Tourism forecasting by search engine data with noise-processing

Li Xiaoxuan1*, Wu Qi2, Peng Geng1 and Lv Benfu1

1School of Economics and Management, University of Chinese academy of sciences, China.
2Institute of Geology and Geophysics of the Chinese Academy of Sciences, China.

Received 9 October, 2015; Accepted 22 January, 2016

In many studies, search engine data were efficient to analyze and forecast as an explanatory variable, including the tourism volumes predictions. However, the search data and the tourism volumes were always interfered by the noise. Without noise-processing, the predictive ability of search engine data might be weak, even invalid. As a method of noise-processing, Hilbert-Huang Transform (HHT) could deal with non-linear and non-stationary data. This study proposed a model with denoising and forecasting by search engine data, namely CLSI-HHT. The search queries were composited into an index first, then the noise were extracted from the index and tourism volumes sequences by HHT. The study further forecast the tourism volumes with the effective series. The results demonstrated that CLSI-HHT model outperformed the baselines significantly while the index model without denoising performs nearly same as the time series model. Moreover, wavelet transform and filtering were compared with HHT on denoising and the results implied that HHT had higher signal noise ratio (SNR) and forecast more accurately. The study concluded that noise-processing was necessary for the tourism forecasting with search engine data, and HHT could be an effective method on denoising.

Key words: Hilbert-Huang Transform (HHT), search engine data, noise-processing, wavelet transform.

INTRODUCTION

In the past few decades, the tertiary industry has developed fast with the driving force of macroeconomics. Leisure entertainment demands are increasing while tourism is one of the most important components. However, tourism development could present a significant challenge for the nature, environment and scene management as well as boom the local economy. Tourists exploding and the upsurge of holidays both make it difficult to manage the scene and distribute the limited resource appropriately. The perishable nature of tourism products make forecasting an important subject for future success (Gunter and Onder, 2015), especially the peak forecasting. Meanwhile, the internet changes the decision process and behavior. The search engines are able to capture and record the online behavior of netizen. Many previous studies have effectively analyzed and forecast the social economics with search engine data, including tourism volumes, stock prices, countries’ risks and...
exchange rate (Smith, 2012; Zhang et al., 2013; Fondeur and Karame, 2013). However, despite predicting tourism volumes by web data, not many studies have investigated the noise-processing before predictions. Either the tourism volumes or the search engine data, both of them are easy to be interfered by the noise from outside. This noise will be contained in the statistical sequences, which could impact on the analysis of this study. The actual behavior and decisions may be misunderstood by the jamming of noise, which will further fail the tourism forecasting. All these characteristics of tourism products and web data make noise-processing a necessary issue for both academics and practitioners.

In signal processing and physics, spectral analysis is widely utilized to deal with the noise. Fourier transform combines the signal features of time and frequency, and it respectively analyzes in the time and frequency domain. But Fourier is only defined on the natural domain of the global integration and cannot reflect the signal spectrum characteristics of the local time. Both of them make Fourier transform contradictory between the time domain and the frequency domain in analyzing the signals. Moreover, Fourier transform requires the signals to be stationary and have obvious difference in the spectrum characteristics of the noise.

These preconditions are always dissatisfied for sequences in tourism and other social science objectives. Wavelet transform recovers some restrictions of Fourier transform. It can handle the non-stationary signals and present the distinct features of the effective signals compared with the noise (Huang, 1998). Hilbert-Huang transform (HHT) is new developed time-frequency analysis in recent years and can deal with the non-stationary and non-linear signals freely. HHT absorbs the advantages of multiresolution of Wavelet without choosing the Wavelet basis. All these characteristics make HHT suitable on filtering and denoising. Two main parts complete HHT: Empirical Mode Decomposition proposed by Huang and Hilbert spectral analysis (Huang et al., 1998).

Empirical mode decomposition (EMD) is an adaptive signal decomposition by their own characteristic time scales. Compared to Wavelet transform and Fourier transform, EMD has higher signal noise ratio (SNR) in dealing with non-linear and non-stationary series, and it is widely applied in noise-processing and forecasting. EMD decomposes the original signal into several Intrinsic Mode Functions (IMFs) and a residue. Each IMF contains the local characteristics of the original signal in different time scales (Chen et al., 2012). IMFs, derived from EMD, satisfy the narrow-band requirements of Hilbert transform and can be further calculated to obtain the Hilbert instantaneous frequency and the power spectrum. The study preprocess the noise of tourism volumes and search engine data by HHT, then forecast the tourism volumes and analyze the role of denoising in tourism forecasting with search engine data.

