sustainability

MDPI

Article

Weekly Hotel Occupancy Forecasting of a Tourism Destination

Muzi Zhang [1,2,*], Junyi Li [1,3,*], Bing Pan [4] and Gaojun Zhang [5]

[1] School of Geography and Tourism, Shaanxi Normal University, Xi'an 710062, China
[2] Qionglai Prefectural Bureau of Culture, Sport, Radio and TV, Press and Publication, and Tourism, Chengdu 611530, China
[3] Shaanxi Key Laboratory of Tourism Informatics, Xi'an 710062, China
[4] Department of Recreation, Park, and Tourism Management, College of Health and Human Development, Penn State University, University Park, PA 16801, USA; bingpan@psu.edu
[5] Shenzhen Tourism College, Jinan University, Shenzhen 518055, China; zhang_gj@sz.jnu.edu.cn
* Correspondence: muzik@snnu.edu.cn (M.Z.); lijunyi9@snnu.edu.cn (J.L.); Tel.: +86-18202806740

Received: 10 October 2018; Accepted: 19 November 2018; Published: 22 November 2018

Abstract: The accurate forecasting of tourism demand is complicated by the dynamic tourism marketplace and its intricate causal relationships with economic factors. In order to enhance forecasting accuracy, we present a modified ensemble empirical mode decomposition (EEMD)–autoregressive integrated moving average (ARIMA) model, which dissects a time series into three intrinsic model functions (IMFs): high-frequency fluctuation, low-frequency fluctuation, and a trend; these three signals were then modeled using ARIMA methods. We used weekly hotel occupancy data from Charleston, South Carolina, USA as an empirical test case. The results showed that for medium-term forecasting (26 weeks) of hotel occupancy of a tourism destination, the modified EEMD–ARIMA model provides more accurate forecasting results with smaller standard deviations than the EEMD–ARIMA model, but further research is needed for validation.

Keywords: time series; ensemble empirical model decomposition; demand forecasting; signal decomposition; spectral analysis

1. Introduction

Tourism significantly contributes to the world economy. However, the industry is often influenced by many economic factors, creating volatility and causing difficulty in tourism forecasting. Two of the most popular forecasted variables are tourist arrivals and expenditures, which are crucial for tourism businesses and organizations to meet the needs of tourists, allocate limited resources, and formulate appropriate market strategies and policies [1]. In addition, hotel occupancy reflects one of the most important sectors in the tourism industry. At a business level, hotel occupancy forecasting helps individual hotels in revenue management practices and decision making in marketing. As a result, hotels can fully utilize their existing inventories with less waste and thus contribute to the sustainable development of a property; at the macro level, accurate forecasting is beneficial to strategic policy making and the sustainable development of tourism destinations by allocating appropriate resources for supporting hospitality operations.

The main time-series methods for tourism forecasting include an autoregressive integrated moving average (ARIMA) model, autoregressive conditional heteroscedasticity model, generalized autoregressive conditional heteroscedasticity model, and stochastic volatility model. These methods have been widely used in various economic and financial studies due to their speed, convenience, practicality, and relative accuracy, particularly for stable time-series data [2].

However, past studies have shown that no one forecasting method is superior to others under all scenarios. Some studies have shown that a combination of various methods gives more accurate results than a single forecasting method, such as a combination of qualitative and quantitative, and linear and nonlinear methods, or a combination of several elements, such as tourism cycles, seasonality, social events, and risks [3].

Accurate tourism demand forecasting usually relies on consistent patterns in the historical data. However, time series of tourism demand are usually nonlinear and nonstationary, and are affected by random factors that generate a considerable amount of noise due to market dynamics, which makes accurate prediction difficult.

Empirical mode decomposition (EMD) is a novel method of adaptive time series signal analysis [4]. Based on Fourier transformation, this method is considered the most critical breakthrough in linear and spectrum analysis since 2000 [5]. Ensemble empirical mode decomposition (EEMD) was an improvement to the empirical mode decomposition (EMD) that emerged in 2005 [6]. It is believed that combining the EEMD method and ARIMA model helps remove interference signals from the original tourism time series, resulting in a more accurate trend for better forecasting [7,8]. The present study introduced a modified EEMD–ARIMA model, and used it to generate predictions of hotel occupancy. This technique eliminates problems such as nonlinearity and instability that cannot be resolved in traditional ARIMA models. It also employs a self-adaptive time-frequency analysis and does not require a priori assumptions, and thus can be widely applied.

2. Literature Review

The earliest tourism demand forecasting research dates from the 1960s. Beginning in the 1990s, the rapid development of the tourism industry led to an expansion of empirical research on the topic, and a large amount of studies subsequently emerged.

2.1. Time Series and Econometric Methods

In the past three decades, time-series and econometric models have been the two main methods in tourism forecasting. Song and Li reviewed 121 studies of tourism forecasting, among which 72 studies used time-series models to predict tourism demand. Other studies have validated that the ARIMA model is superior to other models [9,10]. In more current studies, ARIMA models are proved to be adaptable to new types of data series, such as Google search engine volume data [11,12]. Pan et al. used Google search query volume data to study hotel demand in Charleston, South Carolina, USA. They compared ARMA family models with autoregressive moving-average models, including some VAR (vector autoregressive) models, with search engine data as an explanatory variable (ARMAX). The results indicated that all three ARMAX models outperformed their ARMA counterparts [11].

A notable feature of econometric models is that they analyze not only single time-series, but also the relationships between the time series and other independent and explanatory variables, such as variables in tourism, markets, economies, or policies. Researchers surveyed the econometric models employed in 84 empirical tourism studies, confirming that from the 1960s to the 1980s, advanced methods such as VAR, autoregressive distributive lag, time-varying parameter (TVP), an almost ideal demand system (AIDS), and cointegration and error correction models (CI/ECM) populated econometric methods. Their results showed that TVP models provided relatively higher accuracy than the alternative models. As a result, researchers further developed this model [13]. Gunter and Onder [14] compared seven different models in forecasting the tourism demand of Paris from its major source markets. Naïve-1 model served as a benchmark. All seven models, including error-correction formulation of autoregressive distributed lag model (EC-ADLM), classical and Bayesian VAR, TVP, ARIMA, and error, trend, seasonality (ETS), significantly outperformed the benchmark model in all cases of source markets and forecast length.

2.2. Combined Methods

Researchers constantly work on the comparison of models; however, no single model is proven superior to others in all cases. Some studies have shown that combined models are more accurate than single ones [3].

Makridakis et al. conducted a series of studies that aimed to compare the forecasting accuracy of different methods by examining a large amount of samples. The termed Makridaskis Competitions (M-competition) testified that accurate forecasting depended on a good match between method and the type of time scale, the type of series (macro, micro, etc.), and the time horizon of forecasting. It also showed that the combination of a few methods helped improve the overall forecasting accuracy. The following studies (known as M2-competition and M3-competition) selected different samples on scales and types, and adopted diverse methods [15–17].

