To Cite:

**IMPROVING DAILY OCCUPANCY FORECASTING ACCURACY FOR HOTELS BASED ON EEMD-ARIMA MODEL**

Gaojun Zhang, Ph.D.
1. Shenzhen Tourism College, Jinan University, Shenzhen, China
2. College of Tourism and Environment, Shaanxi Normal University, Xi’an, China
   Telephone: 86-189-3807-5205
   E-mail: zhang_gj@sz.jnu.edu.cn

Jinfeng Wu*, Ph.D.
Professor
College of Tourism and Environment, Shaanxi Normal University, China
   Telephone: 86-130-8754-1092
   E-mail: jfwu@snnu.edu.cn

Bing Pan*, Ph.D.
Associate Professor
Department of Recreation, Park, and Tourism Management
College of Health and Human Development
Penn State University, University Park, PA 16802, USA
   Telephone: 1-814-865-1851 Fax: 1-814-867-1751
   E-mail: bingpan@psu.edu

Junyi Li, Ph.D.
Professor
College of Tourism and Environment, Shaanxi Normal University, Xi’an, China
   Telephone: 86-133-79263086
   Fax: 86-29-85310495
   Email: lijunyi9@snnu.edu.cn

Minjie Ma, M.S.
Xi’an Special Education School, China
   Telephone: 86-189-9196-0469

Muzi Zhang, M.S.
School of Tourism and Environment
Shaanxi Normal University, China
   Telephone: 86-182-0280-6740
   E-mail: 695210761@qq.com
Jian Wang, Ph.D.
The Rainmaker Group, USA
4550 North Point Pkwy #400, Alpharetta, GA 30005,
Phone: (678) 578-5700
Email: jwang@letitrain.com

*Corresponding Authors

March 21, 2017

Acknowledgements
This research is partially supported by a grant from the National Natural Science Foundation of China (Grant # 41428101). We want to thank the anonymous company that provided the valuable data. Dr. Qing Zhu of the International Business School of Shaanxi Normal University offered tremendous help in research methodology.

Short Title: Improving Hotel Forecasting with EEMD-ARIMA Model
Abstract

Predicting daily occupancy is extremely important for the revenue management of individual hotels. However, daily occupancy can fluctuate widely and is difficult to forecast accurately based on existing forecasting methods. In this paper, Ensemble Empirical Mode Decomposition (EEMD)—a novel method—is introduced, and an individual hotel is chosen to test the effectiveness of EEMD in combination with an autoregressive integrated moving average (ARIMA). Result shows that this novel method, EEMD-ARIMA, can improve forecasting accuracy better than the popular ARIMA method, especially for short-term forecasting.

Keywords: forecasting; occupancy; individual hotels; EEMD; ARIMA
IMPROVING DAILY OCCUPANCY FORECASTING ACCURACY FOR HOTELS BASED ON EEMD-ARIMA MODEL

Introduction

Accurate forecasting plays an important role in human society, ranging from dealing with natural disasters to the development of economies. Accurate forecasting can help individuals or organizations allocate resources optimally (Song and Li, 2008). Private businesses and local governments can use accurate forecasting to make decisions on guiding traffic, designing better infrastructure, adjusting prices, and formulating policies (Song and Li, 2008). This also applies to the tourism industry. For tourism businesses, accurate forecasting can help them reduce the costs of excess capacity or unfilled demand due to the perishable nature of tourism products (Archer, 1987; Frechtling, 2001; Xie, 2010). Indeed, revenue cannot be realized from a product that is unsold the following day.

Moreover, accurate forecasting of tourism demand has been one of the most important fields in tourism economics over the past few decades (Song et al., 2012; Witt and Witt, 1995). Tourism forecasting research has gone through several methodological developments, and numerous models have been proposed, including time series models and multivariate models. However, as an increasing number of methods were employed in tourism demand forecasting, researchers started to compare and discuss the best-performing
models (Schwartz, 1999; Weatherford and Kimes, 2003; Zakhary et al., 2011). However, after extensive studies, no consensus has been reached (Peng et al., 2014). Based on empirical evidence, no single model can consistently outperform others on all occasions (Peng et al., 2014; Schwartz and Hiemstra, 1997). Different methods suit different contexts, and the latest research tends to combine different methods with the purpose of improving forecasting accuracy (Li et al., 2006; Shen et al., 2011; Wong et al., 2007). For example, Chen KH (2011) combined linear and non-linear models, while Shen et al. (2011) used six linear combination methods; Chan et al. (2010) employed programming approaches to determine the weights of combination; and Song et al. (2013) combined statistical and judgmental approaches to improve forecasts. All these empirical studies generally provide evidence of favorable outcomes by combining forecasting methods.