LITERATURE REVIEW

Forecasting with search engine data

Search engine data offers a new source for the behavioral analysis. As the antecedent actions of decision-making in real life, online search can reflect users’ trends in the future. Search engines have become the main sources for netizen to gather information and are helpful for researchers to find the behavior patterns and decision-making process.

As early as 2009, Ginsberg et al. (2009) successfully predicted flu outbreaks in the United States with Google Search, then the studies with search engine data have sprung up in the next decades. The scholars applied search engine data to predict economics (Vosen and Schmidt, 2011; McLaren and Shanbhogue, 2011; Dzielinski, 2012; Fondeur and Karame, 2013), finance (Bollen et al., 2011; Bordini et al., 2012; Smith, 2012; Zhang et al., 2013), disease control (Rothberg et al., 2014) and so on. In Japan, Takeda and Wakao (2014) collected and recorded search queries of 189 stocks in Google Trends as the online search intensity. They then analyzed the relationship among search data, stock trade volumes and stock returns, finding that Google Trends data had a great positive effect on stock trade volumes, enabling them to predict the trade returns in the next periods.

Unlike the other studies to predict the future with the present search engine data, Varian and Choi (2009) claimed the important role of Google Trends data in predicting the present rather than the future to make up the lagging calculation of the statistical data. They combined the Google search data and time series to predict the volumes in several different industries, including retail sales, automobile sales, home sales and travel.

Meanwhile, Search engines extend the information dissemination channels for tourism industry and altered the way that people gather travel information (Beldona, 2005; Buhals and Law, 2008). Gawlik et al. (2011) based on the research of Varian and Choi (2009) used query-specific search data to forecast the monthly tourism volumes in Hong Kong between 2005 and 2010. They extracted the features and chose the highest relevant queries of key words, then evaluated the performance of the prediction with k-fold cross validation. It turned out that the forecast result was better than the study of Varian and Choi (2009) in either the training error or forecast error. The research of Bangwayo-Skeete and Skeete (2015) gathered the composite queries of “hotel and flight” and applied Autoregressive Mixed-Data Sampling (AR-MIDAS) model. Contrast with seasonal autoregressive integrated moving average (SARIMA) model and autoregressive (AR) model, AR-MIDAS model performed better in most out-of-sample forecast tests. The result also demonstrated that search data had a
significant benefit in predictions. Song et al. (2013) proposed a web-based tourism demand forecasting system (TDFS), including the data module, the quantitative forecast and forecast evaluation. They also uniquely combined the quantitative prediction and judgmental forecast together in the TDFS.

Most analysts forecast with data from Google Trends. But Baidu owns overwhelming users in comparison to Google search engine in China. There is a growing trend after Google quit the Chinese market. Since 2006, Baidu has opened its search database and published "Baidu Index" online, which enable researchers to forecast the tourism trends and other social issues with Baidu Index. Yang et al. (2015) compared Google and Baidu search data to predict the tourism volumes in Hainan Province of China. The indicators suggested that Baidu Index outperformed the Google Trends and the benchmark ARMA model due to the overwhelming users for Baidu search engine in China. Vaughan and Chen (2015) forecast the quality of Chinese universities and companies by using Google and Baidu search data respectively and obtained similar conclusion. Huang et al. (2013) predicted the tourism flow in the Forbidden City during the "golden week" by using Baidu Index and autoregressive distribution lad (ARDL) model. They claimed that Baidu index data and ARDL model could produce the more accurate forecast results than the traditional ARMA model.

HHT in the noise processing

The forecast ability of the web search data has been confirmed by many previous studies. However, either web search data or the social society object contains some inevitable noise. The predictions become more difficult because of this noise. Signal science mainly processes the noise by applying spectrum analysis which maps the signals in frequency domain. The most commonly methods include Fourier Transform and Wavelet Transform. As mentioned earlier, Fourier and Wavelet transform have too many restrictive conditions to suit the social science. HHT was firstly proposed by Huang et al. (1998) and consisted of two parts: EMD and Hilbert Transform. Since then, HHT has begun to be widely used in the field of building structure (Roveri and Carcaterra, 2011), mechanical fault (Bin et al., 2011; Wu et al., 2014), medicine (Yan and Lu, 2014) and geophysics (Ni et al., 2013). Song et al. (2012) applied HHT and threshold to denoise the electrocardiogram (ECG) signal. EMD decomposed the ECG signal into the noise layers and the effective signal layers. The noise layers were calculated by the energy spectrum and high frequency average periods. The experiment results implied that several main noise included in ECG signal were effectively identified and extracted by HHT and threshold method.

