Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 27 (2011) 870–886
international journal of forecasting
www.elsevier.com/locate/ijforecast

Combination of long term and short term forecasts, with application to tourism demand forecasting

Robert R. Andrawis[a], Amir F. Atiya[a,*], Hisham El-Shishiny[b]

[a] *Department of Computer Engineering, Cairo University, Giza, Egypt*
[b] *IBM Center for Advanced Studies in Cairo, IBM Cairo Technology Development Center, Giza, Egypt*

Available online 15 September 2010

Abstract

Forecast combination is a well-established and well-tested approach for improving the forecasting accuracy. One beneficial strategy is to use constituent forecasts that have diverse information. In this paper we consider the idea of diversity being accomplished by using different time aggregations. For example, we could create a yearly time series from a monthly time series and produce forecasts for both, then combine the forecasts. These forecasts would each be tracking the dynamics of different time scales, and would therefore add diverse types of information. A comparison of several forecast combination methods, performed in the context of this setup, shows that this is indeed a beneficial strategy and generally provides a forecasting performance that is better than the performances of the individual forecasts that are combined.

As a case study, we consider the problem of forecasting monthly tourism numbers for inbound tourism to Egypt. Specifically, we consider 33 individual source countries, as well as the aggregate. The novel combination strategy also produces a generally improved forecasting accuracy.

Keywords: Time series forecasting; Tourism forecasting; Tourism demand; Forecasting tourism in Egypt; Forecast combination; Exponential smoothing; Holt's model; Bayesian forecasting

1. Introduction

It is well-documented that forecast combinations are often superior to their constituent forecasts. Typically, time-varying conditions among time series, such as regime switching or simply parameter drifts, make identifying the best model among various competitors almost like a moving target. The problem is also aggravated by parameter estimation errors and model misspecification. Forecast combination reduces these unfavorable effects. In several studies (such as those of Clemen, 1989; Makridakis & Hibon, 2000; Stock & Watson, 2004), combined forecasts have

* Corresponding author.
E-mail addresses: robertrezk@yahoo.ca (R.R. Andrawis), amir@alumni.caltech.edu (A.F. Atiya), shishiny@eg.ibm.com (H. El-Shishiny).

0169-2070/$ - see front matter
doi:10.1016/j.ijforecast.2010.05.019

generally been shown to outperform the forecasts from the single best model.

One favorable feature to have in a forecast combination system is diversity in the underlying forecasting models, as a safeguard against focusing over much on a narrow specification (see Armstrong, 2001). Diversity is typically achieved by using different forecasting models, different explanatory variables, or possibly non-overlapping estimation periods. However, most methods for combining forecasts consider time series with identical timings. It is possible, however, that different time frames will introduce additional complementary information that will only help to improve the forecasting performance. In this paper we investigate the benefits of combining forecasts obtained using different time aggregations. For example, we could have a monthly time series where we need a long horizon forecast, such as 12 or 24 months ahead. We aggregate the time series (timewise) to obtain a yearly series. By forecasting both the monthly and time-aggregated yearly series, and combining their forecasts, we make use of both the short-term dynamics (exemplified by the monthly series) and the long-term dynamics (exemplified by the yearly series). Moreover, the short-term forecast should have a greater influence on the time period a short time ahead. Conversely, the long-term forecast will probably be more influential in the later months of the horizon. These aspects can be tuned by variable weighting according to the number of steps ahead being forecasted.

The other topic that we consider is tourism demand forecasting. Tourism is a major sector in the economies of many countries. In fact, tourism is one of the fastest growing sectors in the world economy. Tourist arrivals grew by 6% during 2007, reaching 900 million worldwide, and producing over 600 billion dollars in revenue.[1] It is therefore very important for decision makers to have accurate forecasts of tourism numbers. We apply the developed long-term/short-term combination methodology to the problem of forecasting the inbound tourism demand for Egypt. Specifically, the goal is to forecast tourist arrivals from 33 major source countries, as well as total tourist arrivals.

[1] World Tourism Organization (http://www.unwto.org/aboutwto/why/en/why.php?op=1).

Forecast combination in the context of tourism forecasting has not been considered until recently. In fact, we have been able to identify only a few such studies, despite the importance of the topic and its potential impact on forecasting accuracy. A recent major survey article by Song and Li (2008) recommends that the research community conduct more studies on forecast combination in the context of tourism forecasting. Specifically, they say:

"more efforts are needed to look at the forecasting accuracy improvement through forecast combinations. For example, more complex combination techniques, additional advanced individual forecasting methods and multiple forecasting horizons should all be considered in future studies".

We hope that this study will be one more step toward exploring such an aspect. In summary, the contributions of this work are:

- We make the point that combining short-term and long-term forecasts is likely to lead to a forecast performance which is better than that of either one on its own. This is confirmed by conducting tests on two large business time series benchmarks. As such, this strategy could be a serious contender for forecasting problems involving monthly time series (with long enough forecast horizons).
- We compare 15 major forecast combination methods, to determine which methods are best suited to this different time aggregation combination framework.
- Some of the forecast combination methods considered are novel, and thus this study is also a contribution to the general topic of forecast combination. Examples of such methods are the combination method based on testing performance differences, and the hierarchical forecast combination (i.e., a combination of several linearly and nonlinearly combined forecasts).
- We apply the proposed short-term/long-term combination approach to the tourism forecasting problem. Thereby, we make use of the lessons learned from the experiments above to determine a methodology for applying the combination framework to the problem of tourism forecasting. This tourism application also confirms the superiority of the proposed approach.

2. Previous work

There has been very little work in the literature on combining short-term and long-term forecasts. The few works that we have found are described in what follows. Trabelsi and Hillmer (1989) developed an approach for combining forecasts when the timing is not the same (for example combining monthly forecasts from one source with yearly forecasts from a different source). They considered an ARIMA-modeled monthly series, and obtained an analytical solution to the combination problem, essentially by estimating the covariances of the error terms of the short-term and long-term time series. Greene, Howrey, and Hymans (1986) also investigated the problem of combining forecasts when the timing is not the same, specifically modifying quarterly forecasts from an econometric model using additional monthly time series information. Cholette (1982) likewise discussed the use of benchmark forecasts for modifying forecasts from an ARIMA model applied to a monthly time series. For example, they considered forecasting a monthly time series using a model like ARIMA, and then combining the forecasts with quarterly forecasts available from experts. Engle, Granger, and Hallman (1989) considered the problem of combining short-term and long-term forecasts. They considered a separate set of forecasts made by each of the short-term and long-term models, then merged the two sets of forecasts, and the combined forecasts outperformed either set alone. They used the concept of cointegration. A good analysis of the effects of time aggregation and how they can affect cointegration is also given by Granger (1993). Casals, Jerez, and Sotoca (2009) considered the case where a number of time series are observed at different frequencies. With the help of a state space formulation, they studied the way in which aggregation over time affects both the dynamic components of the time series and their observability. They analytically related the forecast variances of the high frequency series to those of the time-aggregated series, and if followed through, their analysis has interesting implications for combining forecasts for different time aggregations. Riedel and Gabrys (2007) considered the concept of so-called multi-level forecasting. In their approach they consider time series that can be grouped in a hierarchical manner, and combine the forecasts obtained at the different hierarchical levels. For example, airline reservations can be viewed by fare class, by origin–destination itinerary, and by point of sale. By performing different aggregations and combining the corresponding forecasts, they obtain an improved forecasting performance.