In general, the improvement of forecasting accuracy between these combined methods is similar. In fact, some researchers believe model complexity does not necessarily improve forecasting accuracy (Lim et al., 2009; Pan et al., 2012), and, as Koupriouchina et al. (2014) noted, the future of forecasting performance is possibly leading to a more comprehensive forecasting evaluation framework. Therefore, in order to improve forecasting accuracy significantly, it is necessary to uncover the origin of forecasting errors. In the field of tourism demand forecasting, indicators such as tourist volumes and
occupancy rates fluctuate wildly with no obvious regularity. This is due to the fact that the factors influencing tourism demand fluctuate themselves. If the irregularity of tourism demand data can be reduced, or irregular fluctuations can be converted to regular patterns, it would be easier to capture the rules for a forecasting model. This would lead to an improvement in forecasting accuracy. The present work mainly focuses on combining two different methods—ARIMA and a novel method, named EEMD—to improve forecasting accuracy.

**Theoretical background**

In hotel revenue management practices, accurate forecasts can facilitate decision making on pricing for an individual property or a group of hotels (Yang et al., 2014). The amount of revenue generated from a hotel often depends on its ability to utilize its capacity efficiently; therefore, knowing how many guests would arrive in the coming days is very useful (Heo et al., 2013). This makes high quality forecasting of guest arrivals and occupancy rates a key aspect for hotel revenue management (Koupriouchina et al., 2014; Weatherford and Kimes, 2003). As Talluri and Van Ryzin (2004: 407) observed: “A revenue management system requires forecasts of quantities such as demand, price sensitivity, and cancellation probabilities, and its performance depends critically on the quality of these forecasts.” Despite the importance of occupancy forecasting, research in the field is still lacking (Heo et al., 2013).
The literature has been rather undeveloped before the 2000s, with a few notable published studies after 2000 (Weatherford et al., 2001; Zakhary et al., 2011).

**Multivariate Models**

Multivariate models might be among the most used in tourism demand forecasting. For example, Jeffrey and Barden (2000) induced a stepwise regression analysis with nine variables to forecast the occupancy peak and trough of chain or group hotels. Also, a similar multivariate model can be seen in Yüksel’s (2007) study, which combined 25 variables to forecast the monthly occupancy of a five-star hotel. Concentrating on tourism data only, Zakharya et al. (2011) forecasted a hotel occupancy rate with many variables such as reservation arrivals, cancellations, length of stay, no-shows, group reservations, seasonality, trends and so on. Conversely, Chen MH (2011) measured the forecasting ability of a single variable—crisis events—to hotel occupancy.

**Time Series Models**

In addition to multivariate models, time series models are another important method for forecasting hotel occupancy (Gunter and Önder, 2015; Weatherford and Kimes, 2003; Zakharya et al., 2011). Specifically, one form of the time series model is the (seasonal) autoregressive (integrated)
moving-average model ((S)AR(I)MA). In addition, novel time series models have also emerged in tourism demand forecasting, such as error trend seasonal or exponential smoothing (ETS) (Hyndman et al., 2002, 2008); cointegration and error correction (Haliçioğlu et al., 2010; Song et al., 2011b); the vector autoregressive method (VAR); autoregressive distributed lag (ADL) (Song et al., 2013); the time-varying parameter (TVP) (Song et al., 2011a); an exponentially weighted moving average (E(W)MA) combined with the Holt-Winters method (Aghazadeh, 2007); and autoregressive mixed data sampling (AR-MIDAS) (Bangwayo-Skeete and Skeete, 2015). All of them have performed well in forecasting future tourism demand under certain scenarios.

Many researchers have tried to forecast the average occupancy of hotels in different areas, countries, regions, cities, and even for some specific hotels. For example, Andrew, Cranage and Lee (1990) employed time series and exponential smoothing methods to forecast the monthly occupancy of a major city-center hotel. They found that both methods function very well and produce a very low mean sum of squared residuals; moreover, they stated that the Box-Jenkins approach outperforms exponential smoothing. Later, Law and Au (1999) published work focused on regional occupancy rate, and he validated the value of the neural network approach to forecast annual hotel occupancy in Hong Kong. He compared the accuracy of the neural network approach with with multiple regressions, and concluded that the
neural network method performed the best among the three methods. Post-2000, Choi (2003) identified the leading economic indicators that can affect the US hotel industry, then built a model based on those indicators and used the novel indicator system model to forecast hotel demand in the United States.

In addition, Schwartz and Hiemstra (1997) proposed a method to forecast daily occupancy based on previous booking curves. Other researchers updated a special version of the ESM method (the Holt-Winters method) to forecast daily hotel room demand in a particular property (Rajopadhye et al., 1999). In order to forecast a series of occupancies of Norwegian guests in Swedish hotels, the authors concluded that models based on differenced series were preferred when measured by goodness-of-fit. In Yüksel's (2007) work, the analytical hierarchy process (AHP) was employed to adjust the forecast of ESM models, which integrated quantitative and qualitative methods, and obtained a lower error rate between real and forecast values. In comparison, Koupriouchina et al. (2014) utilized system forecasting (prepared via the computer) and manual forecasting (prepared by hotel managers) to investigate a better method. Interestingly, no single method has outperformed others in all cases. Although the time series method usually performs the best, it failed to do so in Italy, where pick-up models achieved the best results (Ellero and Pellegrini, 2014). Therefore, we need more real-life cases and the
In the past, annual, quarterly, or monthly data was often used in tourism demand forecasting, as well as hotel demand forecasting (Song and Li, 2008; Table 1). However, forecasting with daily data is the most crucial approach due to fierce market competition in the hospitality industry. Compared to other indicators, such as international tourist arrivals, the occupancy of hotels is more relevant to the industry (Pan et al., 2012). Therefore, in this paper, we have selected daily hotel occupancy as the dependent variable. As noted by Weatherford et al. (2001), hotels should forecast occupancy in high frequency in order to fulfill the needs of revenue management. Consequently, this paper would attempt to improve the forecasting the daily occupancy of an individual hotel, with the assistance of a new and unique method known as ensemble empirical model decomposition (EEMD). This paper is organized as follows: the next section is an introduction to EEMD; the following one shows the data and forecasting results calculated by EEMD and ARIMA; and the last section summarizes the paper and discusses the results as well as the new method.
Ensemble Empirical Mode Decomposition