Recent studies attempted to introduce HHT into business management and have a remarkable effect. Ju et al. (2014) combined HHT and time series analysis to study the impact mechanism between stock prices of Chinese crude oil and macroeconomics. HHT was used to screen the target stocks for each event. The results turned out that HHT was valid as the screen method for event studies. Chen et al. (2012) and Yao et al. (2014) decomposed the tourism volumes signal and extracted features by EMD. They both found that EMD was a reliable method to deal with the non-linear and non-stationary signal, and could improve the predicted accuracy. Chen and Wei (2011) proposed a time variants exploration method that includes EMD and Hilbert spectral analysis (HSA) to extract the frequency features of Taipei short-term passenger flow. Also they compared the results of HHT with that of fast Fourier transform (FFT) and concluded that HHT could obtain the narrower frequency band, accurately capture time-frequency-energy distribution and help to enhance the performance of transportation systems.

Overall, there was one main limitation in the previous studies, related to forecasting with search engine data. Most of these studies have focused on the predicted ability and the relationship between web search data and social society. When the web search data and forecast objects includes a large number of noise signals, the predicted ability and the correlations might be weakened and misled. How to deal with this noise signal before the study analyze and predict by web search data. Can HHT denoise the web data and improve the predicted ability effectively? The study proposes a novel method of denoising the web data and the predicted object to improve the forecast accuracy, especially during the peak periods.

METHODOLOGY

Empirical mode decomposition

As the first stage of HHT, EMD is a signal analysis method which can deal with non-linear and non-stationary data. The basic principle of EMD is to decompose the time series into a sum of oscillatory functions, namely intrinsic mode functions (IMFs). Every IMF contains the local features in different time scales. Besides, the IMF must satisfy the following two basic conditions:

1. In the whole function time range, the number of the local extrema (including the maxima and the minima) and the number of zero crossing points must be equal, or differ only by one;
2. At any time, the local average is zero.

The first condition is similar to the traditional narrow band requirements for a stationary Gaussian process (Chen et al., 2012). This condition chooses the local requirement to avoid the unnecessary fluctuation effects. The latter condition is more complicated. The envelope of the maxima forms the upper envelope, and the envelope of the minima forms the lower envelope. The mean of the upper and lower envelopes is set to zero, in order to make sure that the signal wave is a local symmetry.
The main steps of EMD can be broken down into two steps:

**Step 1**: Take the experimental data as \( Y(t) \) and calculate all the local extrema. Then connect all the maxima into the upper envelope line \( U(t) \) and all the minima into the lower envelope line \( L(t) \). All the data will be contained between \( U(t) \) and \( L(t) \) and the mean of these two envelopes are

\[
M_i(t) = (U(t) + L(t))/2
\]  

Eliminate \( M_i(t) \) from the original series \( Y(t) \) and obtain the first line \( Q_1(t) \), which is the IMF:

\[
Q_1(t) = Y(t) + M_1(t)
\]  

However, if \( Q_1(t) \) is unsatisfied to the two conditions, it cannot be called as the IMF. Step 1 process is to be repeated till the series \( Q_k(t) \) meets the conditions.

\[
Q_{1k}(t) = Q_{1(k-1)}(t) + M_{1k}(t)
\]  

Then the residue can be treated as the trends of the original time series.

\[
Y(t) = \sum_{i=1}^{n} IMF_i + R_n
\]

If the residue \( R_n \) is omitted, the original signal \( Y(t) \) can be presented as:

\[
Y(t) = \text{Re} \sum_{i=1}^{n} a_i(t) e^{j\phi_i(t)} = \text{Re} \sum_{i=1}^{n} a_i(t) e^{j\frac{d\phi_i(t)}{dt}}dt
\]

**Step 2**: Subtract IMF1 from the original series,

\[
P_1(t) = Y(t) + IMF_1
\]

Then the study takes \( P_1(t) \) as the original time series and iteratively perform the whole step 1 to decompose other IMFs. This process ends till the residue becomes a constant or a monotonic function. The residue can be treated as the trends of the original time series.