However, our approach will differ from the reviewed literature. We will not assume any specific data generating model, such as ARIMA, and we will use (and compare) the standard forecast combination strategies (as well as some novel ones). As such, our work is more empirical, as it seeks to answer the question of whether combining short- and long-term forecasts is beneficial for general business applications, and for tourism forecasting in particular. The purpose of this is to provide the forecaster with a useful tool for improving the forecasting performance.

The past work on tourism forecasting, on the other hand, is extensive. Tourism forecasting can be categorized as following one of two main approaches (see Frechtling, 1996; Li, Song, & Witt, 2005; Song & Li, 2008; Witt & Witt, 1995; and Wong & Song, 2002). The first is qualitative forecasting, such as judgemental forecasting and Delphi-style methods. The second approach, quantitative forecasting, can be further subdivided into two major categories. The first (and the one considered most often in the literature) is econometric forecasting. In this approach the tourism demand is forecast using a number of explanatory variables (for example GDP of originating country, CPI of inbound country, etc.). The goal in this approach is not just to produce accurate forecasts, but also to model the relationships among the variables. The second approach, which is closely related to our work, is the time series approach.

In the business time series forecasting literature there have typically been two competing methodologies: exponential smoothing type models and the ARIMA/Box-Jenkins approach. These two methodologies have also been applied extensively to the tourism forecasting problem. These applications cover a variety of possible tourist destinations/origins, and a variety of forecast horizons. Examples from the first category include the work of Lim and McAleer (2001), who applied various types of exponential smoothing models to tourist arrivals to Australia; Witt, Newbould, and Watkins (1992), who applied exponential smoothing to domestic tourism to Las Vegas and showed that it obtains a level of accuracy comparable to those of other, more sophisticated models; and

Bermudez, Segura, and Vercher (2007), who applied Holt-Winters model (the seasonal version of Holt's exponential smoothing) to UK arrivals by air.

There have also been many studies applying the second approach. For example, Chu (2009) applied three univariate ARMA-based models to tourism demand for a number of Asian countries, and showed that these models perform very well. Chang, Sriboonchitta, and Wiboonpongse (2009) applied a Box-Jenkins methodology to inbound tourism in Thailand. They also included a test for the presence of unit roots and seasonal unit roots. More sophisticated approaches have also been tested in the tourism forecasting domain. For example, du Preez and Witt (2003) applied multivariate models to forecasting tourist arrivals to the Seychelles from a number of European countries, viewing the multi-origin time series as a vector process. Song and Witt (2004) considered vector autoregressive (VAR) modeling of inbound tourism to Macau, where the multivariate process contained the exploratory variables, and found that the VAR model produces superior results for medium- and long-term horizons. Wong, Song, and Chon (2006) applied a Bayesian version of the VAR to tourism demand for Hong Kong (the so-called BVAR model), and showed that this model invariably outperforms its unrestricted VAR counterpart. Goh and Law (2002) applied a seasonal ARIMA model to inbound tourism in Hong Kong. Gil-Alana, Cunado, and de Gracia (2008) considered the problem of seasonal analysis for inbound tourism to the Canary Islands, Spain. They considered both deterministic and stochastic seasonality. For the latter they employed seasonal unit roots and seasonally fractionally integrated models. Chu (2008) applied ARFIMA to inbound tourism to Singapore. They showed that their proposed model gives convincingly better results than traditional approaches. Athanasopoulos and Hyndman (2008) considered a hybrid econometric/time series model for domestic Australian tourism, using a state space approach, while for the same problem of forecasting domestic Australian tourism, Athanasopoulos, Ahmed, and Hyndman (2009) considered a hierarchical forecasting model based on disaggregating the data for different geographical regions. They proposed two new methods for estimating the forecasts at the different levels of aggregation.

Novel models such as neural networks have also been investigated in the tourism forecasting literature. For example, Kon and Turner (2005) applied neural networks to tourism demand for Singapore, and showed that their neural network obtained better results than traditional approaches. Medeiros, McAleer, Slottje, Ramos, and Rey-Maquieira (2008) considered tourist arrivals for the Balearic Islands, Spain, using a neural network forecasting model that incorporates the time-varying conditional volatility. This is accomplished through the use of the so-called neural network regression with GARCH errors (NN-GARCH), where the parameters are estimated using the quasi-likelihood. Other novel approaches include the work of Petropoulos, Nikolopoulos, Patelis, and Assimakopoulos (2005), who applied the technical analysis methods of stock forecasting to tourist arrivals to Greece and Italy. Technical analysis is a forecasting methodology that is typically applied to financial market forecasting (for example, they used the relative strength indicator (RSI) and the trend lines).

Forecast combination has also been considered in the context of tourism forecasting, though not as much as it deserves, given its importance and the significant impact it has on accuracy. We have identified only the following studies on the application of forecast combination to tourism forecasting.

The earliest work was performed by Fritz, Brandon, and Xander (1984), who studied the combination of time series and econometric forecasts. Chu (1998) later developed a combined seasonal ARIMA and sine wave nonlinear regression forecasting model for inbound tourism demand to Singapore. More recent work includes the study by Oh and Morzuch (2005), who showed that a combined forecast using a simple average always outperforms the poorest individual forecasts, and often outperforms the best individual model as well. Wong, Song, Witt, and Wu (2007) compared the results of three different forecast combination strategies for tourism arrivals in Hong Kong from ten major sources, and found that even though forecast combination strategies do not always beat the best single model, they are almost always better than the worst model. This suggests that forecast combination can considerably reduce the risk of forecasting failure. Song, Witt, Wong, and Wu (2009) also explored forecast combination for tourism data, and obtained even more favorable results than Wong et al. (2007).

Specifically, all forecast combination strategies are more accurate than the average individual model for all horizons, with the improvement being more significant for longer-term forecasting. Shen, Li, and Song (2008) compared three forecast combination methods, namely the simple average, the variance-covariance combination method and the discounted mean square error method, and found that the variance-covariance combination method performs best. Moreover, they found that forecast combination methods are superior to the best of the individual forecasts.

This is just a snapshot of the published work related to tourism forecasting. It is beyond the scope of this paper to provide a thorough review, as the body of literature is vast. A very thorough and up to date review is given by Song and Li (2008). Interestingly, there has been little work on forecasting tourism for Egypt, either inbound or outbound (there have been a few econometric forecasting models, such as those of Hilaly & El-Shishiny, 2008, Kamel, Atiya, El Gayar, & El-Shishiny, 2008, and Zaki, 2008, but no time series approaches). As such, this study forms one of a very small group of investigations of this aspect, especially the time series approach.

3. The proposed setup

Let x_t be a monthly time series, where t indexes the month number. We perform a time aggregation step to convert it to a yearly time series y_τ, preferably by calendar year. This means that y_τ equals the sum of the x_t's whose month falls in the year τ considered. We apply a forecasting model to the monthly time series to forecast 12 or 24 months ahead, for example. Similarly, we forecast the yearly time series one or two years ahead. To combine the two forecasts, we have to interpolate the forecasts for the yearly series to obtain forecasts at a monthly frequency. The forecasts from the monthly time series are then combined with the monthly forecasts derived from the yearly time series. This conversion of yearly forecasts to a monthly frequency is performed by a simple linear interpolation scheme. This scheme has the following conditions: (a) the forecast for the year equals the sum of the constituent monthly forecasts; (b) the monthly forecasts for a specific year follow a straight line; (c) the line's slope can change from one year to the next, generally leading to a piecewise linear function; (d) furthermore, this piecewise linear function is a continuous function; that is, the ending point for one year is the starting point for the next year. The computation required to perform this translation from yearly to monthly forecasts is fairly simple.