Peng, Song, and Crouch found that various combinations of factors influence forecasting outcomes. They performed a meta-analysis of tourism demand forecasting and offered practical suggestions. The reviewed 65 studies from 1980 to 2011, and showed that the accuracy of forecasting models is influenced by tourist origins, destinations, duration of stay, modeling methods, data frequency, the types of demand variables and their measures, and sample sizes. However, they revealed that a combination of methods usually showed superior accuracy compared with single ones [18].

Bangwayo-Skeete and Skeete incorporated Google search query volume in autoregressive mixed-data sampling (AR-MIDAS) models, and demonstrated its superiority in forecasting the amount of tourists from three main source countries to five Caribbean destinations [12]. Li et al. proposed a generalized dynamic factor model (GDFM) with search engine data to forecast tourist demand in Beijing. By using the common components of search trends data to construct a better index, they compared the new index with an ARIMA model, and a model with an index created by principal component analysis. The results showed that the combination of a composite search index and GDFM resulted in more accurate results [19].

Linear and nonlinear methods were combined by Chen to forecast outbound tourism demand. The three linear forecasting methods were naïve, exponential smoothing, and ARIMA models. These were combined with two nonlinear methods, back-propagation neural networks (BPNNs) and support vector regression (SVR); the directional change accuracy (DCA) test was used to forecast turning points. The study produced forecasts using the three linear methods, in combination with BPNNs, SVR, and the DCA test. The result revealed that combined methods outperform single linear methods [3].

2.3. Other Methods

Besides time-series and normal econometric models, researchers have been trying to apply or combine more methods from other fields.

By the late 1990s, the neural network method had become widely adopted in scientific and business fields. Before a study by Law [20], few studies had used this method to forecast hotel demand. Law showed that the neural network model outperformed multiple regression and naïve extrapolation.

Simulation methods have been used by Zakhary et al. By making accurate estimations of the algorithm's parameter, the Monte Carlo simulation method can simulate the actual physical processes that are related to hotel demand and occupancy. Through the application of this method, hotel reservation was simulated forward in time, and these future Monte Carlo paths yielded forecast densities. This method attained superior outcomes compared to other methods [21].

Pai, Hung, and Lin employed a novel method to forecast tourism arrivals in Hong Kong and Taiwan from 1969 to 2010. They applied the fuzzy c-means clustering algorithm combined with logarithm least-squares support vector regression (LLS-SVR), and used genetic algorithms (GA) to select the parameters. They compared this with the traditional ARIMA method, and revealed that their novel method was superior to traditional ones [22].

Hassani et al. applied Singular-Spectrum Analysis (SSA) to forecast tourist arrivals to the United States (U.S.). By comparing to ARIMA, exponential smoothing and neural networks, SSA was superior to alternative models over both short and long periods [23].

Caicedo-Torres and Payares surveyed several machine learning models, such as ridge regression and kernel ridge regression, to forecast the daily occupancy rates for a hotel. They discussed the approaches related to dataset construction and model validation, and found that machine learning models are good tools to forecast daily hotel occupancy [24]. Researchers also succeeded in increasing daily hotel occupancy forecasting accuracy by introducing simulated scenario analysis. A competitive set's aggregated forecast was set as the input to the process; therefore, the individual forecast absorbed external factors in the market, and thus improved the accuracy [25].

EMD is a new self-adaptive algorithm that can decompose a series of data. Few studies have focused on adopting EMD in tourism research. EMD and BPNNs were applied to examine tourism demand forecasts in Taiwan. Chen, Lai, and Ye reviewed samples of tourists from Japan, Hong Kong, and Macao in 1971, and then contrasted EMD–BPNN analysis with BPNNs and an ARIMA model alone. The EMD–BPNN method attained more accurate outcomes than did the BPNN or ARIMA methods [26]. Zhang et al. adopted EEMD–ARIMA for daily hotel occupancy forecasting for an individual hotel. This research validated that the EEMD–ARIMA model had better forecasting ability than the ARIMA model, especially in the short term [7].

In conclusion, researchers have adopted time series, econometric models, and artificial intelligent methods to forecast tourism demand. A combination of methods has been proven superior in some cases. As a novel data decomposition method, EMD has been combined with neural network as well as ARIMA, and proven superior in daily hotel occupancy forecasting. However, the existing EEMD–ARIMA model showed a satisfactory result only in a very short period of a few days. It is unclear in its applicability in other time series in different scales. Therefore, the present research verified the existing EEMD–ARIMA model, and introduced a modified EEMD–ARIMA model to achieve accurate medium and long-term forecasting for hotel occupancy.

3. Methodology

ARIMA models first turn unstable time series into stable time series by d differences (Equation (1)), and then regress dependent values on lag values and the random error's present and lag values. Depending on the stability of the original time series and its differencing methods, an ARIMA model can be expressed as an AR model, MA model, or ARMA model.

Time series of tourism demand are often nonlinear and nonstationary with white noises, and can cause difficulties in forecasting tourism demand. Therefore, the white noises were reduced by EEMD, so that irregular fluctuations could be transferred to more mild and understandable series. Thus, EEMD is applied to abstract main trend and stable cycles from a signal, and seems to be an ideal candidate.

Unlike existing EEMD research in tourism, the previous empirical study often adopted nonstationary and nonlinear time series data, while hotel occupancy data often reflects local tourism seasonality, thus relatively stable. As this kind of data were decomposed by EEMD, it would generate many high-frequency signals. When these signals were forecast by ARIMA models separately, the summation of respective errors would further accumulate, resulting in misleading predictions. Therefore, we introduce a modified EEMD–ARIMA model by partially combining signals decomposed by EEMD, and subsequently decreased the errors.

This research aimed at validating a modified EEMD–ARIMA method for accurately forecasting tourism demand in terms of hotel occupancy.

3.1. Autoregressive and Moving Average Model (ARMA)

In general, an ARIMA (p,d,q) model can be expressed as:

$$\Delta^d \ln y_t = \mu + \sum_{i=1}^{p} \phi_i \Delta^d \ln y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{1}$$

ARMA and AR models are a more general form of ARIMA (p,d,q). These models are standard time series models [10].

3.2. Empirical Mode Decomposition (EMD)

In the EMD method, Wu and Huang used intrinsic model function (IMF) to convert time signals into narrowband frequencies [6]. They believe that all signals are composed of IMFs in different frequencies, and the compounded IMFs make up a natural signal. The goal of the EMD method is to decompose IMFs from signals with Hilbert transformation. The following describes the specific steps of the EMD method:

(1) Plot the original data signal x(t);

(2) Select all the maxima points from the original data, and connect them to compose an upper envelope $\mathrm{e_{max}(t)}$ with spline interpolation; then connect all of the minima points the same way to form the lower envelope $\mathrm{e_{min}(t)}$;

(3) Calculate the mean $\mathrm{a_1(t)}$ between the upper and the lower envelopes;

$$a_1(t) = [e_{\max}(t) + e_{\min}(t)]/2 \tag{2}$$

(4) Calculate a new data column $\mathrm{x_1(t)}$ by subtracting $\mathrm{a_1(t)}$ from x(t);

$$x(t) - a_1(t) = x_1(t) \tag{3}$$

(5) Then, $\mathrm{x_1(t)}$ is deemed to be the first IMF (written as $\mathrm{c_1(t)}$); and steps 1 to 4 should be repeated as many times as $\mathrm{x_n(t)}$ meets the stopping criterion of IMFs [6];

(6) The residue:

$$r_1(t) = x(t) - c_1(t) \tag{4}$$

is a new dataset excluding the high-frequency signal and subjected to the same sifting process as described before for the next IMF from $\mathrm{r_1(t)}$. Finally, the procedure continues until the residue r(t) becomes a constant or a monotonic function, and no more IMFs can be extracted. At the end of this sifting procedure, the original data signal x(t) can be expressed as the sum of IMFs and the residue of x(t) as:

$$\mathrm{x(t)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{c_i(t)} + \mathrm{r(t)} \tag{5}$$

where n is the number of IMFs, r(t) is the final residue, and $\mathrm{c_i(t)}$ are almost orthogonal to each other, and all their means are zero.