EEMD is derived from empirical mode decomposition (EMD) (Wu and Huang, 2009). EMD was proposed by Huang et al. (1998) and is considered a major breakthrough in linear and stable analysis based on Fourier transformation, which won the NASA Government Invention of the Year in 2003 (NASA, 2003). The method is proficient in dealing with non-stationary data by identifying the different scales of volatility in the origin wave and decomposing it into a series of new waves; the new wave is called an intrinsic mode function (IMF), and the IMF with the lowest frequency represents the trend of the data series. Owing to the self-adaptive nature of IMF, its frequency and Hilbert spectrum can be calculated by Hilbert transformation.

The application of EMD for the analysis of wave $x(t)$ should satisfy three pre-conditions:

1. Oscillating and periodic patterns are repeated;
2. The local mean is zero and the signal is symmetric to its local mean;
3. One cycle of oscillation should comprise the following: the sinusoidal function should start at zero, end at zero, and pass through zero between two crossings. Alternately, it should start at the local maximum and terminate at consecutive local maxima, passing through two zeros and the local minimum.

When the original signal meets the requirements given above, EMD can be
used to systematically decompose the signal. In short, the EMD decomposition is a screening process of the vibration mode. The basic principle of EMD is to use the cubic spline function to fit the partial maximum and minimum values. By doing so, one can confirm the instantaneous equilibrium position of the signal. Consequently, several IMFs and a residual that satisfy some particular criteria can be extracted, and finally, the original signal is decomposed. The unique advantage of EMD is that IMFs can highlight partial characteristics, and the residual can reveal the internal trend of the original signal. By analyzing the IMFs, the characteristics of the origin signal information can emerge.

Since IMF is an important indicator in EMD, Huang et al. (1998) defined how an oscillating wave can be considered as IMF. It should satisfy two conditions:

1. The number of extrema and the number of zero crossing must differ only by one;
2. The local average is zero, meaning that the mean of the upper envelope and lower envelope is zero.

Specifically, the process of EMD analysis is as follows:

1. Identify the local extrema, including the maximum and minimum;
2. Connect all of the local maxima to form a new signal with a cubic spline function as the upper envelope $U_i(t)$; similarly, all of the local
minima forms the lower envelope \( L_i(t) \).

(3) Calculate the mean curve of the upper and lower envelope \( M_1(t) \)
\[
M_1(t) = \frac{1}{2} [U_i(t) - L_i(t)]
\]

(4) Calculate a signal \( S_{1,1}(t) \) by \( X(t) \) minus \( M_1(t) \)
\[
S_{1,1}(t) = x(t) - M_1(t)
\]

(5) Check the residue \( S_{1,1}(t) \) to make sure it satisfies the two conditions of IMF in the previous statements, and, if so, we obtain the first IMF, and, if not, repeat the process from (1) to (5) while \( S_{1,1}(t) \) is taken as the new signal until the residue satisfies the two properties. However, in practice, the actual meaning would be lost if the process is to be repeated several times, which would make the IMF an amplitude constant frequency modulation signal.

Therefore, Peng et al. (2005) noted that the standard deviation of the adjacent signals \( S_{1,k-1}(t) \) and \( S_{1,k}(t) \) can be used as stopping criterion (the experience value is 0.2-0.3). When the standard deviation (SD) meets the threshold value, the screening process would stop, and the IMF is abstracted;

(6) Separate the residue \( R_1(t) \) from \( x(t) \)
\[
R_1(t) = x(t) - IMF_1
\]

However, \( R_1(t) \) may contain some IMFs with a longer period, and therefore another sifting is needed for \( R_1(t) \). More IMFs would be separated, namely IMF_2, IMF_3, IMF_4 … IMF_n and \( R_2(t), R_3(t), R_4(t) ... R_n(t) \).

The entire process can be represented as follows:
\[
\begin{align*}
R_1(t) &= x(t) - IMF_1 \\
R_2(t) &= R_1(t) - IMF_2 \\
& \quad \ldots \ldots \ldots \\
R_{n-1}(t) &= R_{n-2} - IMF_{n-1} \\
R_n(t) &= R_{n-1} - IMF_n \\
\end{align*}
\]

When the residue \( R_n(t) \) is a constant or a monotonic function, the process ends, and, at the same time, the latent trend of the signal can be judged by \( R_n(t) \), which is also known as trends function.

