83	−50.33	−57.71	−41.61	−23.00	−38.35	−44.54	−45.11	−39.81

5.6 Further warnings about the MAPE

The MAPE is clearly the most commonly used forecast accuracy measure in the forecasting literature. Hyndman and Koehler (2006) provide some warnings regarding this measure and highlight the conditions under which it is unsuitable and should not be used. The MAPE is defined only when all of the actual values being forecast are non-zero. In the M3-competition (Makridakis and Hibon, 2000), excessively large or undefined MAPEs were avoided by using positive data only. Also, the MAPE (and all other measures based on percentage errors) assumes a meaningful zero. Both of these conditions hold with the tourism data we use here, and there is no reason for us to think that the MAPE is unsuitable.

However, the MAPEs for the pure time series methods applied to the yearly data (Table 6) are less than half the size of the MAPEs obtained for those cases where explanatory variables are used (Table 15). This is due to nine series that contain low values and are amongst the cases with explanatory variables. These cases produced very large percentage errors, which caused the distribution of the MAPEs to be highly positively skewed. For example, the one-step-ahead forecast errors from the ETS models have a MAPE for all yearly series of 23.6%, whereas it is 51.6% for the cases with explanatory variables. When these nine series are excluded, the MAPEs are 16.3% and 23.5% respectively. Furthermore, when recalculating the MAPEs for the yearly series, excluding these nine cases, we find that the rankings of the methods change. In particular, the Theta method forecasts more accurately than Naïve, which makes the rankings of the methods identical to the rankings given by the MASE.

Consequently, even though the MAPE can formally be applied here, we caution against its use due to the numerical instability that results whenever some series contain small values.

6 Conclusion

We have designed a forecasting competition that uses data from the field of tourism only. The forecast methods we consider are three fully automated time series algorithms (Forecast Pro, ARIMA and exponential smoothing based algorithms), two method-specific approaches (the Theta method and the damped trend), and five general frameworks that incorporate explanatory variables (static and dynamic regression, autoregressive distributed lag models, time varying parameter models and vector autoregressions). We conclude that pure time series approaches forecast tourism demand data more accurately than methods that use explanatory variables. This is a similar result to Witt and Witt (1995) and Kulendran and King (1997) but in contrast to Allen and Fildes (2001) and Song et al. (2003a).

This has immediate practical consequences, as models with explanatory variables are commonly used in both the tourism literature and the tourism industry (especially for scenario-based forecasting). In the most recent study involving models with exogenous variables, Fildes et al. (2010) find some improvements in forecasting accuracy from using a "correctly specified structural econometric model". However, they conclude that simple time series alternatives would probably be appropriate in most cases. In this study (which is much broader than any of the aforementioned studies) we not only find no improvement from using models with exogenous variables for forecasting tourism data, but we find significant improvements from using sophisticated, fully automated time series approaches.

One possible reason for the forecasting inferiority of the models with explanatory variables in comparison to the time series approaches could be that it is more challenging to forecast some of the explanatory variables (e.g., GDP) than directly forecasting the dependent variable. However, we think this can only be part of the explanation as the ex post forecasts were, in general, worse than the ex ante forecasts, whereas the reverse would be true if the difficulty of forecasting the explanatory variables was the only factor.

Another possible reason that may have caused the forecasting failure of the models with explanatory variables is possible model misspecifications. The variables and the specifications used in this paper for the models with explanatory variables are general and typical of the models used in tourism. For example, tourism demand is mostly measured by tourist arrivals to some destination or tourist departures from some destination. This measure is limited in terms of reflecting the actual demand for tourism as it ignores the duration that tourists stay in the destinations and their expenditure while they visit the destination. Furthermore, the price variable is measured by the consumer price index of the destination in most cases (adjusted for exchange rates), which may not exactly capture the movements of tourism prices as the baskets of goods consumed by tourists tend to be different from those consumed by the consumers in the destination.

We understand the need and the usefulness of models with exogenous variables in policy analysis, however we do suggest that the forecasting performance of these models should be evaluated against pure time series alternatives before they are put in use. The most consistent performance of any of the methods that use explanatory variables came from allowing for time varying parameters — a result that is similar to previous studies (see for example Garcia-Ferrer et al., 1987; Song et al., 2003a; Li et al., 2006, among others).