3.3. Ensemble Empirical Mode Decomposition (EEMD)

Research has shown that EMD has a mode mixing problem because of the noise in the signal [6]. Mode mixing is defined as either a single IMF consisting of different time scales, or a component of similar scales distributed in different IMFs [27]. This process makes the waveform of two adjacent IMFs mixed together, generating difficulty in implementing feature extraction. The EEMD, which uses noise-assisted data analysis (NADA), as proposed by Wu and Huang, overcomes this problem. In this method, the added white noise changes the distribution feature of extremum points in low-frequency composition, facilitating an average separation of extremum points in frequency scales, and mitigates the mode mixing problem [6]. Based on previous EMD method, the procedure of EEMD can be described as follows [28]:

(1) Add white noise series to the targeted signal several times. The white noise series has a mean of zero and a constant standard deviation:

$$x_i(t) = x(t) + n_i(t) \tag{6}$$

where $x_i(t)$ represents the signal when white noise is added at time i; and $n_i(t)$ refers to the added white noise at time i.

(2) Decompose the new series with the added white noise by the EMD method into IMFs;

(3) Repeat steps (1) and (2), but add different white noise series each time;

(4) Calculate the ensemble means of the corresponding IMFs above, and then obtain the final IMFs by EEMD decompositions.

$$c_j(t) = 1/N\sum_{i=1}^{n} c_{ij}(t) \tag{7}$$

where N denotes the times of the added white noise series, and $c_j(t)$ represents the number of IMFs decomposed through EEMD at number j.

In addition, two common measurements were used to examine the accuracy of the models: the mean absolute percentage error (MAPE) and root mean square error (RMSE). They are respectively expressed as:

$$MAPE = \frac{1}{m}\sum_{t=1}^{m}\left[\frac{|\hat{y}_t - y_t|}{y_t}\right] \tag{8}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(\hat{y}_t - y_t)^2} \tag{9}$$

4. Data Description

This study selected Charleston, South Carolina, USA as an empirical case due to the authors' easy access to the data sources. Charleston is located in the southeastern U.S.; approximately five million tourists visit this port city and its resorts every year [29].

Smith Travel Research, Inc. (STR, Hendersonville, TN, USA) is a company that tracks supply and demand data for the hotel industry and provides market share analysis for all of the major hotel chains and brands. It covers 110 hotels from a total of around 190 in the area, accounting for approximately 60% of the market. Using these data, STR computes average hotel occupancy, which is regarded as a representative of the overall hotel occupancy in Charleston. Therefore, these data were considered suitable for use in the present study [30].

The time scale in tourism demand research studies is usually daily, weekly, monthly, quarterly, or annual data, among which annual data is the most common. For a tourism destination, the smaller the time scale data is, the more helpful the results are for managers to see market dynamics, and for policy makers to make decisions [9]. However, it is rare to see medium-small scale (weekly) data in hotel occupancy research due to the unsteadiness and indeterminacy in a short time scale. Therefore, to enrich research in medium to small-scales, weekly hotel occupancy data of a tourism destination is tested in the present study.

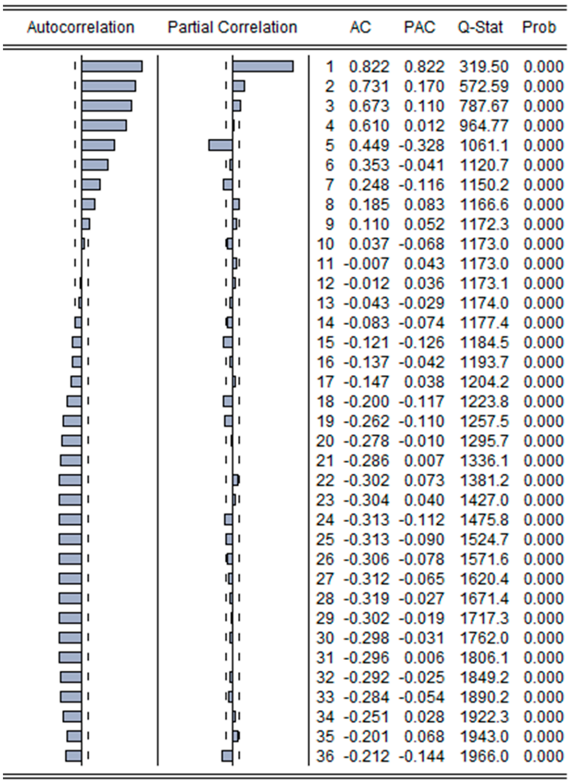
The Charleston area is divided into four sub-areas: North Charleston, East Cooper, West Ashley, and the Peninsula. The total hotel occupancy in the Charleston area was first used as the main empirical case for the proposed model; data series from the four sub-areas were then used to verify the model's reliability.

5. Empirical Results

In this section, we first conducted one of the most popular methods, ARIMA, to forecast the hotel occupancy in the whole Charleston area as a benchmark model; second, we adopted EEMD method to compare its accuracy with the benchmark model; third, we combined EEMD method with ARIMA, and tried to reach a better forecasting accuracy. When the third method failed, we adopted a modified EEMD–ARIMA model and achieved the best forecasting results.