4. Forecast combination methods

Let the short- and long-term time series be $x_t^{(1)}$ and $x_t^{(2)}$ respectively (with monthly frequencies), and let their h-step-ahead forecasts be $\hat{x}_{t+h}^{(1)}$ and $\hat{x}_{t+h}^{(2)}$ respectively. For example, $\hat{x}_{t+h}^{(1)}$ would be the monthly forecasts and $\hat{x}_{t+h}^{(2)}$ would be the yearly forecasts, converted to a monthly frequency (as described in the previous section). These two forecasts are usually combined using two combination weights w_1 and w_2 to produce the combined forecast $\hat{z}_{t+h}$, as follows:

$$\hat{z}_{t+h} = w_1 \hat{x}_{t+h}^{(1)} + w_2 \hat{x}_{t+h}^{(2)}. \tag{1}$$

We tested the following forecast combination methods (closely following the terminology of Timmermann, 2006).

4.1. Simple average (AVG)

In this scheme, the forecast is the simple average of the two individual forecasts, that is $w_1 = w_2 = \frac{1}{2}$.

4.2. Variance based (VAR)

In this approach, we assume that the two forecasts are unbiased, with variances equal to σ_1^2 and σ_2^2 respectively. Let σ_{12} denote the covariance of the two forecasts. Then, assuming that the weights sum to 1, the optimal weights can be obtained as (see Timmermann, 2006, for the derivation):

$$w_1 = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} \tag{2}$$

$$w_2 = 1 - w_1. \tag{3}$$

The error from estimating the covariance could possibly have an impact on the accuracy of the combined forecast, and therefore, one way is to ignore the covariance between the two forecasts. The optimal weights

can then be written as:

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{4}$$

$$w_2 = 1 - w_1. \tag{5}$$

We abbreviate this method as *VAR-NO-CORR*.

4.3. Inverse of the mean square error (INV-MSE)

Stock and Watson (1999) introduced a method whereby the weights are proportional to the inverse of the mean square error (MSE). It is closely related to the VAR-NO-CORR method. We modified this method such that the weights vary with the forecast horizon. This means that we use different combination weights for every month ahead (of the 12 or 24 months to be forecast). However, if we were to compute the mean square error (as a measure of performance) specifically for each month ahead case, then the data would not be sufficient to obtain accurate estimates. To combat that problem, we have computed the MSE using some kind of moving average. This means that the MSE pertaining to some step ahead h is estimated as the MSE over steps ahead $h-k$ to $h+k$; that is, over a window of size $2k+1$ around h. This enables us to make use of more data, while the expanded data that we use are still relevant, since the inherent MSEs for values at neighboring steps ahead should not differ much. In our situation, we took $k = 1$.

The weights will then be given by:

$$w_1^h = \frac{\sum_{j=-k}^{k} \mathrm{MSE}_{h+j}^{(2)}}{\sum_{j=-k}^{k} \mathrm{MSE}_{h+j}^{(1)} + \sum_{j=-k}^{k} \mathrm{MSE}_{h+j}^{(2)}} \tag{6}$$

$$w_2^h = \frac{\sum_{j=-k}^{k} \mathrm{MSE}_{h+j}^{(1)}}{\sum_{j=-k}^{k} \mathrm{MSE}_{h+j}^{(1)} + \sum_{j=-k}^{k} \mathrm{MSE}_{h+j}^{(2)}}, \tag{7}$$

where $\mathrm{MSE}_l^{(i)}$ is the mean square error for forecast model i in the case of forecasting step l. Note that the MSE is estimated from the evaluation set, which is extracted from the in-sample period for parameter estimation purposes. More details of its structure will be given in the next section. The superscript h in w_i^h reflects the fact that this is the weight for the specific step h case.

4.4. Rank based weighting (RANK)

Aiolfi and Timmermann (2006) proposed a combination method based on setting the weights proportional to the inverse of their performance ranks. This is expected to be more robust and less sensitive to outliers than inverse MSE-based methods. However, the drawback is the discrete nature of the method, as it limits the weights to only a few possible levels.

As with the inverse MSE method, our implementation uses weights that vary with the forecast horizon, likewise using a moving window for MSE measurement.

Let w_i^h be the weight for forecast model i for the step h forecasting case. Then,

$$w_1^h = \frac{R_{2,h}}{R_{1,h} + R_{2,h}} \tag{8}$$

$$w_2^h = \frac{R_{1,h}}{R_{1,h} + R_{2,h}}, \tag{9}$$

where $R_{i,h}$ denotes the performance rank for forecasting model i for horizons $h-k$ to $h+k$, with 1 being the best and 2 the worse. Here again, we used $k = 1$ and also the rank is estimated from the evaluation set. The performance measure used for ranking purposes is the MSE.

4.5. Least squares estimation

A common approach is to estimate the combination weights using linear regression. There are typically three variations, which vary according to the amount of flexibility allowed in the weights. These are given by the following linear regression formulations (Granger & Ramanathan, 1984):

$$x_{t+h} = w_0 + w_1 \hat{x}_{t+h}^{(1)} + w_2 \hat{x}_{t+h}^{(2)} + \epsilon_{t+h} \tag{10}$$

$$x_{t+h} = w_1 \hat{x}_{t+h}^{(1)} + w_2 \hat{x}_{t+h}^{(2)} + \epsilon_{t+h} \tag{11}$$

$$x_{t+h} = w_1 \hat{x}_{t+h}^{(1)} + w_2 \hat{x}_{t+h}^{(2)} + \epsilon_{t+h} \quad \text{s.t. } w_1 + w_2 = 1. \tag{12}$$

The first model, Eq. (10), contains an intercept that can be useful in correcting any possible existing biases. It is beneficial if there is reason to believe that

the individual forecasts could be biased. The second and third models, Eqs. (11) and (12), assume that the underlying forecasting models are unbiased. The third model involves estimating only one variable (e.g. w_1), and is therefore expected to impart a smaller weight estimation error. On the other hand, Timmermann (2006) makes the point that it could lead to an insufficient specification (leading to correlated forecasts and errors). We refer to the three linear regression models in Eqs. (10)–(12) as *LSE1*, *LSE2* and *LSE3*, respectively. For these three models we consider fixed weights for all steps ahead.

4.6. Shrinkage method (SHRINK)

In the shrinkage method, the combination weights are shrunk toward the equal weight solution. One approach, proposed by Diebold and Pauly (1990), is based on a Bayesian analysis. Another approach, proposed by Stock and Watson (2004) is based on shrinking the weights linearly toward the equal weight solution. This is the one we used in our comparison. We used the shrinkage method in conjunction with the INV-MSE combination weights.