**Hilbert transform and spectrum analysis**

Then the second stage of HHT is Hilbert spectral analysis, which is performed to obtain the time-frequency-energy distribution. The IMFs, decomposed by EMD, perfectly satisfy the narrow band requirements of Hilbert transform. Set each IMF as \( c_i \) and structure the analytic signal for \( c_i \):

\[
z_i(t) = c_i(t) + j\dot{c}_i(t) = a_i(t)e^{j\phi_i(t)}
\]

The amplitude function, the phase function and the instantaneous frequency are:

\[
a_i(t) = \sqrt{c_i^2(t) + \dot{c}_i^2(t)}
\]

\[
\phi_i(t) = \arctan(\dot{c}_i(t) / c_i(t))
\]

\[
f_i(t) = \frac{1}{2\pi} \frac{d\phi_i(t)}{dt}
\]

**Empirical test**

**Data**

As one of the famous scenic spots in China, Jiuzhaigou attracts tens of thousands visitors every year, which leads to the potential problems of safety and being stranded for tourists. Particularly, the peak seasonal management is much more difficult for the authorities and policy makers. On October 1st, 2013, the first day of the National Day holiday, four thousands tourists were stuck at the entrance of Jiuzhaigou for five hours due to overcrowding (Qiu, 2013).

This signifies the urgency for accurate prediction for anticipating and managing influxes of tourists during the peak seasons. Moreover, in the network search rankings of Chinese scenic areas, Jiuzhaigou is far ahead and gets the most concern of tourists among the most popular scenic spots. This demonstrates that tourists prefer to collect information and arrange their travel plans with Jiuzhaigou being an attractive site for tourists. As for thesearch engine, Baidu overwhelms other search engines in China. Lots of studies proved that Baidu search data is much better and appropriate than Google and other search engines in analyzing
Regarding these aspects, the study proposes a forecast model with noise-processing, namely CLSI-HHT, to predict the tourism volumes based on search engine data and empirically test the model by taking the case of Jiuzhaigou.

The tourism traffic data of Jiuzhaigou is published by its official website every day. In the empirical study, the time span is from 1 June, 2012 to 31 December, 2014, including 944 data in total. During the empirical test, the series will be divided into two sets: training data and testing data. To achieve more reliable and accurate results, a long period is chosen as the training period (Chen et al., 2012). Based on this, the first 760 data spanning from 1 June, 2012 to 31 June, 2014 (80% of the total sample points) are used as the training sample while the remaining 184 data spanning from 1 July, 2014 to 31 December 2014 (20% of the total sample points) are treated as the testing sample. This testing period covers the peak tourism season (from August to October) and the off season (November and December) in Jiuzhaigou.

**Composite leading search index**

Baidu index of each query is in absolute numbers which reflects the search times. CLSI contains four main steps:

1. According to the tourists’ information demand, the study initially choose 15 basic search keywords, such as “Jiuzhaigou”, “the weather of Jiuzhaigou”, “the hotels in Jiuzhaigou”, “Jiuzhaigou airport”(in Chinese) and so on. Input these basic keywords to Baidu search engine, then the search engine will recommend more relevant queries based on your search query. This helps to expand the tracking scope of search terms. For example, when we search “Jiuzhaigou”, Baidu search engine recommends the queries including “the travel guides of Jiuzhaigou”, “travel agents in Jiuzhaigou”, “the ticket in Jiuzhaigou” and so on.

2. Taken these 15 basic queries as the seed search queries, the recommended queries are recaptured from Baidu Index. This round is repeated to obtain iteratively the search queries and ends till no new queries are recommended. Delete the queries without query traffic or without embodied. Only 146 search queries are retained in all.

3. Calculate the Pearson correlation coefficient among the search queries and Jiuzhaigou tourist volumes with different lag periods to identify the leading search queries. In total, 32 correlation coefficients are calculated among the search queries and Jiuzhaigou tourist arrivals, 0 to 31 days ahead, respectively. Based on these coefficients, the study further chooses the queries with the maximum correlation values in the modeling process. Confirm the threshold in order to reserve the appropriate number of the leading search queries. Excessive leading search queries contain more noise and useless information, while deficient leading search queries lose too much information. Only 6 search queries is selected if the threshold is 0.8. And if we choose 0.6, more than 60 search queries satisfy the standard. With the threshold of 0.7 and at least one lag period prior to the tourist arrivals, a total of 24 search queries are selected appropriately.