5.1. ARIMA and SARIMA Models

Since ARIMA models require a stable series to produce accurate results, an augmented Dickey–Fuller (ADF) test was used to determine its stability. The results demonstrated that no unit root was present in the time series, implying that computing difference was not necessary, and the value of d in ARIMA (p,d,q) was set to zero. According to Box and Jenkins, p and q are usually confirmed by autocorrelation function (ACF) and partial autocorrelation function (PACF) testing. The ACF and PACF results reveal that if a time series conforms to the AR (p) model, then PACF is truncated by p steps; if a time series conforms to the MA (q) model, then ACF is truncated by q steps [31]. The PACF of hotel occupancy was truncated and converged with the ACF in the confidence interval by three or five steps (Figure 1). This was considered sufficient evidence to establish an AR (3) or AR (5) model. However, the PACF and ACF methods are not always applicable with mixed ARMA models. Hence, the auto.arima forecast function in R was used to fit the appropriate p and q values automatically [32]. The function does not only compare among ARIMA models, but ARMA models as well. It can also compare different orders of the models, which results in a more accurate identification of a optimal model than by choosing the ACF and PACF artificially. As the data is stable, seasonality is included when modeling, and the number of observations per year is 52. The final model was a special case of ARIMA, a seasonal ARIMA (SARIMA) model, and ARIMA(2,0,4)(0,0,2)$_{52}$ was selected.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.822	0.822	319.50	0.000
		2	0.731	0.170	572.59	0.000
		3	0.673	0.110	787.67	0.000
		4	0.610	0.012	964.77	0.000
		5	0.449	-0.328	1061.1	0.000
		6	0.353	-0.041	1120.7	0.000
		7	0.248	-0.116	1150.2	0.000
		8	0.185	0.083	1166.6	0.000
		9	0.110	0.052	1172.3	0.000
		10	0.037	-0.068	1173.0	0.000
		11	-0.007	0.043	1173.0	0.000
		12	-0.012	0.036	1173.1	0.000
		13	-0.043	-0.029	1174.0	0.000
		14	-0.083	-0.074	1177.4	0.000
		15	-0.121	-0.126	1184.5	0.000
		16	-0.137	-0.042	1193.7	0.000
		17	-0.147	0.038	1204.2	0.000
		18	-0.200	-0.117	1223.8	0.000
		19	-0.262	-0.110	1257.5	0.000
		20	-0.278	-0.010	1295.7	0.000
		21	-0.286	0.007	1336.1	0.000
		22	-0.302	0.073	1381.2	0.000
		23	-0.304	0.040	1427.0	0.000
		24	-0.313	-0.112	1475.8	0.000
		25	-0.313	-0.090	1524.7	0.000
		26	-0.306	-0.078	1571.6	0.000
		27	-0.312	-0.065	1620.4	0.000
		28	-0.319	-0.027	1671.4	0.000
		29	-0.302	-0.019	1717.3	0.000
		30	-0.298	-0.031	1762.0	0.000
		31	-0.296	0.006	1806.1	0.000
		32	-0.292	-0.025	1849.2	0.000
		33	-0.284	-0.054	1890.2	0.000
		34	-0.251	0.028	1922.3	0.000
		35	-0.201	0.068	1943.0	0.000
		36	-0.212	-0.144	1966.0	0.000

Figure 1. autocorrelation function (ACF) and partial autocorrelation function (PACF) results of Charleston hotel occupancy.

Several fit measures were considered applicable, such as Akaike's Information Criterion (AIC), Bayes Information Criterion (BIC), and Schwarz Criterion (SC). AIC, SC, and R^2 were chosen to select the best model. In a comparison of those measures among AR (3), AR (5), and ARIMA (2,0,4) $(0,0,2)_{52}$, ARIMA (2,0,4) $(0,0,2)_{52}$ was found to be the most effective model (Table 1).

Table 1. Comparison of model's robustness index. AIC: Akaike's Information Criterion, AR: autoregressive, ARIMA: autoregressive integrated moving average, SC: Schwarz Criterion.

	AR (3)	**ARIMA (2,0,4) $(0,0,2)_{52}$**	**AR (5)**
R^2	0.698	0.737	0.732
AIC	−2.462	−2.586	−2.573
SC	−2.427	−2.524	−2.520

Finally, the ARIMA (2,0,4) $(0,0,2)_{52}$ model was diagnosed in R. As shown in Figure 2, the first test is the standardized residual, the second is the autocorrelation function, and the third is the p values of the Ljung–Box statistic. The results showed no volatility cluster in the standardized residual, and no significant autocorrelation in the residual autocorrelation function. Additionally, the p values of the Ljung–Box statistic remained over 0.8, indicating no apparent patterns in the residual. Since the model extracted all useful information with the exception of noise, ARIMA (2,0,4) $(0,0,2)_{52}$ was confirmed as the most accurate ARIMA model for predicting Charleston hotel occupancy.

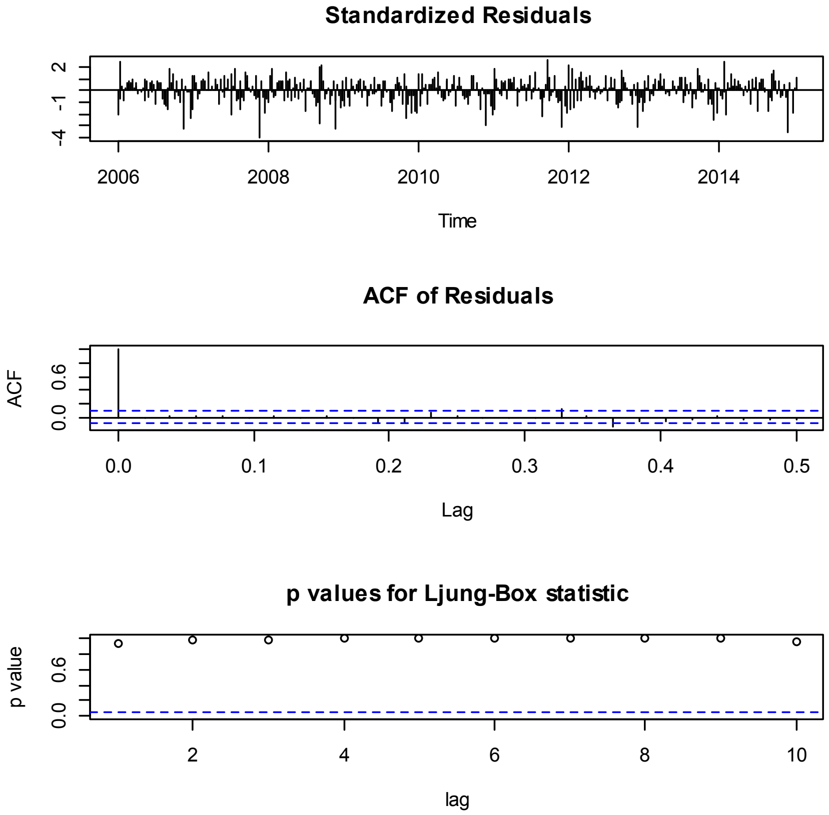


Figure 2. Diagnose of ARIMA (2,0,4) $(0,0,2)_{52}$ model.

A 52-week data forecast was computed by the ARIMA (2,0,4) $(0,0,2)_{52}$ model (Figure 3).

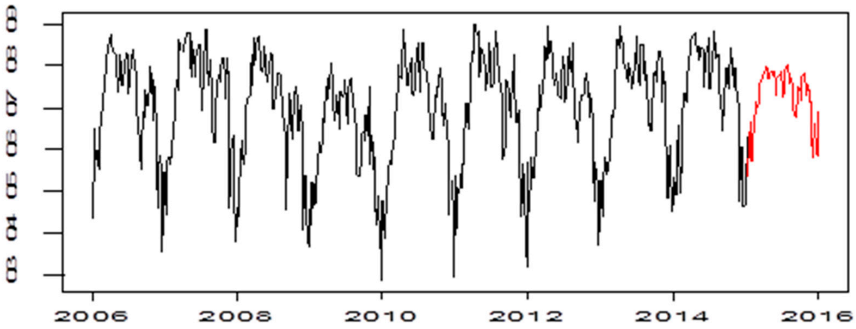


Figure 3. Forecasting results of hotel occupancy by ARIMA (2,0,4) $(0,0,2)_{52}$ model.