Now the signal is decomposed into IMF1, IMF2, IMF3 … IMFn and \( R_n(t) \)
\[
x(t) = \sum_{j=1}^{n} IMF_j + R_n(t).
\]

As a self-adaptive signal analysis method, EMD has been successfully adopted in various fields. However, many problems exist. The most serious one is model mixing (also termed modal aliasing). This happens when either:

(1) one IMF mixes two different models with distinct cycles; or (2) two or more IMFs contain the same model with one cycle.

To deal with this limitation, Huang and Wu (2008) upgraded EMD to EEMD, which can deal with illusive components or mode mixing and overlap effectively (Huang et al., 2009; Wu et al., 2007, 2009; Wu and Huang, 2004). Further, the decomposition of signals by EEMD is better than that of EMD (Wu and Huang, 2009). Consequently, IMFs would be more distinct and reliable. EEMD is essentially a method to handle signals with white noise and is known as noise assisted data analysis. The method is the opposite of noise
reduction, where white noises $w(t)$ would be added to the original signal $x(t)$ to form a new signal $X(t)$.

$$X(t) = x(t) + w(t)$$

$w(t)$ is a small white Gaussian noise that obeys normal distribution:

$$\xi_n = \frac{\xi}{\sqrt{N}}$$

In the formula above, $N$ is the total number of noise series; $\xi$ is the amplitude of the white Gaussian noise; and $\xi_n$ is the standard deviation, which is calculated from the refactoring signal and original signals. In other words, $\xi_n$ is the deviation between input signal and the corresponding result that is obtained from the refactoring IMF components. $X(t)$ is the decomposing signal, and many white Gaussian noises are added to $x(t)$ to simulate the observation:

$$X_i(t) = x(t) + W_i(t)$$

Then the new signal $X_i(t)$ is decomposed, we obtain a group of IMFs and residual components:

$$X_i(t) = \sum_{j=1}^{q} IMF_{ij} + R_{ij}(t)$$

According to the zero-mean features of the white Gaussian noises’ uniform distribution, we count the average of all the IMFs in order to offset the impact of the white Gaussian noise. At the same time, the modal aliasing problem can be solved (Wu et al., 2009).
\[
IMF_i = \frac{1}{n} \sum_{j=1}^{n} IMF_{ij}
\]

\[
R_x(t) = \frac{1}{n} \sum_{j=1}^{n} R_{nj}
\]

In the process, the amplitude of added white noise is limited, rather than infinite or infinitesimal. Wu et al. (2009) pointed out that the amplitude of noise has little effect on the decomposition result, and the application of EEMD has no subjectivity and is also self-adaptive. With the intention of ensuring that the IMFs are closer to the real signal, a sufficient number of noise series should be added to the decomposition process. Moreover, the amplitude of noise should be as small as possible, so the error can be reduced.

\[
T = \frac{N_1}{N_2}
\]

In the formula above, \(T\) denotes the period, \(N_1\) denotes the number of original data, and \(N_2\) denotes the number of maximums or minimums. In this paper, we choose the number of maximums as \(N_2\).

As a novel method to deal with non-stationary data, EEMD has been applied to the natural and social sciences for forecasting purposes. In astronomy, Monjoly et al. (2016) attempted to conduct hourly forecasting of global solar radiation with EEMD and significantly improved solar forecast accuracy; in meteorology, Qi et al. (2016) combined EEMD and SVR (support vector regression) to forecast monthly rainfall; while Zang et al. (2016) forecasted the short-term wind power interval with the combination of EEMD, a runs test (RT), and a relevance vector machine (RVM). In the context of the rapid
development of high-speed rail (HSR) in China, Jiang et al. (2014) adopted EEMD and combined it with a grey support vector machine (GSVM) to forecast the daily ridership of the Wuhan-Guangzhou HSR, and found that the forecasting error was much lower than the other two existing forecasting approaches (a support vector machine and autoregressive integrated moving average). Other studies have also adopted the EEMD method (Basha et al., 2015; Li et al., 2008; Plakandaras et al., 2014; Tang et al., 2012). All of them proved that EEMD has distinct advantages in dealing with time series data and integrating with existing forecasting methods.

Tourism researchers have also adopted EMD for tourism demand forecasting. Cao et al. (2016) adopted EMD and wavelet decomposition methods to analyse the characteristics of multi-scale fluctuations and calculate tourism efficiency in Chinese national scenic areas. Moreover, Lai et al. (2013) integrated EMD and support vector regression (SVR) for forecasting international tourist arrivals to Taiwan, and Shabri (2015) forecasted tourism demand with the Group Method of Data Handling (GMDH) and ARIMA. He added EMD in the process and showed that both of the proposed models—EMD-GMDH and EMD-ARIMA—were better than traditional single methods. Wang (2015) combined EMD with ARMA and also demonstrated the superior performance of the new model.
However, the improved model of EEMD has not been widely adopted in tourism research with a few exceptions. When forecasting tourist arrivals from Singapore to Malaysia, Shabri (2016) blended a Least Square Support Vector Machine (LSSVM) and EEMD, and indicated that EEMD can efficiently improve prediction accuracy. In this paper, EEMD will be combined with a popular and useful method, ARIMA, to forecast the occupancy rate of an individual hotel, with the purpose of improving forecasting accuracy.