4. The study further combines the search index with moving summation. Each of the leading queries will be shifted according to the lag order of the maximum Pearson correlation value. Then sum up all of the moved queries to obtain the search index, namely INDEX7. Figure 2 presents the correlations between INDEX7 and Jiuzhaigou tourist volumes. Two series have similar change trend, which generally implies the possibility of forecasting the tourism volumes with composite search index.

**HHT in noise-processing**

On the basis of CLSI, EMD is conducted to the Jiuzhaigou tourism volumes and INDEX7. The decomposition derives out eight IMF functions and a residue of Jiuzhaigou tourist volumes and seven IMF functions and a residue of INDEX7. Figure 3 and Figure 4 respectively presents the series of all the component functions and the residues.

EMD adaptively decomposes the signal into several IMF components and a residue. Each IMF function contains the different characteristics time scales which presents the feature information of
Figure 2. The correlations between IDENX7 and Jiuzhaigou tourist volumes.

Figure 3. The IMFs and residue of tourist volumes.
the signal with multiresolution. Specifically, the resolution size varies with different signals. In Figure 3 and 4, IMF1 has the highest frequency. The frequency of all IMF functions is decreased in turn and non-overlapping while the residue is closed to a monotone function. The horizontal axis represents the time range, and the vertical axis shows the function values of each IMF.

Furthermore, the Hilbert transform is conducted to obtain the instantaneous frequency for IMFs of tourism volumes. The results are illustrated in Figure 5. As discussed earlier, the Hilbert spectrum represents the energy distribution of the time series data in both frequency and time scale. It provides more information about amplitude variants of measured time series data. In Figure 5, the instantaneous frequency of IMF1 has no obvious frequency band. The value distributes uniformly in the high frequency range and IMF1 is identified as the noise layer. The instantaneous frequency gradually reduces and tends to be smooth from IMF2 to IMF8.

The main frequency of IMF2 is located at 0.15 cycles per day (7 days per cycle), which demonstrates that there are 7 days period in the signal’s feature of Jiuzhaigou tourism volumes. This is consistent with the weekend effect of Jiuzhaigou. The main frequency of IMF3 appears in 0.07 cycles per day (14 days per cycle) and could be treated as the harmonic wave of IMF2. IMF4 have the similar main frequency of 0.03 cycles per day (about 30 days per cycle). This IMF reveals the monthly period of the tourism of Jiuzhaigou. The tourist traffic increase or decrease much similarly month to month. The main frequency of IMF5-IMF7 is located at around 0.01 cycles per day (100 days per cycle). This illustrates the seasonal period of the tourist flow.

Particularly, this seasonal characteristic is significant because of the high altitude of Jiuzhaigou (nice and cool in summer and snowed in winter). The lowest main frequency of IMF8 is 0.003 cycles per day (about 360 days per cycle). The residue is almost the monotone function, and also explain the long-term trend of the series. Based on the results of EMD and Hilbert spectrum analysis, IMF1 will be extracted as the high-frequency noise. The rest of IMF components and the residue are aggregated as the effective signal of Jiuzhaigou tourist volumes.

Similarly, Hilbert instantaneous frequency of individual IMF1-IMF7 is illustrated in Figure 6 for INDEX7. The instantaneous frequency of IMF1 has no obvious frequency band and distributes uniformly in the high frequency range. So IMF1 is identified as the noise layer. The rest of IMF components and the residue are
Training models

In the tourist traffic predictions, most scholars choose to apply time series models and econometric models. The study forecast the tourist volumes of Jiuzhaigou with the search engine data on the basis of considering and extracting the noise. The effective signal layers, obtained from HHT, are used to fit the models and forecast. Meanwhile, the time series model of Jiuzhaigou tourist volumes, the predicted model of web search data without denoising and BP neural network are chosen as the baseline models.

\[
visitor_t = c + \alpha_1 visitor_{t-4} + \alpha_2 visitor_{t-7} + \mu_t \quad (13)
\]

\[
visitor_t = c + \alpha_1 Index7_t + \mu_t \quad (14)
\]

\[
visitor_{t-1} = c + \alpha_1 Index7_{t-1} + \mu_t \quad (15)
\]