To analyze the forecasting effect of the ARIMA model, prediction length was separated into a medium and a long term, or 26 and 52 weeks, respectively. The two models were tested according to MAPE and RMSE. As shown in Table 2, for both MAPE and RMSE, the results from the 52-week predictions were superior to those of the 26-week predictions, indicating that the ARIMA model is more effective in longer-term forecasting.

Table 2. Forecasting performance of ARIMA model for different prediction lengths. MAPE: mean absolute percentage error, RMSE: root mean square error.

Prediction length	MAPE	RMSE
26 weeks	6.4	5.8
52 weeks	6.2	5.5

5.2. EEMD

EEMD was used to split the original hotel occupancy data into several IMFs and the trend term T. The decomposition results are shown in Figure 4. Hotel occupancy in the Charleston area was decomposed into seven IMFs ranging from high to low frequency, and one residual. All of the series were independent from each other. The residual series was considered to be the trend, since it shows the movement in the largest scale. As shown in Figure 4, the volatility of the sequences gradually decreased and cycles grew longer, and clear annual patterns emerged in IMF1 to IMF5 (Figure 4).

Four indices were used to analyze the outcomes of EEMD, which were the average cycle of IMFs, the correlation coefficient between IMFs and the original series, the variance percentage of IMFs in the original series, and the variance percentage of IMFs in the series.

IMF3 had a 17.41-week average period, implying that the original signal exhibited a four-month regular fluctuation. In contrast, IMF4 possessed a near annual fluctuation, with a 42.73-week period. The Pearson product moment correlation coefficient was applied to measure the correlation between the IMFs and the original series. As shown in Table 3, IMF4 and IMF3 are the most closely related to the original signal, followed by IMF5, IMF2, and IMF1. The trend had a weak correlation with the original signal. This may have been caused by fluctuations in the original series. As shown in Figure 4, the trend reaches a trough between 150–200 weeks, and constantly rises afterward. This period corresponds with the 2008 financial crisis in the U.S., which affected the entire hospitality industry.

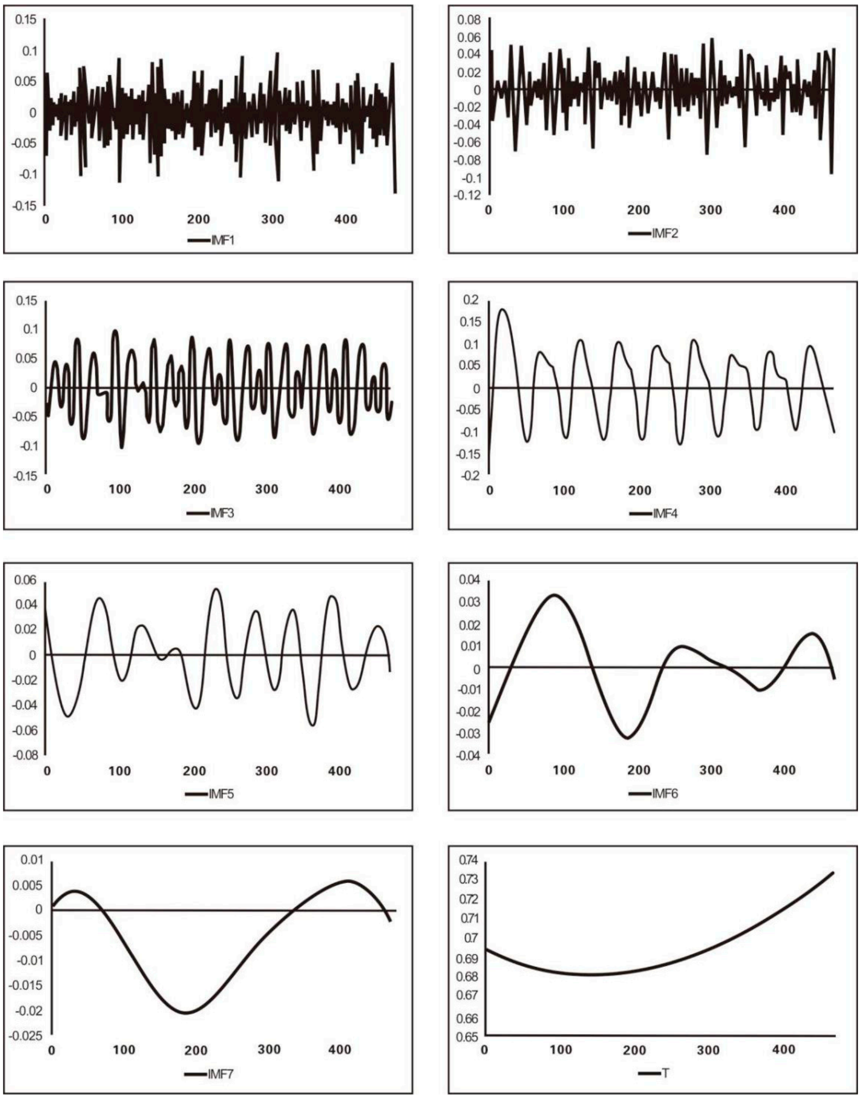


Figure 4. Ensemble empirical mode decomposition (EEMD) of Charleston hotel occupancy signal.

Table 3. Analysis of intrinsic model functions (IMFs) and trend (T) of Charleston hotel occupancy based on EEMD.

	Average Cycle	**Correlation Coefficient**	**Variance Percentage (Decomposed by EEMD)**	**Variance Percentage (Original Series)**
IMF1	3.03	0.14	22.75%	15.48%
IMF2	6.33	0.24	6.91%	4.70%
IMF3	18.27	0.63	20.51%	13.96%
IMF4	36.54	0.74	28.54%	19.42%
IMF5	52.78	0.64	8.19%	5.57%
IMF6	118.75	0.20	1.60%	1.09%
IMF7	237.50	0.19	2.14%	1.46%
T	475.00	−0.02	9.36%	6.37%

5.3. EEMD–ARIMA Method

To further increase the forecasting accuracy, the proposed EEMD–ARIMA model decomposes the original signal into different levels of frequency (IMFs), and simplifies the forecasting process of complicated original data by forecasting each IMF first, thereby improving the accuracy. The EEMD–ARIMA model operates as follows:

(1) The Charleston hotel occupancy signal was decomposed into several IMFs and the trend;
(2) Since each IMF is independent, and their summation is equal to the original signal, ARIMA was used to model every IMF and obtain relevant forecasting values;
(3) All of the forecasting values were summed to obtain the final prediction.

Throughout this process, seven IMFs and T emerged from X(t), and then R was used to model all of the IMFs and *t* values with ARIMA method. The ADF test was examined before the modeling process. To ensure reliable results, the different functions in R were compared, and Table 4 displays the outcomes.

Table 4. Augmented Dickey–Fuller (ADF) test of IMFs and T.