Data resource

The daily occupancy data in this study is provided by a leading revenue management software company offering pricing solutions for hotels and gaming casino resorts as well as apartments.

Figure 1 shows the annual fluctuation of occupancy in the test hotel (LMN), from July 12, 2007 to August 14, 2014. The daily occupancy fluctuates with no apparent patterns. The only observable regularity is that the occupancy declines first and rises later—twice in December every year.

------------------- Insert Figure 1 here -------------------

Hotel occupancy forecasting with EEMD-ARIMA

First, we need to set up a few parameters to instruct the model on how to decompose the data. According to past studies (Wu et al., 2009), the number
of IMFs is set to 10 since the real occupancy data has 2591 data points; the standard deviation of white noise is set to 0.2, and the number of integrations is 100. The results of the EEMD are displayed in Figure 2; the first sub-plot named signal is the real occupancy, followed by IMF1 to IMF10, and the last one is the residue, also termed “trend.” Figure 2 shows that the randomness is decreasing from signal to trend, and, in turn, patterns emerge at higher levels of aggregation.

In total, the IMFs and trend decomposed by EEMD can explain 75.5% of the variance of occupancy. This is because of the nonlinear wave added in the original data series (Peel et al., 2005). The fluctuation period of IMF1 is about 3 days, which contributes the most to the variance explanation and accounts for 26.3% and 34.9% of the original and decomposed data, and it has the widest oscillation. The oscillation was obvious from July 2007 to May 2009, during the early stages of hotel operations. After operating for 600 days, the oscillation decreased, but the amplitude increased, suggesting that the hotel’s occupancy began to stabilize.
The fluctuation period of IMF2 is 7 days, which explains the second-most variance to the model, which accounts for 18.9% of the origin variance and 25.0% of the decomposed data. Similar to IMF1, IMF2 also experienced two stages: an oscillation stage followed by a different amplitude stage. The difference is that the later oscillations of IMF2 became more drastic, and the degree of fluctuation was even more than those during the hotel’s initial operations. The fluctuation period of IMF3 is about 14 days, and interestingly, IMF3 contributes the fourth-most to the result while IMF4 with a period of 29 days contributes the third-most, suggesting that the period of 29 days was more significant than the 14 days’ cycle.

The variance contribution rate of IMF5 to IMF10 is small: the total being 4.3%, and the period ranging from 63 days to 1,296 days. This suggests that these IMFs are low-frequency, long periods, slow fluctuating, and have a limited effect on explanatory power for hotel occupancy. In other words, a long period has little effect on hotel occupancy forecasting, which also implies that the factors influencing occupancy are complicated. As stated by Huang et al. (2003), the trend term reveals the latent trajectory of the origin data. In the case of LMN, it decreases at first and then keeps increasing from 1,000 days onward because, as the trend evolves to a steady-state growth, the occupancy of LMN keeps increasing in the foreseeable future.
The correlation coefficient between the IMFs showed that the IMFs and the trend term as well as the real occupancy have significantly positive correlations (see Table 3), with the exception of IMF10. This suggests that the decomposition is effective, of which the correlation coefficients of IMF1 to IMF4, with the real hotel occupancy data, are more than 0.4, and therefore moderately significant positive correlations. The correlation coefficients between IMF5 to IMF10 for the trend and real occupancy are small, which suggests they are significant but weak correlations. The results of correlation coefficients are consistent with the analysis of variance contribution rates, suggesting that short-period cycles of less than a month are more meaningful.

**Forecasting with EEMD-ARIMA**

According to the analysis above, the forecasting steps are set to four periods within a month: 3 days, 7 days, 14 days, and 29 days. These correspond to the major period of hotel operations: half a week, a week, two weeks and four weeks. The origin data is divided into a training set and a testing set (see Table 4). Each of the four training sets is decomposed by EEMD, and the most popular forecasting method—ARIMA—is adopted as the benchmark. The top rows in Table 4 and Table 5 represent the original ARIMA models. The IMFs and trends are modeled and forecasted separately and integrated later; the forecasting results of ARIMA and EEMD-ARIMA are compared to determine if the new method can indeed improve the forecasting accuracy. The related
Figures 3 through 6 compare the forecast results, which show that the EEMD-ARIMA method clearly outperforms the individual ARIMA over three or seven days. Besides, the former can follow the directional change of real occupancy better than ARIMA method. However, the forecasts made by EEMD-ARIMA do not function very well beyond 7 days, as the forecast occupancy would keep increasing or decreasing. Further, the forecasting error becomes progressively larger. This indicates that EEMD-ARIMA may not be suitable for direct forecast of long-term occupancy rates.