In equation (13) to (15), \( visitor_t \) denotes the Jiuzhaigou tourist volumes. \( index7_t \) denotes the index sequence \( INDEX7 \). Equation
13) and (14) are the baseline models, denoted M(1) and M(2). visitor_lt and Index7_lt denote the effective series of tourist volumes and INDEX7, respectively. Stationary tests (unit root tests) with Augmented Dickey-Fuller test method (ADF) are performed for all four variables. visitorLt and Index7Lt are stationary sequences. visitor_lt and Index7_lt are non-stable but the first difference are. Two pairs of time series from M(1) and M(2) are co-integration series with the same order. This supported Granger causality analysis and ARMAX (Autoregressive Moving Average with External Variables) models. Equation (15), denoted M(3), is the predicted model after the noise-processing of HHT. Thus, the regression models are constructed based on equation (13) to (15) and illustrated in Table 1.

The coefficients of all explanatory variables are significant at the 5% level. Both Index7 and Index7_Lt are statistically significant at the 1% level. The residue test implies that there is sequence correlations. So ARMA adjustment is conducted based on the regression. In Table 1, M(2) with the search engine data performs better than the time series model of M(1). Further, M(3) with noise processing with HHT overwhelms M(1) and M(2). All three models can be expressed as equation (16) to (18) respectively.

$$\text{visitor}_t = 5.193 + 0.785\text{visitor} _{t-1} + 0.161\text{visitor} _{t-7} + \mu_t$$

$$\mu_t = \epsilon_t + 0.103\epsilon_{t-1}$$

(16)

$$\text{visitor}_t = -44.306 + 1.296\text{Index7}_t + \mu_t$$

$$\mu_t = 0.673\mu_{t-1} + \epsilon_t + 0.185\epsilon_{t-7}$$

(17)
Table 1. Regression comparison of Model 1 to 3.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Historical data</th>
<th>Index7</th>
<th>Noise processing with HHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>M(1)</td>
<td>M(2)</td>
<td>M(3)</td>
</tr>
<tr>
<td>visitor _t</td>
<td>0.785**</td>
<td>1.296***</td>
<td>1.177***</td>
</tr>
<tr>
<td>-</td>
<td>(31.216)</td>
<td>(16.846)</td>
<td>(13.762)</td>
</tr>
<tr>
<td>visitor _7</td>
<td>0.161***</td>
<td>c</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>(6.996)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>c</td>
<td>5.193***</td>
<td>AR(1)</td>
<td>AR(1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.673***</td>
<td>2.011***</td>
</tr>
</tbody>
</table>

Independent variables:

- visitor \_t \ -1 0.785**
- Index7 1.296***
- Index7 \_t 1.177***
- visitor \_7 0.161***
- Index7 1.296***
- Index7 \_t 1.177***
- visitor \_7 \ -1 0.161***
- Index7 1.296***
- Index7 \_t 1.177***
- c 5.193***
- AR(1) 0.673***
- AR(1) 2.011***
- MA(1) 0.571***
- MA(1) 0.571***
- MA(1) 0.571***
- AR(1) 0.673***
- AR(1) 2.011***
- MA(1) 0.571***
- MA(1) 0.571***
- MA(1) 0.571***
- AR(1) 0.673***
- AR(1) 2.011***
- MA(1) 0.571***
- MA(1) 0.571***
- MA(1) 0.571***

R^2 0.831 0.847 0.994
DW 1.99 1.95 1.91
1% -3.439 -3.849 -3.495
5% -2.659 -2.887 -2.8897
10% -2.569 -2.581 -2.5819

() shows the t-value; *, ** and *** means that it’s remarkable on the level of 10, 5 and 1% respectively.