	Test for Unit Root	ADF	Significance Level			Probability Value
			1%	5%	10%	
IMF1	Level	−14.7157	−3.444	−2.868	−2.570	0.0000
IMF2	Level	−16.9913	−3.444	−2.868	−2.570	0.0000
IMF3	Level	−15.1463	−3.444	−2.868	−2.570	0.0000
IMF4	Level	−11.4492	−3.444	−2.868	−2.570	0.0000
IMF5	Level	−6.0545	−3.444	−2.868	−2.570	0.0001
IMF6	Second difference	−3.4075	−3.444	−2.868	−2.570	0.0399
IMF7	Second difference	−2.9771	−3.444	−2.868	−2.570	0.0000
T	Second difference	−22.8449	−3.444	−2.868	−2.570	0.0000

Next, each IMF a nd *t* value was modelled in R, with the outcomes shown in Table 5.

Table 5. ARIMA models of IMF and T from X(t).

	ARIMA	Sigma^2	Log likelihood	AIC
IMF1	(3,0,2) $(0,0,1)_{52}$	0.000941	970.13	−1926.26
IMF2	(3,0,4) $(0,0,2)_{52}$	3.43×10^{-5}	1744.4	−3470.8
IMF3	(4,0,5) $(0,0,2)_{52}$	2.28×10^{-7}	2920	−5818
IMF4	(1,0,2) $(0,0,2)_{52}$	6.92×10^{-6}	2113.24	−4216.48
IMF5	(1,0,2) $(0,0,2)_{52}$	5.11×10^{-7}	2725.2	−5440.39
IMF6	(0,2,5) $(1,2,0)_{52}$	1.01×10^{-12}	5786.86	−11561.73
IMF7	(0,2,2) $(0,2,1)_{52}$	8.38×10^{-13}	5835.36	−11664.72
T	(0,2,5) $(0,0,2)_{52}$	1.03×10^{-13}	6332.75	−12653.5

The forecasted values of seven IMFs and the trend were summed up to obtain the final forecast, and then compared with the actual data. As shown in Figure 5, the predicted series does not predict the actual data series accurately. In the previous EEMD–ARIMA forecasting [7], the original signals were nonstationary. When they were decomposed into relatively stationary signals, the additive effect of all of the IMFs enabled them to overcome the drawback of the complexity in the original signals and improved the forecasting accuracy. However, in this current study, EEMD–ARIMA has less forecasting accuracy than the original ARIMA model. Therefore, a modified EEMD–ARIMA model is proposed.

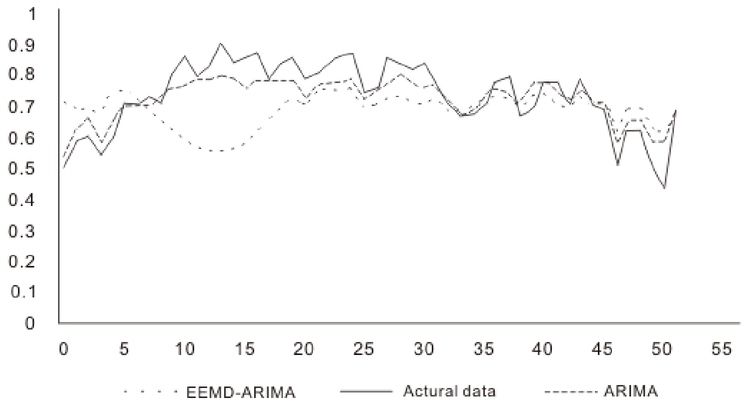


Figure 5. Forecasting results of the modified EEMD–ARIMA model by the summation of IMFs and T in future 52 weeks.

5.4. Modified EEMD–ARIMA Method

Since the original EEMD–ARIMA failed to increase forecasting accuracy compared to the original ARIMA models, partially combining IMFs may reduce the accumulation of errors and increase the strengths of EEMD, thereby improving the accuracy. Previous studies have often applied a t test on each signal, and examined their means to distinguish between high and low frequencies. If the mean of one signal does not equal zero, then all of the subsequent signals are identified as low-frequency signals, and the former ones are high-frequency signals; these are named as the short term, long term, and trend [4]. All of the IMFs from X(t) have a mean of near zero; therefore, the traditional method is not applicable. Since this data series has obvious seasonal fluctuation, these IMFs were divided into high and low-frequency signals by cycle. IMFs with an average cycle smaller than 52 (larger than annual cycles) were considered high frequency signals, and all of the others were low-frequency signals. Accordingly, IMF1–IMF4 were classified as high-frequency signals, whereas IMF5–IMF7 were low-frequency signals. This novel research model is depicted in Figure 6.

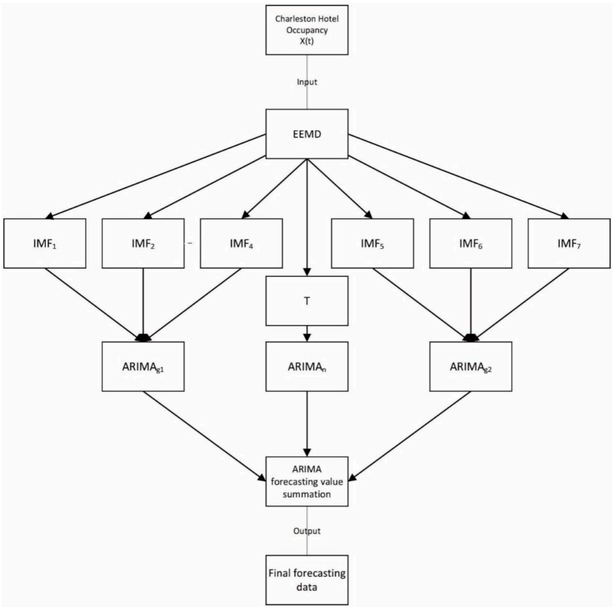


Figure 6. Modified model structure by EEMD–ARIMA.

ARIMA was applied to the IMFs of the high-frequency group, low-frequency group, and trend to produce a 52-week forecast. The results are shown in Table 6.

Table 6. ADF test and ARIMA model making based on the distinction of high frequency and low frequency.

	Test for Unit Root	ADF	Significance Level			Probability
			1%	5%	10%	
High frequency g1	Level	−6.900	−3.444	−2.868	−2.570	0.0000
Low frequency g2	1st difference	−3.267	−3.444	−2.868	−2.570	0.0170
	ARIMA	Sigma^2	Log likelihood			AIC
High frequency g1	(5,0,4) (0,0,2)$_{52}$	0.004322	611.58			−1201.16
Low frequency g2	(0,0,1) (1,1,1)$_{52}$	0.0003733	1056.52			−1486.54

Next, we compare the prediction calculated with the traditional ARIMA models with those of the Modified EEMD-ARIMA. As shown in Figure 7, in short-term forecasting (10 weeks or less), the forecasting values of both the ARIMA and EEMD–ARIMA models were higher than the real values. However, in medium-term forecasting, the EEMD–ARIMA predictions were clearly closer to the actual values. Specifically, in the MAPE and RMSE tests shown in Table 7, the EEMD–ARIMA model reduced 31.25% and 31.03% of the error in medium-term forecasting, respectively. The effects of long-term forecasting were inferior to those of medium-term forecasting, reducing only 9.68% of the error in MAPE and 16.36% in RMSE. Thus, the modified EEMD–ARIMA successfully improved the forecasting accuracy for weekly hotel demand, compared to the ARIMA models.