From the analysis above, it seems that direct application of EEMD-ARIMA is not suitable for long-term forecasting. Thus, we introduce the rolling forecasting method. As an example, we take the next 30 days’ forecasts, which
can be divided into 10 three-day periods. The first three days’ occupancy is forecasted and then added to the training set in order to forecast occupancy for the next three days, and so on. The forecast results can be seen in Figure 7; the rolling forecasting method that divides longer periods into shorter periods achieved better results with both ARIMA and EEMD-ARIMA. Further, EEMD-ARIMA outperforms ARIMA again. It is important to note that, since ARIMA is not proficient in capturing turning points, a sudden decrease or increase in real occupancy will result in a significant error in the forecasts. On the other hand, EEMD-ARIMA outperforms ARIMA in capturing the turning points in the data series.

Conclusions

Individual hotels are basic business units of the hospitality industry, occupancy rate is a key indicator in hotel performance, and thus its forecasting is crucial for revenue management. Forecasting for a single hotel on a daily basis is especially difficult due to complex and erratic fluctuations (Jeffrey and Barden, 2000). We introduce a novel method, EEMD, to reduce the fluctuation and capture the regularities in the patterns of hotel occupancy. Eventually, a forecasting model integrated by EEMD-ARIMA has been presented.
The comparison between forecasting results of ARIMA and EEMD-ARIMA shows that the novel model performs better especially for short-term forecasts, whether measured by diagnostic parameters or the ability to predict turning points. This unique method can effectively handle data fluctuation by decomposing the complicated occupancy data into a set of relative regular waves. Therefore, an EEMD-ARIMA model is appropriate to forecast daily occupancy for individual hotels.

The practical implications of this study rest on its application in individual hotels. There is a lack of studies focusing on daily occupancy forecasting for single properties. EEMD is proficient in dealing with non-stationary and fluctuating data and has been popular in other disciplines. After the exploratory and empirical testing, we found that the introduction of EEMD can bring more accurate daily forecasting. This provides a more efficient and accurate way for individual properties in forecasting daily occupancy.

For an individual hotel, the within-week occupancy pattern can be complex and erratic, since its patterns will depend on the range and type of markets served. National holidays, local school holidays, occasional sporting events, one-off exhibitions or tourism promotions, unforeseen events, and even extreme weather conditions can produce day-to-day variations in occupancy rates. However, our results show that consistent patterns underlie this
complexity (Lomanno, 1995). It is towards the identification of these consistent and aggregate patterns, along with their effects in the daily occupancy of individual hotels, that this study contributes to literature.

However, one limitation of this model is its complexity. Parsimonious models are usually preferred in forecasting (Makridakis and Hibon, 2000). The complexity of EEMD-ARIMA will make it hard to explain to laypeople and thus limit its adoption and usage. In addition, this model could be improved further. The EEMD method does not have a satisfactory solution to deal with the ending effect due to EEMD envelopes. Since it’s unlikely that the IMFs’ ends are all at their maximum or minimum values, the upper and lower envelope would radiate at both ends, which would consequently affect IMFs and trends, and the forecasts would deviate from the actual trends. For example, when the current occupancy is a local extrema, the next forecasting step would increase or decrease as well. The ending effect has little influence on short-term forecasts, but it would exert great influence over long-term forecasts. This explains why the results of the 14-day or 29-day period’s direct forecasting are poor in Figure 5 and Figure 6. Further, the forecasting model cannot determine the turning point automatically, and, therefore, it is necessary to add human intervention or divide long-term forecasting into short periods to obtain accurate occupancy results. Future studies should examine these problems.
References


Chan CK, Witt SF, Lee YCE and Song H (2010) Tourism forecast combination using


Wu Z and Huang NE (2004) A study of the characteristics of white noise using the


## Tables and Figures

**Table 1. Summary of studies on hotel demand forecasting**

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Data type</th>
<th>Research area</th>
<th>Main methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>William et al. (1990)</td>
<td>Monthly</td>
<td>Individual hotel, USA</td>
<td>Time series model</td>
</tr>
<tr>
<td>Schwartz and Hiemstra (1997)</td>
<td>Daily</td>
<td>3 hotels, USA</td>
<td>Curves similarity model</td>
</tr>
<tr>
<td>Law (1998)</td>
<td>Annual</td>
<td>Hong Kong, PRC</td>
<td>Neural network</td>
</tr>
<tr>
<td>Rajopadhye et al. (2001)</td>
<td>Daily</td>
<td>Individual hotel, USA</td>
<td>Exponential smoothing model</td>
</tr>
<tr>
<td>Choi (2003)</td>
<td>Annual</td>
<td>USA</td>
<td>Economic indicator system</td>
</tr>
<tr>
<td>Lim et al. (2009)</td>
<td>Monthly</td>
<td>New Zealand</td>
<td>Exponential smoothing model</td>
</tr>
<tr>
<td>Zakhary et al. (2011)</td>
<td>Daily/Weekly</td>
<td>Individual hotel, Egypt</td>
<td>Monte Carlo simulation</td>
</tr>
<tr>
<td>Song et al. (2011b)</td>
<td>Quarterly</td>
<td>Hong Kong, PRC</td>
<td>Error correction method</td>
</tr>
<tr>
<td>Pan et al. (2012)</td>
<td>Weekly</td>
<td>Charleston, USA</td>
<td>Time series model</td>
</tr>
<tr>
<td>Yang et al. (2014)</td>
<td>Weekly</td>
<td>Charleston, USA</td>
<td>Time series model</td>
</tr>
<tr>
<td>Ellero and Pellegrini (2014)</td>
<td>Daily</td>
<td>5 hotels, Italy</td>
<td>Pick-up models</td>
</tr>
<tr>
<td>Koupriouchina et al. (2014)</td>
<td>Daily</td>
<td>Hotel group, Netherlands</td>
<td>System and manual</td>
</tr>
</tbody>
</table>
Table 2. Analysis of IMFs and TREND Based on EEMD