Let the weights of the underlying combination method (i.e. INV-MSE) for the h-step-ahead forecast for forecasting model i be w_i^{*h}. Let w_i^h denote the corresponding weight after applying the shrinkage method. It can be evaluated as:

$$w_i^h = \psi w_i^{*h} + (1 - \psi)(1/N) \tag{13}$$

$$\psi = \max(0, 1 - \alpha N/(T - h - N - 1)), \tag{14}$$

where α is the strength of the shrinkage (we took $\alpha = 0.5$), N is the number of forecasting models (in our case $N = 2$), and T is the sample size used to estimate the weights.

4.7. Geometric mean

The arithmetic mean is one conventional way of combining forecasts. However, other types of means could also be valuable for forecast combinations. We tested the geometric mean here. One advantage of the geometric mean is that it always gives a lower value than the arithmetic mean. Thus, in a way it provides some type of shrinkage, which is a desirable property. Another advantage of the geometric mean is that the combination is nonlinear, and thus provides diversity in the selection of forecast combination methods available. The geometric mean has rarely been considered in the forecast combination literature, and has mostly been applied in special situations, such as volatility forecasting (Patton & Sheppard, 2009), combining forecast densities (Faria & Mubwandarikwa, 2008), and grey forecasting (Chen, Ding, & Zhang, 2007).

The combined forecast is given by

$$\hat{z}_{t+h} = \sqrt{\hat{x}_{t+h}^{(1)} \hat{x}_{t+h}^{(2)}}. \tag{15}$$

We abbreviate this method as *GEOM*.

Another more flexible approach is to consider the weighted geometric mean (we call it *GEOM-WTD*), defined by

$$\hat{z}_{t+h} = \left[\hat{x}_{t+h}^{(1)}\right]^w \left[\hat{x}_{t+h}^{(2)}\right]^{1-w}. \tag{16}$$

In this approach, the optimal weight w is obtained by performing a one-dimensional search in [0, 1] and choosing the value that minimizes the MSE. The search is performed on the evaluation set, which is the set extracted from the in-sample period for parameter estimation purposes.

4.8. Harmonic mean

We also considered the harmonic mean as another nonlinear way of combining the forecasts. As in the case of the geometric mean, its value is also lower than the arithmetic mean, thus providing some shrinkage. It has also rarely been investigated in the literature (we found one study, by Chen et al., 2007).

We abbreviate this approach by *HARM*. It is given by:

$$\hat{z}_{t+h} = \frac{2\hat{x}_{t+h}^{(1)} \hat{x}_{t+h}^{(2)}}{\hat{x}_{t+h}^{(1)} + \hat{x}_{t+h}^{(2)}}. \tag{17}$$

We also tested its weighted counterpart (*HARM-WTD*), given by

$$\hat{z}_{t+h} = \frac{\hat{x}_{t+h}^{(1)} \hat{x}_{t+h}^{(2)}}{(1 - w)\hat{x}_{t+h}^{(1)} + w\hat{x}_{t+h}^{(2)}}. \tag{18}$$

Again, the optimal weight w is obtained using a one-dimensional search, performed on the evaluation set, and selecting the value that minimizes the MSE.

4.9. A method based on testing the performance difference

It has been argued in the literature that if there is one dominant forecasting model (performance-wise), then one would be better off simply using this forecasting model on its own (even if this selection is based on the ex ante forecasting performance). In such a situation, combining forecasts would not be very beneficial. At the other extreme, when the individual forecasting models' performances are comparable, or do not differ much, then forecast combination is probably the most beneficial strategy. We propose a novel approach whereby we discriminate between these two situations, and on that basis, decide whether to employ forecast combinations. Specifically, we employ a statistical significance test to test the hypothesis: "The two individual forecasting models give equal performance". If this hypothesis is accepted, then the forecasts are equally weighted. If not, then we select the best forecasting model, rather than combining the forecasts. The statistical test that we use is based on Wilcoxon's signed rank test at the 90% significance level (explained in some detail in the next section). Because this model is based on switching between a forecast combination strategy and a no-combination strategy, we abbreviate it as *SWITCH*.

4.10. Hierarchical forecast combination (HIER)

We propose here a novel approach whereby we take the forecast combination one level higher. Specifically, we consider combining combined forecasts. This means that we identify some of the forecast combination methods and have the overall forecast being a weighted combination of the forecasts obtained by these forecast combination methods. If we were only dealing with linear forecast combination methods, then this approach would be meaningless. It will simply lead to a "composite" forecast combination method. However, it becomes meaningful when nonlinear forecast combination methods are included.

We therefore designed the following method. From among all of the forecast combination methods described previously, we select the two best linear methods and the two best nonlinear methods, and combine these four methods' combined forecasts (using a simple average). This selection is based on the methods' performances in the evaluation set.

5. Simulation experiments

5.1. Simulations on the benchmark data

To test the proposed concept of combining short- and long-term forecasts, we considered some of the standard business-type time series benchmarks. The first benchmark is the M3 time series competition data. This was a major forecasting competition which was organized by the *International Journal of Forecasting* (Makridakis & Hibon, 2000). We have considered all monthly series in the M3 data set that have more than 80 data points. The range of lengths of the time series considered turned out to be between 81 and 126, and the number of time series considered turned out to be 1020.

We also considered another business-type time series benchmark, namely that of the NN3 time series competition.[2] This competition was organized in 2007, and targeted computational-intelligence type forecasting approaches. The data consist of monthly time series. We only considered time series containing more than 80 data points. We ended up with 61 time series, with lengths varying from 115 to 126. Both the M3 and NN3 data sets share many features with the monthly tourism time series that we will consider. They exhibit analogous types of trends and seasonality, and they are also comparable in length. We therefore hope that the conclusions obtained using these benchmarks will have useful implications when considering the tourism time series.

We considered the problem of forecasting 24 months ahead, and therefore held back the last 24 months from each time series as an out-of-sample set (to be forecast in a multi-step-ahead fashion). In addition, for many of the methods we need an extra evaluation data set in order to determine optimal values for the parameters (the parameters are mainly the combination weights). If we had taken 24 more points from the in-sample set, the remaining data might not have been sufficient for an accurate assessment of the performances of the different parameter sets. We therefore used the multiple time origin test (see Tashman, 2000). The time origin is the point from which the

[2] Forecasting Competition for Artificial Neural Networks & Computational Intelligence, 2007 (http://www.neural-forecasting-competition.com/NN3/results.htm).

multi-step ahead forecasts are generated. In the multiple time origin test we shift the time origin a few times, performing the multi-step ahead forecasting each time and computing the error. The average of these errors will then be used as the evaluation criterion. We used a three time origin test, with each forecast period being 24 months long and each time origin separated by one month. Of course, these evaluation data are extracted from the in-sample data (they correspond to the last 26 points). Once this evaluation is complete, we fix the parameter set according to the optimal values determined from the evaluation set, then recalibrate the forecasting model on the entire in-sample period.