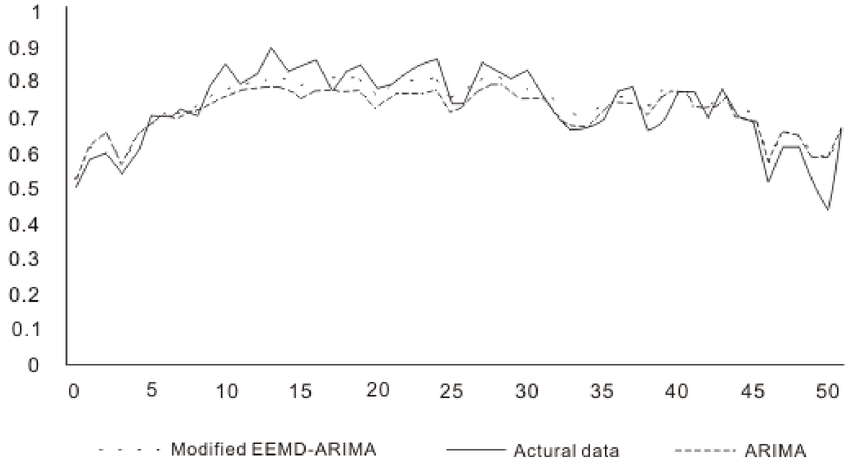


Figure 7. Forecasting results by modified EEMD–ARIMA model in future 52 weeks.

Table 7. Forecasting performance comparison between ARIMA and modified EEMD–ARIMA.

Forecasting Length	ARIMA		EEMD–ARIMA	
	MAPE	RMSE	MAPE	RMSE
26 weeks	6.4	5.8	4.4	4.0
52 weeks	6.2	5.5	5.6	4.6

5.5. Testing Modified EEMD–ARIMA with More Data Series

To test the efficacy of the improved modified EEMD–ARIMA model, four sub-areas in the city of Charleston were selected for a comparative analysis. They are the peninsula, West Ashley, East Cooper, and North Charleston (Table 8).

Table 8. Descriptive statistics of the four sub-areas.

	Mean	**Median**	**Maximum**	**Minimum**	**Range**	**Std. Dev.**	**Skewness**	**Kurtosis**
Charleston total	0.703	0.735	0.902	0.285	0.617	0.128	−0.807	2.988
Peninsula	0.756	0.792	0.949	0.315	0.634	0.132	−0.965	3.382
West Ashley	0.723	0.760	0.962	0.320	0.642	0.140	−0.660	2.586
East Cooper	0.661	0.700	0.930	0.216	0.714	0.161	−0.586	2.391
North Charleston	0.671	0.691	0.913	0.245	0.668	0.128	−0.738	3.273

MAPE and RMSE test results are provided in Table 9. The trend and accuracy of the forecasting of the four areas are very similar to those for the whole area shown in Figure 7, due to the highly correlated nature of the four data series of sub-areas and the whole Charleston area. The forecasting accuracy of 26 weeks' hotel demand all increased compared to the ARIMA models; for two areas of East Cooper and North Charleston, the forecasting accuracy of 52 weeks' data deteriorated, while those of the other two areas improved. Compared with the descriptive statistics in Table 8, the data series from East Cooper and North Charleston areas had relatively larger ranges, especially lower minimum values. This might indicate that the EEMD–ARIMA model has a greater effect on time series that are more stable.

Table 9. Forecasting performance comparison between ARIMA and modified EEMD–ARIMA of four sub-areas.

	Forecasting Length	**ARIMA**		**Modified EEMD-ARIMA**	
		MAPE	RMSE	MAPE	RMSE
Charleston Total	26 weeks	6.4	5.8	4.4	4
	52 weeks	6.2	5.5	5.6	4.6
Peninsula	26 weeks	10	9.9	10	9
	52 weeks	8.8	8.6	8.6	8.1
West Ashley	26 weeks	8	6.8	7.7	6.5
	52 weeks	7.2	6.2	6.8	6
East Cooper	26 weeks	12	9.7	11.8	9.3
	52 weeks *	10.9	8.8	12.2	9.2
North Charleston	26 weeks	6.7	6.1	5.7	4.9
	52 weeks *	7.9	6.5	10.3	7.9

* bold numbers indicate non-improvement.

6. Conclusions

This study tested a modified EEMD–ARIMA model for tourism demand forecasting by combing time-series models with time-frequency analysis. Charleston, South Carolina, USA was used as an empirical case by forecasting its weekly hotel occupancy. The overall Charleston area, as well as data series in four specific sub-areas inside Charleston, were tested. The overall prediction results were compared to traditional time series models. The modified EEMD–ARIMA model universally improved forecasting accuracy for 26 weeks' ahead, but failed to do so in two of the five areas tested for 52 weeks' forecasting. Thus, the results are mixed: the modified EEMD–ARIMA forecasting model performed better with medium-term forecasting and with data series with smaller ranges.

ARIMA was employed as a benchmark to model hotel occupancy data and forecast 52-week data sets. Through a test of a self-adaptive time-frequency analysis tool, EEMD, hotel occupancy signals were split into several IMFs, which assisted in exploring the intrinsic regularities of the data. However, a traditional EEMD–ARIMA model did not result in more accurate forecasting; thus, the model was modified by combining seven IMFs and a residual into three series, which were labeled high-frequency fluctuation, low-frequency fluctuation, and trend. The three fluctuations were modelled with ARIMA, and their forecasting values were summed to obtain the final results. Validated with the models of the four sub-areas, the modified model markedly increases the forecasting accuracy for 26 weeks' ahead forecasting, and for data series with a relatively small range and standard deviation.

7. Discussion and Future Research

This study demonstrated that fluctuation patterns can be extracted accurately from hotel occupancy signals by using EEMD. This may assist researchers in analyzing fluctuation regularities in different frequencies, through which the trend and patterns can be identified. Patterns of different frequencies were shown to correlate with economic development cycles.

Although EEMD–ARIMA does work well on decomposing fluctuated and nonlinear signals, the model has its limitations. The current study showed that the EEMD–ARIMA model was not as universally applicable as it was thought to be, due to the different types of data. The EEMD–ARIMA model didn't show accurate prediction even in the short term. After being modified and tested by relatively stable weekly data, a modified EEMD–ARIMA can achieve more accurate medium-long term forecasting. In addition, the previous study [7] had not tested other data samples, while the modified EEMD–ARIMA model are considered more reliable, since it was tested on four other data series. Last but not least, since the hotel occupancy is a symbol of tourism demand, the empirical results can not only support hotel managers for decision making, they can also provide support for tourism destinations' management, resource allocation, and sustainable development.