<table>
<thead>
<tr>
<th>Percentage of variance before decomposition (%)</th>
<th>IMF1</th>
<th>IMF2</th>
<th>IMF3</th>
<th>IMF4</th>
<th>IMF5</th>
<th>IMF6</th>
<th>IMF7</th>
<th>IMF8</th>
<th>IMF9</th>
<th>IMF10</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of variance after decomposition (%)</td>
<td>26.32</td>
<td>18.85</td>
<td>7.88</td>
<td>7.99</td>
<td>4.27</td>
<td>2.71</td>
<td>2.09</td>
<td>1.20</td>
<td>1.45</td>
<td>0.17</td>
<td>2.54</td>
</tr>
<tr>
<td>Period (days)</td>
<td>3.2</td>
<td>7.0</td>
<td>13.7</td>
<td>28.8</td>
<td>64.8</td>
<td>117.8</td>
<td>235.5</td>
<td>431.8</td>
<td>1295.5</td>
<td>2591.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IMF1</td>
<td>IMF2</td>
<td>IMF3</td>
<td>IMF4</td>
<td>IMF5</td>
<td>IMF6</td>
<td>IMF7</td>
<td>IMF8</td>
<td>IMF9</td>
<td>IMF10</td>
<td>Trend</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>P-value</td>
<td>.591*</td>
<td>.575*</td>
<td>.429**</td>
<td>.408**</td>
<td>.358**</td>
<td>.312**</td>
<td>.153**</td>
<td>.188**</td>
<td>.175**</td>
<td>.008</td>
<td>.127**</td>
</tr>
</tbody>
</table>

*the significant level of p-value is 0.05 (two-tailed);

**The significant level of p-value is 0.01 (two-tailed).
Table 4. The Size of Each Forecast Data Series

<table>
<thead>
<tr>
<th></th>
<th>3d</th>
<th>7d</th>
<th>14d</th>
<th>29d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>2588</td>
<td>2584</td>
<td>2577</td>
<td>2562</td>
</tr>
<tr>
<td>Testing set</td>
<td>3</td>
<td>7</td>
<td>14</td>
<td>29</td>
</tr>
</tbody>
</table>
Table 5. Adopted Structure and Parameters of ARIMA for 3- and 7-day Forecasts

<table>
<thead>
<tr>
<th>Model</th>
<th>Stationary Test</th>
<th>ARIMA (p, d, q)</th>
<th>Adjusted R²</th>
<th>R²</th>
<th>RMS E 3d</th>
<th>MAPE 3d</th>
<th>MAPE 7d</th>
<th>Normalized BIC 3d</th>
<th>Normalized BIC 7d</th>
</tr>
</thead>
<tbody>
<tr>
<td>signal</td>
<td>1</td>
<td>(1,0,14)</td>
<td>0.444</td>
<td>0.444</td>
<td>0.178</td>
<td>27.934</td>
<td>27.957</td>
<td>-3.435</td>
<td>-3.434</td>
</tr>
<tr>
<td>IMF1</td>
<td>1</td>
<td>(0,0,8)</td>
<td>0.234</td>
<td>0.234</td>
<td>0.107</td>
<td>423.965</td>
<td>424.269</td>
<td>-4.449</td>
<td>-4.448</td>
</tr>
<tr>
<td>IMF2</td>
<td>1</td>
<td>(3,0,6)</td>
<td>0.963</td>
<td>0.963</td>
<td>0.020</td>
<td>101.546</td>
<td>101.672</td>
<td>-7.820</td>
<td>-7.819</td>
</tr>
<tr>
<td>IMF3</td>
<td>0</td>
<td>(0,2,8)</td>
<td>0.962</td>
<td>0.999</td>
<td>0.002</td>
<td>293.956</td>
<td>294.593</td>
<td>-12.074</td>
<td>-12.073</td>
</tr>
<tr>
<td>IMF4</td>
<td>0</td>
<td>(1,4,8)</td>
<td>0.637</td>
<td>1.000</td>
<td>0.000</td>
<td>0.650</td>
<td>0.651</td>
<td>-17.589</td>
<td>-17.587</td>
</tr>
<tr>
<td>IMF5</td>
<td>0</td>
<td>(1,4,13)</td>
<td>0.761</td>
<td>1.000</td>
<td>0.000</td>
<td>0.333</td>
<td>0.333</td>
<td>-24.431</td>
<td>-24.430</td>
</tr>
<tr>
<td>IMF6</td>
<td>0</td>
<td>(2,4,1)</td>
<td>0.867</td>
<td>1.000</td>
<td>0.000</td>
<td>0.005</td>
<td>0.005</td>
<td>-30.960</td>
<td>-30.958</td>
</tr>
<tr>
<td>IMF7</td>
<td>0</td>
<td>(1,5,8)</td>
<td>0.159</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-35.944</td>
<td>-35.942</td>
</tr>
<tr>
<td>IMF8</td>
<td>0</td>
<td>(0,5,4)</td>
<td>0.782</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-39.941</td>
<td>-39.940</td>
</tr>
<tr>
<td>IMF9</td>
<td>0</td>
<td>(3,3,7)</td>
<td>0.935</td>
<td>1.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>-39.417</td>
<td>-39.415</td>
</tr>
<tr>
<td>IMF10</td>
<td>0</td>
<td>(2,3,6)</td>
<td>0.781</td>
<td>1.000</td>
<td>0.000</td>
<td>0.010</td>
<td>0.000</td>
<td>-39.097</td>
<td>-39.095</td>
</tr>
<tr>
<td>Trend</td>
<td>0</td>
<td>(2,2,2)</td>
<td>0.484</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-42.549</td>
<td>-42.549</td>
</tr>
</tbody>
</table>
Table 6. Adopted Structure and Parameters of ARIMA for 14- and 29-day Forecasts