The data sets are first preprocessed by checking whether a log transformation is beneficial. To this end, we consider the monthly time series, and perform the 24-step-ahead forecasting on the evaluation data set. We then repeat this exercise on the log-transformed time series, and whichever gives the lower forecasting error (i.e., the original time series or the log-transformed time series) will be used. Of course, the forecasting error is computed after unwinding all preprocessing, including the log transformation (if any). After the log transformation, a seasonality test is performed, in order to determine whether the time series contains a seasonal component or not. The test involves taking the autocorrelation with a lag of 12 months, as well as the partial autocorrelation coefficient with a lag of 12 months (see Box & Jenkins, 1976). Both have to have significant values for the time series to be considered to have a seasonal component. If the test indicates the presence of seasonality, then we use the classical additive decomposition approach (Makridakis, Wheelwright, & Hyndman, 1998, chap. 3) to deseasonalize the data. Once preprocessing has been performed, we consider the deseasonalized series, perform a time aggregation step, in order to create a yearly time series, and apply the forecasting model on both monthly and yearly time series. Once the forecasts have been obtained, we unwind all seasonality and log preprocessing. Then we apply all forecast combination methods considered, to obtain the composite forecast. Please note that the seasonal average will be added back, not only to the forecasts of the monthly time series, but also to the interpolated annual time series. This means that once the annual time series has been forecast, the forecast is interpolated to create month-by-month forecasts, as described in Section 3. Then the monthly seasonal average is added back to these forecasts.

The forecasting model used is a version of Holt's exponential smoothing based on maximum likelihood that was proposed by Andrawis and Atiya (2009). Holt's exponential smoothing model is based on estimating smoothed versions of the level and trend of the time series. Then, the level plus the trend is extrapolated forward to obtain the forecast. We chose the exponential smoothing model because it has been quite successful in forecasting business-type time series (Gardner, 2006), and was ranked near the top in the M3 forecasting competition (Makridakis & Hibon, 2000; for example, one model based on Holt's exponential smoothing is among the top five models for both the annual and monthly data). The version of exponential smoothing that we use in this study is very competitive; as Andrawis and Atiya (2009) pointed out, it outperformed other exponential smoothing approaches. Thus, we have a model which has little room for improvement. This approach uses the single source of error state space formulation put forward by Hyndman, Koehler, Ord, Snyder, and Grose (2002) as a starting point. Both the level and the trend smoothing constants, as well as the initial level and the initial trend, are obtained using the maximum likelihood approach, by converting the problem into a simple two-dimensional search.

We used two error measures. The first was the symmetric mean absolute percentage error, defined as

$$\text{SMAPE} = \frac{1}{MH}\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{2|\hat{z}_{t+h}^{m} - z_{t+h}^{m}|}{\hat{z}_{t+h}^{m} + z_{t+h}^{m}} * 100, \quad (19)$$

where $\hat{z}_{t+h}^{m}$ is the combined h-step-ahead forecast for time series m, z_{t+h}^{m} is the true time series value (for the monthly time series) for series m, H is the forecast horizon (in our case $H = 24$), and M is the number of time series in the benchmark. Note that both error measures are computed after rolling back all of the preprocessing steps performed, such as the deseasonalization and the log transformation.

The second error measure is the mean absolute scaled error (MASE), proposed by Hyndman (2006) and Hyndman and Koehler (2006). For a time series

Table 1
The performance of the single models.

Method	SMAPE (NN3)	SMAPE (M3)	MASE (NN3)	MASE (M3)
ML-Short	19.91	14.49	1.64	3.11
ML-Long	18.09	22.08	1.89	7.10

m, the MASE is defined as:

$$\text{MASE}(m) = \frac{\frac{1}{H}\sum_{h=1}^{H}|\hat{z}^m_{t+h} - z^m_{t+h}|}{\frac{1}{t-1}\sum_{i=2}^{t}|z^m_i - z^m_{i-1}|}. \quad (20)$$

The numerator represents the mean absolute error over the forecast horizon. The denominator is a scaling factor, and represents the mean absolute error for the naïve method, measured on the in-sample set. The reason why the scaling factor is measured on the in-sample set rather than the forecast period is that the in-sample set is typically much larger, and will therefore yield a more reliable factor. The final MASE is the mean of the individual MASE(m)s for all of the time series considered. The MASE has some features which are better than the SMAPE, which has been criticized for the fact that its treatment of positive and negative errors is not symmetric (see Goodwin & Lawton, 1999). However, because of its widespread use, the SMAPE will still be used in this paper.

Table 1 shows the out-of-sample SMAPE and MASE error measures for the M3 and NN3 data sets, for both single models: those based on monthly data and those based on yearly data. Both the SMAPE and the MASE for the latter are computed after converting the forecasts to a monthly frequency, according to the linear interpolation scheme discussed in Section 3. Tables 2 and 3 show the out-of-sample SMAPE and MASE results, respectively, of the different forecast combination methods, applied to the monthly and yearly based models, for both the M3 and NN3 benchmarks.

Note that looking at Tables 1 and 3, we can see that the MASE numbers are higher than 1. The reason for this is as follows. The numerator gives the error for the entire 24-month-ahead period, while the denominator gives the error for the naïve method when performing only one-step-ahead forecasting. For many time series the month-to-month variations are not large. However,

Table 2
The out-of-sample SMAPE values of the different forecast combination methods for the NN3 and M3 data sets. Note that the monthly and yearly based models obtained SMAPE values of respectively 19.91 and 18.09 for the NN3 data set and 14.49 and 22.08 for the M3 data set.

Combination function	NN3	M3
AVG	16.59	15.41
VAR	17.8	14.69
VAR-NO-CORR	16.32	13.8
INV-MSE	16.47	13.41
RANK	16.40	14.10
LSE1	24.82	19.25
LSE2	20.57	15.59
LSE3	16.79	13.62
SHRINK	16.76	14.34
GEOM	17.77	15.62
GEOM-WTD	16.81	13.67
HARM	18.81	15.87
HARM-WTD	16.83	13.72
SWITCH	19.89	14.21
HIER	17.19	14.6

Table 3
The out-of-sample MASE values of the different forecast combination methods for the NN3 and M3 data sets. Note that the monthly and yearly based models obtained MASE values of respectively 1.64 and 1.89 for the NN3 data set and 3.11 and 7.10 for the M3 data set.

Combination function	NN3	M3
AVG	1.59	3.85
VAR	1.56	3.04
VAR-NO-CORR	1.49	2.92
INV-MSE	1.43	2.96
RANK	1.50	3.15
LSE1	2.16	5.07
LSE2	1.72	3.73
LSE3	1.48	2.97
SHRINK	1.42	3.60
GEOM	1.61	4.02
GEOM-WTD	1.48	2.97
HARM	1.62	4.19
HARM-WTD	1.48	2.97
SWITCH	1.55	3.66
HIER	1.59	3.46

Table 4
The Wilcoxon test results for the NN3 data set. The entries are the Wilcoxon statistics W^{+}_{norm}, with the p-values in brackets.

	VAR-NO-CORR	RANK	AVG	INV-MSE	SHRINK	ML-Short	ML-Long
VAR-NO-CORR	0 (0.5)	–	–	–	–	–	–
RANK	0.79 (0.22)	0 (0.5)	–	–	–	–	–
AVG	0.91 (0.18)	0.3 (0.38)	0 (0.5)	–	–	–	–
INV-MSE	0.96 (0.17)	0.74 (0.23)	0.41 (0.34)	0 (0.5)	–	–	–
SHRINK	1.88 (0.03)	1.53 (0.06)	1.42 (0.08)	3 (0)	0 (0.5)	–	–
ML-Short	1.74 (0.04)	1.69 (0.05)	1.34 (0.09)	1.54 (0.06)	0.94 (0.17)	0 (0.5)	–
ML-Long	3.25 (0)	3.35 (0)	3.37 (0)	3.04 (0)	2.64 (0)	0.74 (0.23)	0 (0.5)

because the forecast horizon is large, a minor trend deviation in the forecast will be amplified as the forecast horizon increases, thus leading to a relatively high MASE.