Of the pure time series forecasting approaches, we find that Forecast Pro, ARIMA and ETS consistently forecast more accurately than the seasonal Naïve approach for seasonal data (both monthly and quarterly) and for quarterly data the damped trend method performed extremely well. It is interesting that since the early work of Newbold and Granger (1974), it is the first time in the empirical forecasting literature that an ARIMA based algorithm has produced forecasts as accurate as, or more accurate than, those of its competitors. For both of these seasonal frequencies, Forecast Pro and ETS produce forecast coverage probabilities which are satisfactorily close to the nominal rates. Similar results for ETS were found in the recent study of Kim et al. (2008). We would like to reiterate that not reporting the uncertainty associated with a point forecast can lead to a false sense of accuracy and we encourage forecasters to make the reporting of prediction intervals common practice, especially when these forecasts are directly used for policy making.

For yearly data we find that Naïve produces the most accurate forecasts especially for one year ahead (inline with Witt and Witt, 1995). It is only the Theta method that is competitive to Naïve, a method which also performed extremely well for annual data in the Makridakis and Hibon (2000) M3 forecasting competition where it was first introduced. Aggregating monthly or quarterly forecasts from one of Forecast Pro, ARIMA or ETS to give yearly forecasts also produced more accurate forecasts than Naïve.

Finally, we find that the mean absolute percentage error distribution becomes highly positively skewed, due to a handful of small scaled series. Hence, we verify and endorse statements from previous researchers that the mean absolute scaled error should replace the mean absolute percentage error as the standard measure of forecast accuracy.

7 Acknowledgments

We would like to extend our gratitude to everyone who responded to our "Request for data" letter and supplied us with data. All of the time series data are available from `http://www.forecasters.org/ijf`. In order to adhere to all confidentiality agreements with all parties, the data are presented under coded titles.

We acknowledge valuable comments from Michael Clements (editor), two anonymous referees and Andrew Maurer. We acknowledge Konstantinos Nikolopoulos and Nikolaos Bougioukos for supplying the forecasts from the Theta method. We thank Andrew Maurer and Claude Calero at Tourism Research Australia and Mike Chan and Bruce Bassett at New Zealand's Ministry of Tourism for providing data and explanations. Athanasopoulos and Hyndman acknowledge financial support from Tourism Research Australia and the Sustainable Tourism Cooperative Research Centre. Athanasopoulos acknowledges financial support from Australian Research Council grant DP0984399. Song and Wu would like to thank the Research Grant Committee of Hong Kong for financial support (Grant No. BQ-04H).

References

Allen, P. G. and Fildes, R. (2001). Econometric forecasting. In Armstrong, J., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 271–327. Kluwer Academic Publishing, Boston.

Armstrong, J. (2001). Should we redesign forecasting competitions? *International Journal of Forecasting*, 17:542–545.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: methods for reducing forecast error. *International Journal of Forecasting*, 22:583–598.

Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16:521–530.

Athanasopoulos, G., Ahmed, R. A., and Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, 25:146–166.

Athanasopoulos, G. and Vahid, F. (2008). VARMA versus VAR for macroeconomic forecasting. *Journal of Business and Economic Statistics*, 26:237–252.

Bohara, A. K. and Sauer, C. (1992). Competing macro-hypotheses in the United States: a Kalman filtering approach. *Applied Economics*, 24:389–399.

Bonham, C., Gangnesa, B., and Zhou, T. (2009). Modeling tourism: A fully identified vecm approach. *International Journal of Forecasting*, 25:531–549.

Box, G., Jenkins, G., and Reinsel, G. (2008). *Time Series Analysis*. Wiley, New Jersey, 4th edition.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods*. Springer-Verlag, New York, 2nd edition.

Chatfield, C. (2001). Prediction intervals for time series forecasting. In Armstrong, J., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 475–494. Kluwer Academic Publishing, Boston.

Diebold, F. and Kilian, L. (2000). Unit-root tests are useful for selecting forecasting models. *Journal of Business and Economic Statistics*, 18:265–273.

Fildes, R. and Goodwin, P. (2007). Against your better judgement? How organisations can improve their use of management judgment in forecasting. *Interfaces*, 37:570–576.

Fildes, R. and Ord, J. K. (2004). Forecasting competitions – their role in improving forecasting practice and research. In Clements, M. P. and Hendry, D. F., editors, *A companion to economic forecasting*, pages 322–353. Blackwell, Oxford.

Fildes, R., Wei, Y., and Ismail, S. (2010). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting*, forthcoming.

Garcia-Ferrer, A., Highfield, R., Palm, F., and Zellner, A. (1987). Macroeconomic forecasting using pooled international data. *Journal of Business and Economic Statistics*, 53–67.

Gardner, Jr, E. S. (1985). Exponential smoothing: the state of the art. *Journal of Forecasting*, 4:1–28.

Goodrich, R. (2000). The Forecast Pro methodology. *International Journal of Forecasting*, 16:533–535.

Goodrich, R. (2001). Commercial software in the M3-competition. *International Journal of Forecasting*, 17:560–565.