Despite the ability of the modified EEMD–ARIMA model to significantly increase medium-term forecasting accuracy, we have not tested it on forecasting turning points and dramatic fluctuations. In addition, the improvement of forecasting accuracy is only at most 1–2% of MAPE. Although relatively significant compared to the overall 6–11% of MAPE, it is still limited in practice. The variable is the occupancy rate, which ranges from 0% to 100%. Thus, it has a lower and an upper limit. Thus, the accuracy could be further improved by automatically limiting both boundaries. Furthermore, this study focused only on hotel occupancy data; other tourism demand data such as tourist arrival or spending could be investigated with the method in the future. In addition, since destinations or areas are spatially correlated, future research could adopt spatial correlation in forecasting those areas simultaneously to achieve better results.

Author Contributions: Data curation, B.P.; Formal analysis, M.Z.; Methodology, M.Z., J.L., B.P. and G.Z.; Supervision, J.L. and B.P.; Writing–original draft M. Z.; Writing-review & editing, M. Z., J. L. and B.P.

Funding: This project was jointly funded by Natural Science Foundation Projects No. 41571135, No. 41428101, and Jinan University, Shenzhen Tourism College OSP #193942 and #193940.

Acknowledgments: Qing Zhu of the International Business School of Shaanxi Normal University offered tremendous guidance and help in research methodology.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Frechtling, D.C. *Forecasting Tourism Demand: Methods and Strategies*; Butterworth Heinemann: Oxford, UK, 2001.
2. Deng, Z.L. *Optimal Filtering Theory and Application*; Harbin Institute of Technology Press: Harbin, China, 2000.
3. Chen, K.Y. Combining linear and nonlinear model in forecasting tourism demand. *Expert Syst. Appl.* **2011**, *38*, 10368–10376. [CrossRef]
4. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proc. Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
5. Wang, T. *Research on EMD Algorithm and Its Application in Denoising, Doctoral Dissertation*; Harbin Engineering University: Harbin, China, 2010; unpublished.

6. Wu, Z.; Huang, N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **2009**, *1*, 1–41. [CrossRef]
7. Zhang, G.; Wu, J.; Pan, B.; Li, J.; Ma, M.; Zhang, M.; Wang, J. Improving daily occupancy forecasting accuracy for hotels based on EEMD-ARIMA model. *Tour. Econ.* **2017**, *23*. [CrossRef]
8. Zhao, X.H.; Chen, X. Auto regressive and ensemble empirical mode decomposition hybrid model for annual runoff forecasting. *Water Resour. Manag.* **2015**, *29*, 2913–2926. [CrossRef]
9. Song, H.; Li, G. Tourism demand modelling and forecasting—A review of recent research. *Tour. Manag.* **2008**, *29*, 203–220. [CrossRef]
10. Song, H.Y.; Witt, S.F. *Tourism Demand Modelling and Forecasting: Modern Econometric Approaches*; Pergamon: Oxford, UK, 2000.
11. Pan, B.; Wu, D.C.; Song, H. Forecasting hotel room demand using search engine data. *J. Hosp. Tour. Technol.* **2012**, *3*, 196–210. [CrossRef]
12. Bangwayo-Skeete, P.F.; Skeete, R.W. Can google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tour. Manag.* **2015**, *46*, 454–464. [CrossRef]
13. Li, G.; Song, H.; Witt, S.F. Time varying parameter and fixed parameter linear aids: An application to tourism demand forecasting. *Int. J. Forecast.* **2016**, *22*, 57–71. [CrossRef]
14. Gunter, U.; Önder, I. Forecasting international city tourism demand for paris: Accuracy of uni- and multivariate models employing monthly data. *Tour. Manag.* **2015**, *46*, 123–135. [CrossRef]
15. Makridakis, S.; Andersen, A.; Carbone, R.; Fildes, R.; Hibon, M.; Lewandowski, R.; Newton, H.J.; Parzen, E.; Winkler, R. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *J. Forecast.* **1982**, *1*, 111–153. [CrossRef]
16. Makridakis, S.; Chatfield, C.; Hibon, M. The M2-Competition: A Real-Time Judgmentally Based Forecasting Study. *Int. J. Forecast.* **1993**, *9*, 5–22. [CrossRef]
17. Makridakis, S.; Hibon, M. The M3-Competition: Results, conclusions and implications. *Int. J. Forecast.* **2000**, *16*, 451–476. [CrossRef]
18. Peng, B.; Song, H.; Crouch, G.I. A meta-analysis of international tourism demand forecasting and implications for practice. *Tour. Manag.* **2014**, *45*, 181–193. [CrossRef]
19. Li, X.; Pan, B.; Law, R.; Huang, X. Forecasting tourism demand with composite search index. *Tour. Manag.* **2017**, *59*, 57–66. [CrossRef]
20. Law, R. Room occupancy rate forecasting: A neural network approach. *Int. J. Contemp. Hosp. Manag.* **1998**, *10*, 234–239. [CrossRef]
21. Zakhary, A.; Atiya, A.F.; El-Shishiny, H.; Gayar, N.E. Forecasting hotel arrivals and occupancy using monte carlo simulation. *J. Revenue Pricing Manag.* **2011**, *10*, 344–366. [CrossRef]
22. Pai, P.F.; Hung, K.C.; Lin, K.P. Tourism demand forecasting using novel hybrid system. *Expert Syst. Appl. Int. J.* **2014**, *41*, 3691–3702. [CrossRef]
23. Hassani, H.; Webster, A.; Silva, E.S.; Heravi, S. Forecasting U.S. tourist arrivals using optimal singular spectrum analysis. *Tour. Manag.* **2015**, *46*, 322–335. [CrossRef]
24. Caicedo-Torres, W.; Payare, F. A Machine Learning Model for Occupancy Rates and Demand Forecasting in the Hospitality Industry. In *Advances in Artificial Intelligence, Proceedings of the IBERAMIA 2016, Costa Rica, San José, 23–25 November 2016*; Montes y Gómez, M., Escalante, H., Segura, A., Murillo, J., Eds.; Springer International Publishing: New York, NY, USA, 2016.
25. Schwartz, Z.; Uysal, M.; Webb, T.; Altin, M. Hotel daily occupancy forecasting with competitive sets: A recursive algorithm. *Int. J. Contemp. Hosp. Manag.* **2016**, *28*, 267–285. [CrossRef]
26. Chen, C.F.; Lai, M.C.; Yeh, C.C. Forecasting tourism demand based on empirical mode decomposition and neural network. *Knowl.-Based Syst.* **2011**, *26*, 281–287. [CrossRef]
27. Lei, Y.; He, Z.; Zi, Y. Application of the eemd method to rotor fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.* **2009**, *23*, 1327–1338. [CrossRef]
28. Wu, Z.; Huang, N.E. A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2004**, *460*, 1597–1611. [CrossRef]
29. Charleston Area CVB. *2014–2015 Charleston Area Convention & Visitors Bureau Book*; Charleston Convention & Visitors Bureau: Charleston, SC, USA, 2015.
30. Yang, Y.; Pan, B.; Song, H.Y. Predicting hotel demand using destination marketing organizations' web traffic data. *J. Travel Res.* **2014**, *53*, 433–447. [CrossRef]

31. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*; Holden-Day: San Francisco, CA, USA, 1970.
32. R: The R Project for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 10 October 2018).