To test whether the observed differences have some statistical significance, we employed a Wilcoxon signed rank test. It is a distribution free test for testing the significance of the difference between the performances of a pair of models. It is based on the ranks of the absolute differences (see Hollander & Wolfe, 1973, for a detailed description).

Due to space considerations, we only apply this test to the SMAPE values (not the MASE values). Specifically, consider two models, A and B. Define:

$$u_i = \text{SMAPE}_B(i) - \text{SMAPE}_A(i), \tag{21}$$

where $\text{SMAPE}_A(i)$ ($\text{SMAPE}_B(i)$) is the SMAPE for model A (model B) for the out-of-sample period of time series i. We then order the absolute values $|u_i|$ and compute the rank of each $|u_i|$ (denoted by $\text{Rank}(|u_i|)$). Then the Wilcoxon signed rank statistic is given by:

$$W^{+} = \sum_{i=1}^{M} I\,(u_i > 0)\text{Rank}(|u_i|), \tag{22}$$

where I is the indicator function, and M is the number of time series. When M is large (about $M > 20$), W^{+} is approximately normal. The normalized statistic will approximately follow a standard normal density, as follows:

$$W^{+}_{\text{norm}} \equiv \frac{W^{+} - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} \sim \mathcal{N}(0, 1). \tag{23}$$

Tables 4 and 5 show the Wilcoxon normalized signed rank statistics for each of the pairings of the five top forecast combination methods, as well as the two single forecasts (i.e. the monthly and yearly based forecasts), for the NN3 and M3 data sets, respectively.

From all of the results presented, one can deduce the following observations:

- For the NN3 time series benchmark, the top five methods (according to the SMAPE) turned out to be VAR-NO-CORR, then RANK, then INV-MSE, then AVG, and then SHRINK. The rankings with respect to the MASE measure are: SHRINK and INV-MSE are almost a tie, then LSE3, GEOM-WTD, HARM-WTD, VAR-NO-CORR and RANK almost tie.
- For the M3 time series benchmark, the top five methods (according to the SMAPE) turned out to be INV-MSE, then LSE3, then GEOM-WTD, then HARM-WTD, and then VAR-NO-CORR. The relative ranking changes a little among these top five methods if the Wilcoxon statistic is considered (with VAR-NO-CORR becoming the number 2 method in that respect). The rankings with respect to the MASE measure are: VAR-NO-CORR, then the INV-MSE, LSE3, GEOM-WTD and HARM-WTD methods are almost a tie.
- For both of the benchmarks, the top five models significantly outperform both of the single forecasts (at the 99% level for the M3 benchmark and at the 90% level for the NN3 benchmark for the top four methods). This attests to the value which is added by combining forecasts with different time aggregations. Note that the confidence level is higher for the M3 benchmark because it has many more time series than the NN3 benchmark (1020 versus 61).
- The worst model for both benchmarks is LSE1. This method seems to be too complex, such that the

Table 5
The Wilcoxon test results for the M3 data set. The entries are the Wilcoxon statistics W^{+}_{norm}, with the p-values in brackets.

	INV-MSE	VAR-NO-CORR	LSE3	GEOM-WTD	HARM-WTD	ML-Short	ML-Long
INV-MSE	0 (0.5)	–	–	–	–	–	–
VAR-NO-CORR	1.52 (0.06)	0 (0.5)	–	–	–	–	–
LSE3	2.01 (0.02)	0.64 (0.26)	0 (0.5)	–	–	–	–
GEOM-WTD	2.15 (0.02)	0.88 (0.19)	0.99 (0.16)	0 (0.5)	–	–	–
HARM-WTD	2.04 (0.02)	1.01 (0.16)	1.63 (0.05)	1.91 (0.03)	0 (0.5)	–	–
ML-Short	6.13 (0)	3.23 (0)	6.94 (0)	7.07 (0)	7.04 (0)	0 (0.5)	–
ML-Long	21.26 (0)	22.07 (0)	20.3 (0)	20.14 (0)	20.12 (0)	16.51 (0)	0 (0.5)

parameter estimation error (rather than flexibility) becomes the dominant issue.

- It is conceivable that the reason for the superiority of INV-MSE is that it assigns combination weights that vary with the forecast horizon (or step ahead). As mentioned, the relative strengths of the monthly and yearly forecasts vary with the horizon.
- The interesting surprise is that GEOM-WTD and HARM-WTD are very competitive methods. These methods have rarely (if ever) been considered in the literature. (In fact, we have never come across HARM-WTD in any published work.)
- We know from the literature that the outperformance of forecast combinations over the best or average of the individual forecasts is typically a general phenomenon and does not depend on the type of the underlying forecasting models. However, the degree to which the combination outperforms the individual models depends on the degree of variation in the accuracies of the underlying models, and their average level of performance. If they are very good forecasting models, they will be harder to beat (using a forecast combination strategy). In our situation, as was mentioned earlier, the maximum likelihood model from Andrawis and Atiya (2009) that we use is very competitive, so there is not much room for improvement.
- The simple average, long known to be a robust forecast combination method, is among the top five for the NN3 benchmark, but not the M3 benchmark. This could be due to the fact that the monthly forecasting generally outperforms the yearly based forecasting by a large margin for the M3 benchmark. Thus, weighting them equally would probably drag down the forecast performance. This is not the case for the NN3 benchmark, where the performances of the monthly forecast and the yearly-based forecast are comparable.

Concerning the last point, one might be tempted to draw the conclusion that if the two individual forecasts have comparable performances, then the simple average would be a good way to go. If there are large differences in performance, then one might be better to opt for a performance-based weighting procedure (such as VAR-NO-CORR, INV-MSE, LSE3, GEOM-WTD, and HARM-WTD). To test, or rather confirm, this hypothesis, we performed the following experiment. From among the 1020 time series in the M3 benchmark, we isolated the 200 time series with the greatest differences in SMAPE performances between the two individual forecasts (i.e., the long term forecast and the short term forecast). Let us call this group DIFF. We also isolated the 200 time series with the lowest absolute difference in SMAPE performance between the two individual forecasts (call this group SIM). We then applied the considered forecast combination methods to each of the DIFF and SIM groups.

Table 6 shows the results. As we can see, for the DIFF group, the worst methods are the equally-weighted ones (AVG, GEOM, and HARM). They are even worse than LSE1, the worst method overall. Also, for the DIFF group, the methods weighted according to performance (specifically, GEOM-WTD, HARM-WTD, and LSE3) turned out to be the best. Concerning the SIM group, we discovered that the equally weighted methods (AVG, GEOM, and HARM) are among the top five methods. However, the performance-weighted methods do not lag far behind and are almost as good. This observed phenomenon agrees with the findings of De Menezes, Bunn, and Taylor (2000). (Please note that the partition into the two groups is based on the evaluation data set, but the

Table 6
The SMAPEs of the different forecast combination methods on the DIFF and SIM groups.

Combination function	DIFF	SIM
AVG	27.32	13.12
VAR	22.6	13.69
VAR-NO-CORR	21.2	13.23
INV-MSE	19.32	12.95
RANK	22.83	13.33
LSE1	24.68	19.99
LSE2	22.25	14.56
LSE3	18.85	13.69
SHRINK	21.87	12.97
GEOM	27.74	12.97
GEOM-WTD	18.99	13.64
HARM	28.2	12.86
HARM-WTD	19.07	13.57
SWITCH	21.51	13.31
HIER	20.6	14.27

Table 7
The 33 source countries.