Granger, C. W. J. (2001). Comments on the M3 forecast evaluation and a comparison with a study by Stock and Watson. *International Journal of Forecasting*, 17:565–567.

Greenslade, J. V. and Hall, S. G. (1996). Modelling economies subject to structural change: the case of Germany. *Economic Modelling*, 13:545–559.

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

Hendry, D. F. (1986). Empirical modelling in dynamic econometrics. *Applied Mathematics and Computation*, 20:201–236.

Hyndman, R. J. and Billah, B. (2003). Unmasking the Theta method. *International Journal of Forecasting*, 19:287–290.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 26:1–22.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22:679–688.

Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2005). Prediction intervals for exponential smoothing using two new classes of state space models. *Journal of Forecasting*, 24:17–37.

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

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

Kim, C. J. (1993). Sources of monetary growth uncertainty and economic activity: the time-varying-parameter model with heteroskedastic disturbances. *Review of Economics and Statistics*, 75:483–492.

Kim, J. H., Song, H., Wong, K., Athanasopoulos, G., and Liu, S. (2008). Beyond point forecasting: evaluation of alternative prediction intervals for tourist arrivals. Working Paper 11/08, Department of Econometrics and Business Statistics, Monash University.

Koehler, A. B. (2001). The asymmetry of the sAPE measure and other comments on the M3-Competition. *International Journal of Forecasting*, 17:570–574.

Kulendran, K. and King, M. L. (1997). Forecasting international quarterly tourist flows using error-correction and time-series models. *International Journal of Forecasting*, 13:319–327.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178.

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

Li, G., Wong, K., Song, H., and Witt, S. (2006). Tourism demand forecasting: A time varying parameter error correction model. *Journal of Travel Research*, 45:175–185.

Makridakis, S. and Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 16:451–476.

Makridakis, S., Hibon, M., Lusk, E., and Belhadjali, M. (1987). Confidence intervals: an empirical investigation of the series in the M-Competition. *International Journal of Forecasting*, 3:489–508.

Makridakis, S. and Winkler, R. (1989). Sampling distributions of post-sample forecasting errors. *Applied Statistics*, 38:331–342.

Newbold, P. and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, 137(2):131–165.

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

Ord, K. (2001). Commentaries on the M3-Competition: an introduction, some comments and a scorecard. *International Journal of Forecasting*, 17:537–584.

Osborn, D. R., Heravi, S., and Birchenhall, C. R. (1999). Seasonal unit roots and forecasts of two-digit European industrial production. *International Journal of Forecasting*, 15:27–47.

Pegels, C. C. (1969). Exponential smoothing: some new variations. *Management Science*, 12:311–315.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

Song, H. and Li, G. (2008). Tourism demand modelling and forecasting – A review of recent literature. *Tourism Management*, 29:203–220.

Song, H., Witt, S., and Jensen, T. (2003a). Tourism forecasting: accuracy of alternative econometric models. *International Journal of Forecasting*, 19:123–141.

Song, H., Witt, S., and Li, G. (2003b). Modelling and forecasting the demand for Thai tourism. *Tourism Economics*, 9:363–387.

Song, H. and Witt, S. F. (2000). *Tourism demand modelling and forecasting: modern econometric approaches*. Oxford, Pergamon.

Song, H. and Witt, S. F. (2003). Tourism forecasting: The general-to-specific approach. *Journal of Travel Research*, 42:65–74.

Song, H. and Witt, S. F. (2006). Forecasting international tourist flows to Macau. *Tourism Management*, 27:214–224.

Song, H. and Wong, K. F. (2003). Tourism demand modeling: A time-varying parameter approach. *Journal of Travel Research*, 42:57–64.

Tashman, L. (2001). The M3-Competition and forecasting software. *International Journal of Forecasting*, 17:578–580.

Tay, A. S. and Wallis, K. F. (2002). Density forecasting: A survey. In Clements, M. P. and Hendry, D. F., editors, *A companion to economic forecasting*. Wiley-Blackwell.

Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, 19:715–725.

West, K. D. (2006). Forecast evaluation. *Handbook of Economic Forecasting*, 1:99–134.

Witt, S. F., Song, H., and Louvieris, P. (2003). Statistical testing in forecasting model selection. *Journal of Travel Research*, 42:151–158.

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

Wong, K., Song, H., Witt, S., and Wu, D. (2007). Tourism forecasting: to combine or not to combine? *Tourism Management*, 28:1068–1078.

World Tourism Organization (2008). *UNWTO World Tourism Barometer*, 6:1–44.

Zellner, A. (1963). Estimators for seemingly unrelated regression equations: Some exact finite sample results. *Journal of the American Statistical Association*, 58:977–992.