<table>
<thead>
<tr>
<th>Mode</th>
<th>Stationary Test</th>
<th>ARIMA (p,d,q)</th>
<th>Adjusted R²</th>
<th>R²</th>
<th>RMSE</th>
<th>MAPE</th>
<th>MAE</th>
<th>Normalized BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14d</td>
<td>29d</td>
<td>14d</td>
<td>29d</td>
<td>14d</td>
<td>29d</td>
<td>14d</td>
<td>29d</td>
</tr>
<tr>
<td>signal</td>
<td>(1,0,13)</td>
<td>(1,0,14)</td>
<td>.421</td>
<td>.444</td>
<td>.421</td>
<td>.444</td>
<td>.182</td>
<td>.17828</td>
</tr>
<tr>
<td>IMF1</td>
<td>1</td>
<td>(0,0,8)</td>
<td>.233</td>
<td>.231</td>
<td>.233</td>
<td>2.231</td>
<td>.107</td>
<td>.424</td>
</tr>
<tr>
<td>IMF2</td>
<td>1</td>
<td>(3,0,8)</td>
<td>.933</td>
<td>.962</td>
<td>.933</td>
<td>.962</td>
<td>.027</td>
<td>.020</td>
</tr>
<tr>
<td>IMF3</td>
<td>1</td>
<td>(0,2,8)</td>
<td>(2,0,8)</td>
<td>.890</td>
<td>.999</td>
<td>.996</td>
<td>.999</td>
<td>.004</td>
</tr>
<tr>
<td>IMF4</td>
<td>0</td>
<td>(1,4,8)</td>
<td>.637</td>
<td>1.000</td>
<td>.000</td>
<td>.653</td>
<td>.656</td>
<td>.000</td>
</tr>
<tr>
<td>IMF5</td>
<td>0</td>
<td>(1,4,13)</td>
<td>.761</td>
<td>1.000</td>
<td>.000</td>
<td>.033</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>IMF6</td>
<td>0</td>
<td>(2,4,1)</td>
<td>.867</td>
<td>1.000</td>
<td>.000</td>
<td>.005</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>IMF7</td>
<td>0</td>
<td>(1,5,8)</td>
<td>.159</td>
<td>.158</td>
<td>1.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>IMF8</td>
<td>0</td>
<td>(0,5,4)</td>
<td>.782</td>
<td>1.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>IMF9</td>
<td>0</td>
<td>(3,3,7)</td>
<td>.935</td>
<td>1.000</td>
<td>.000</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>IMF10</td>
<td>0</td>
<td>(2,3,6)</td>
<td>.781</td>
<td>1.000</td>
<td>.000</td>
<td>.010</td>
<td>.011</td>
<td>.000</td>
</tr>
<tr>
<td>Trend</td>
<td>0</td>
<td>(2,2,2)</td>
<td>.482</td>
<td>1.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>
Table 7. Forecasting Error Measures of the 3 Periods

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3d</td>
<td>7d</td>
<td>30d</td>
</tr>
<tr>
<td>ARIMA</td>
<td>0.194</td>
<td>0.196</td>
<td>0.057</td>
</tr>
<tr>
<td>EEMD-ARIMA</td>
<td>0.113</td>
<td>0.114</td>
<td>0.039</td>
</tr>
</tbody>
</table>
Figure 1: Annual Fluctuation of Occupancy

Occupancy from 2007 to 2014 is depicted, showing the fluctuation of occupancy over the years.
Figure 2. Ensemble Empirical Mode Decomposition of Occupancy Rate
Figure 3. Comparison of Forecasting (3 days)
Figure 4. Comparison of Forecasting (7 days)
Figure 5. Comparison of Forecasting (14 days)

- **real occupancy**
- **ARIMA**
- **EEMD-ARIMA**
Figure 6. Comparison of Forecasting (29 days)
Figure 7: Comparison of Forecasting (10 times over a three-day period)