Germany	Italy	United Kingdom
Russian Federation	France	Israel
Saudi Arabia	Libya A. J.	United States
Palestine	Netherlands	Switzerland
Spain	Belgium	Turkey
Austria	Oman	U.A. Emirates
Denmark	Sweden	Norway
Finland	Greece	Jordan
Syrian A.R.	Lebanon	Tunisia
Canada	Australia	Kuwait
Qatar	Bahrain	Morocco

test results in the table are based on the out-of-sample set.)

5.2. Tourism forecasting

We considered the problem of forecasting the inbound tourism demand for Egypt. Specifically, we consider the monthly tourist numbers originating from 33 major source countries, in addition to the total monthly tourist numbers. (All tourist numbers include Egyptian expatriates.) Table 7 gives the names of these 33 source countries, which are essentially the top 33 source countries for inbound tourism to Egypt. We have 34 time series spanning the period from 1993 to 2007. We obtained these data from the Egyptian Ministry of Tourism. They are therefore very reliable data.

Table 8
The out-of-sample SMAPE and MASE error measures for the different forecast combination methods for the tourism data set. Note that the monthly and yearly based models obtained SMAPEs of 32.76 and 64.88 respectively, and MASEs of 2.28 and 3.85 respectively.

Combination function	SMAPE	MASE
AVG	32.95	2.26
VAR-NO-CORR	33.05	2.10
INV-MSE	28.79	1.82
RANK	29.69	1.96
LSE3	28.44	1.89
SHRINK	30.37	1.90
GEOM-WTD	28.94	1.91
HARM-WTD	29.50	1.94

To be sure, we have also applied Tsay's additive outlier test (Tsay, 1988), and no outliers were detected at the 99% level. We would like to mention that this considers only additive outliers, which are basically outliers in the observed series (not in the data generating process), and are usually due to measurement or recording errors (see Chang, Tiao, & Chen, 1988). The other type of outliers, innovation outliers, affect the underlying innovations process and can have an effect which lasts for more than one observation. As it is hard to handle this type of outlier without making major assumptions about the data generating process, this type is not considered here. The forecast horizon is 24 months, and we held back the last 24 months of data as an out-of-sample period. Similar to the benchmark data cases, we used a three-time origin test set as an evaluation data set. We used the last 26 months of the in-sample period for this purpose. We also preprocessed the time series by taking a log transformation, followed by a deseasonalization step. No seasonality test was needed, as we knew beforehand that all of the time series were seasonal. For tourism time series, one generally needs to be careful in handling moving calendar effects. For example, Easter could happen in either March or April. However, we tested the time series and found a weak relationship between the position of Easter and the position of the spring peak of tourism arrivals, and thus no correction was needed in our case. In addition, no log test was needed, since, by inspecting all of the time series, we found an exponential-looking growth curve. We therefore applied the log transformation to all time series.

Table 9
The Wilcoxon test results for the tourism data set. The entries are the Wilcoxon statistics W^{+}_{norm}, with the p-values in brackets.

	LSE3	GEOM-WT	HARM-WT	INV-MSE	SHRINK	RANK	VAR-NO	AVG	ML-Short
LSE3	0 (0.5)	–	–	–	–	–	–	–	–
GEOM-WT	1.74 (0.04)	0 (0.5)	–	–	–	–	–	–	–
HARM-WT	1.36 (0.09)	1.41 (0.08)	0 (0.5)	–	–	–	–	–	–
INV-MSE	0.59 (0.28)	0.11 (0.46)	−0.37 (0.36)	0 (0.5)	–	–	–	–	–
SHRINK	1.17 (0.12)	0.71 (0.24)	0.15 (0.44)	2.28 (0.01)	0 (0.5)	–	–	–	–
RANK	1.72 (0.04)	1.15 (0.12)	0.79 (0.21)	0.73 (0.23)	−0.68 (0.25)	0 (0.5)	–	–	–
VAR-NO	3.15 (0)	2.8 (0)	2.56 (0.01)	2.95 (0)	1.27 (0.1)	3.03 (0)	0 (0.5)	–	–
AVG	2.93 (0)	2.4 (0.01)	1.84 (0.03)	2.4 (0.01)	1.17 (0.12)	3.38 (0)	0.3 (0.38)	0 (0.5)	–
ML-Short	1.85 (0.03)	1.6 (0.05)	1.32 (0.09)	1.29 (0.1)	1.21 (0.11)	0.91 (0.18)	−0.61 (0.27)	−0.11 (0.46)	0 (0.5)

We attempted to make use of the lessons learned from the experiments performed on the benchmark data. For example, there was no point in testing the inferior combination methods (such as LSE1, VAR, and some others). We limited ourselves to the top five models from the NN3 experiment and the top five models of the M3 experiment (according to the SMAPE). We ended up with the following eight models: INV-MSE, RANK, VAR-NO-CORR, AVG, SHRINK, LSE3, GEOM-WTD, and HARM-WTD. Table 8 shows the out-of-sample SMAPE and MASE numbers for these eight methods. Note that the individual models, i.e. the monthly and yearly based models, yielded SMAPEs of 32.76 and 64.88 respectively, and MASEs of 2.28 and 3.85 respectively. One can see from the table that six of the eight methods led to an out-of-sample accuracy better than that of the best of the individual models (with respect to the SMAPE). On the other hand, *all* of the methods led to an improvement over both individual forecasting models (with respect to the MASE). In spite of the large difference in accuracy between the two individual models, the weak model still had something to contribute (in the forecast combination). The top combination methods turned out to be LSE3 and INV-MSE (depending on whether we consider the SMAPE or the MASE). Following this comes GEOM-WTD. Table 9 shows the Wilcoxon signed rank test for the pairwise test of significance between the eight methods and the short-term individual model (applied to the SMAPE numbers). (For brevity, we have not included the long-term individual model, because it was considerably worse than the short-term model.) One can see that the top four models (LSE3, INV-MSE, GEOM-WTD, and HARM-WTD) outperformed the best single model (that is, the short-term model) at the 90% significance level. It is interesting to note that the top four models are among the top 5 models for the M3 benchmark. This may be due to the fact that the two data sets share the common feature that there is generally a large difference in SMAPEs between the short- and long-term models. Thus, this ranking may be partly dictated by this characteristic.

6. Conclusions

In this paper we have investigated the idea of combining forecasts using different time aggregations. The rationale for this idea is that different time scales will capture different dynamics, and therefore this will increase the diversity of the forecasts obtained. The simulation results indicated improvements in accuracy relative to the underlying forecasting models. The other goal of this paper has been to develop a forecasting model for inbound tourism demand for Egypt, using the developed short-term/long-term forecast combination approach. The simulation experiments also showed the extent to which this approach outperformed the individual models. We therefore believe that this is a promising direction, and that it would be beneficial to explore it further, perhaps exploring the combination of more than two time aggregations (for example, monthly, quarterly and yearly).

Acknowledgements

The authors would like to thank the Egyptian Ministry of Tourism for supplying the data and for their assistance. This work is part of the *Cross-Industry Data Mining* research project within the Egyptian Data Mining and Computer Modeling Center of Excellence.

References

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*, 31–53.

Andrawis, R. R., & Atiya, A. F. (2009). A new Bayesian formulation for Holt's exponential smoothing. *Journal of Forecasting*, *28*, 218–234.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.

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

Athanasopoulos, G., & Hyndman, R. J. (2008). Modelling and forecasting Australian domestic tourism. *Tourism Management*, *29*, 19–31.

Bermudez, J. D., Segura, J. V., & Vercher, E. (2007). Holt-Winters forecasting: an alternative formulation applied to UK air passenger data. *Journal of Applied Statistics*, *34*, 1075–1090.

Box, G., & Jenkins, G. (1976). *Time series analysis, forecasting and control*. Holden-Day Inc.

Casals, J., Jerez, M., & Sotoca, S. (2009). Modelling and forecasting time series sampled at different frequencies. *Journal of Forecasting*, *28*, 316–342.

Chang, C.-L., Sriboonchitta, S., & Wiboonpongse, A. (2009). Modelling and forecasting tourism from East Asia to Thailand under temporal and spatial aggregation. *Mathematics and Computers in Simulation*, *79*, 1730–1744.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, *30*, 193–204.

Chen, H., Ding, K., & Zhang, J. (2007). Properties of weighted geometric means combination forecasting model based on degree of logarithm grey incidence. In *Proceedings of the IEEE international conference on grey systems and intelligent services* (pp. 673–677).

Cholette, P. A. (1982). Prior information and ARIMA forecasting. *Journal of Forecasting*, *1*, 375–384.

Chu, F.-L. (1998). Forecasting tourism: a combined approach. *Tourism Management*, *19*, 515–520.

Chu, F.-L. (2008). A fractionally integrated autoregressive moving average approach to forecasting tourism demand. *Tourism Management*, *29*, 79–88.

Chu, F.-L. (2009). Forecasting tourism demand with ARMA-based methods. *Tourism Management*, *30*, 740–751.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.

De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*, 190–204.

Diebold, F. X., & Pauly, P. (1990). The use of prior information in forecast combinations. *International Journal of Forecasting*, *6*, 503–508.

du Preez, J., & Witt, S. F. (2003). Univariate versus multivariate time series forecasting: an application to international tourism demand. *International Journal of Forecasting*, *19*, 435–451.

Engle, R. F., Granger, C. W. J., & Hallman, J. J. (1989). Merging short- and long-run forecasts: an application of seasonal co-integration to monthly electricity sales forecasting. *Journal of Econometrics*, *40*, 45–62.

Faria, A. E., & Mubwandarikwa, E. (2008). The geometric combination of Bayesian forecasting models. *Journal of Forecasting*, *27*, 519–535.

Frechtling, D. C. (1996). *Practical tourism forecasting*. Elsevier Publ.

Fritz, R. G., Brandon, C., & Xander, J. (1984). Combining time-series and econometric forecast of tourism activity. *Annals of Tourism Research*, *11*, 219–229.

Gardner, E. S. (2006). Exponential smoothing: the state of the art—part II. *International Journal of Forecasting*, *22*, 637–666.

Gil-Alana, A. L., Cunado, J., & de Gracia, F. P. (2008). Tourism in the Canary Islands: forecasting using several seasonal time series models. *Journal of Forecasting*, *27*, 621–636.

Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management*, *23*, 499–510.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of symmetric MAPE. *International Journal of Forecasting*, *15*, 405–408.

Granger, C. W. J. (1993). Implications of seeing economic variables through an aggregation window. *Ricerche Economiche*, *47*, 269–279.

Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3*, 197–204.

Greene, M. N., Howrey, E. P., & Hymans, S. W. (1986). The use of outside information in econometric forecasting. In E. Kuh, & D. A. Belsley (Eds.), *Model reliability*. Cambridge, MA: MIT Press.

Hilaly, H., & El-Shishiny, H. (2008). Recent advances in econometric modeling and forecasting techniques for tourism demand prediction. In *Proceedings of Eurochrie Conference*.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. Wiley.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight*, *4*, 43–46.

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

Hyndman, R., Koehler, A., Ord, K., Snyder, R., & Grose, S. (2002). A state space formulation for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439–454.

Kamel, N., Atiya, A. F., El Gayar, N., & El-Shishiny, H. (2008). Tourism demand forecasting using machine learning methods. *ICGST International Journal on Artificial Intelligence and Machine Learning*, *8*, 1–7.

Kon, S. C., & Turner, L. W. (2005). Neural network forecasting of tourism demand. *Tourism Economics*, *11*, 301–328.

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

Lim, C., & McAleer, M. (2001). Forecasting tourist arrivals. *Annals of Tourism Research*, *28*, 965–977.

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

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: methods and applications* (3rd ed.). Wiley.

Medeiros, M. C., McAleer, M., Slottje, D., Ramos, V., & Rey-Maquieira, J. (2008). An alternative approach to estimating demand: neural network regression with conditional volatility for high frequency air passenger arrivals. *Journal of Econometrics*, *147*, 372–383.

Oh, C. O., & Morzuch, B. J. (2005). Evaluating time-series models to forecast the demand for tourism in Singapore: comparing within-sample and post-sample results. *Journal of Travel Research*, *43*, 404–413.

Patton, A. J., & Sheppard, K. (2009). Optimal combinations of realised volatility estimators. *International Journal of Forecasting*, *25*, 218–238.

Petropoulos, C., Nikolopoulos, K., Patelis, A., & Assimakopoulos, V. (2005). A technical analysis approach to tourism demand forecasting. *Applied Economics Letters*, *12*, 327–333.

Riedel, S., & Gabrys, B. (2007). Combination of multi level forecasts. *Journal of VLSI Signal Processing Systems*, *49*, 265–280.

Shen, S., Li, G., & Song, H. (2008). An assessment of combining tourism demand forecasts over different time horizons. *Journal of Travel Research*, *47*, 197–207.

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. (2004). Forecasting international tourist flows to Macau. *Tourism Management*, *27*, 214–224.

Song, H., Witt, S. F., Wong, K. F., & Wu, D. C. (2009). An empirical study of forecast combination in tourism. *Journal of Hospitality and Tourism Research*, *33*, 3–29.

Stock, J. H., & Watson, M. W. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R. F. Engle, & H. White (Eds.), *Cointegration, causality, and forecasting: a festschrift in honour of Clive W.J. Granger*. Cambridge, UK: Cambridge University Press.

Stock, J. H., & Watson, M. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.

Tashman, L. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*, 437–450.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Elsevier Publ.

Trabelsi, A., & Hillmer, S. C. (1989). A benchmarking approach to forecast combination. *Journal of Business and Economic Statistics*, *7*, 353–362.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, *7*, 1–20.

Witt, S. F., Newbould, G. D., & Watkins, A. J. (1992). Forecasting domestic tourism demand: application to Las Vegas arrivals data. *Journal of Travel Research*, *31*, 36–41.

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

Wong, K., & Song, H. (2002). *Tourism forecasting and marketing*. New York: The Haworth Hospitality Press.

Wong, K. F., Song, H., & Chon, K. (2006). Bayesian models for tourism demand forecasting. *Tourism Management*, *27*, 773–780.

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

Zaki, A. (2008). An econometric model forecasting Egypt's aggregate international tourism demand and revenues. *Tourism and Hospitality Planning and Development*, *5*, 215–232.