Table 2. Performance of the two forecasting methods.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE</td>
<td>[ \frac{1}{n} \sum_{i=1}^{n} \left</td>
</tr>
<tr>
<td>RMSE</td>
<td>[ \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_i - Y_i)^2} ]</td>
</tr>
</tbody>
</table>

\( P_i \) presents the predicted values and \( Y_i \) presents the original values.

\[
\begin{align*}
\text{visitor}_t &\approx -30.461 + 1.177 \text{Index7}_t + \mu_t \\
\mu_t &\approx 2.011 \mu_{t-1} - 1.707 \mu_{t-2} + 0.627 \mu_{t-3} + \varepsilon_t + 0.571 \varepsilon_{t-1}
\end{align*}
\] (18)

The regression results show that the search engine data have a positive impact on Jiuzhaigou tourist volumes. The increase of Index7 will cause more tourists accordingly. The reason is that the increase of search engine data indicates more attention of the destination from the tourists. First, the tourists prepared for the upcoming planned travel and gather information in advance. Second, some potential tourists start to pay attention on Jiuzhaigou and it is possible to practice in the future. In equation (17), when Index7 doubles, the future tourism volumes will increase 1.296 times. This coefficient is more than 1 because there are package tours and they have no need to search on the internet. In equation (18), the positive impact weakens and the influence coefficient is 1.177. This explains that in M(2) the impacts are overestimated, calling “Big Data Arrogance”. By contrast, M(3) pre-processes the noise in the search index and the forecast object and adjusts this overestimation.

Forecasting with the models

From the results of the training models, M(2) of forecasting with search engine data performs better than the time series model M(1). Moreover, M(3) of denoising by HHT with search engine data overwhelms the baselines of M(1) and M(2). Based on the training models, the study further forecast the tourist volumes of Jiuzhaigou during the forecast period, including the baselines of M(1), M(2), BP neural network and M(3).

Many different quantitative statistical metrics are applied in evaluating the forecast performance, including Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). Both assessment criteria have been predominately applied in forecast literature (Song and Li, 2008; Chen, 2011; Bangwayo-Skeete and Skeete, 2015). Table 2 shows the metrics and calculations of MAPE and RMSE.

There have been many studies applying the artificial neural networks (ANN) to forecast the tourism demands (Cho, 2003; Palmer et al., 2006; Chen et al., 2012), due to their advantages in
Table 3. The predicted performances of the four models in next seven days.

<table>
<thead>
<tr>
<th>Variable</th>
<th>2014/7/1</th>
<th>2014/7/2</th>
<th>2014/7/3</th>
<th>2014/7/4</th>
<th>2014/7/5</th>
<th>2014/7/6</th>
<th>2014/7/7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original values</td>
<td>ARMA</td>
<td>Index7</td>
<td>BP neural network</td>
<td>CLSI-HHT</td>
<td>153.19</td>
<td>147.02</td>
<td>148.08</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>ARMA</td>
<td>Index7</td>
<td>BP neural network</td>
<td>CLSI-HHT</td>
<td>7.62</td>
<td>6.52</td>
<td>6.29</td>
</tr>
<tr>
<td>RMSE</td>
<td>ARMA</td>
<td>Index7</td>
<td>BP neural network</td>
<td>CLSI-HHT</td>
<td>8059.863</td>
<td>7843.92</td>
<td>5012.02</td>
</tr>
</tbody>
</table>

Table 4. The predicted performances of the four models during the peak periods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original values</td>
<td>-</td>
<td>7913.52</td>
<td>8002.25</td>
<td>5139.94</td>
</tr>
<tr>
<td>ARMA</td>
<td>7378.071</td>
<td>8059.863</td>
<td>5012.02</td>
<td>7158.815</td>
</tr>
<tr>
<td>Index7</td>
<td>7525.11</td>
<td>7843.92</td>
<td>5283.43</td>
<td>7228.83</td>
</tr>
<tr>
<td>BP neural network</td>
<td>6839.11</td>
<td>7509.84</td>
<td>7579.4</td>
<td>5016.31</td>
</tr>
<tr>
<td>CLSI-HHT</td>
<td>7894.1</td>
<td>7970.34</td>
<td>5198.06</td>
<td>7433.35</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>ARMA</td>
<td>Index7</td>
<td>BP neural network</td>
<td>CLSI-HHT</td>
</tr>
<tr>
<td></td>
<td>4.91%</td>
<td>1.98%</td>
<td>2.79%</td>
<td>2.70%</td>
</tr>
<tr>
<td></td>
<td>5.10%</td>
<td>5.28%</td>
<td>2.41%</td>
<td>7.94%</td>
</tr>
<tr>
<td></td>
<td>0.25%</td>
<td>0.40%</td>
<td>1.13%</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

capturing subtle functional relationships within the data (Atsalakis and Valavanis, 2009). Unlike traditional statistical models, neural networks are data-driven and non-parametric models which can depict the non-linear functions without a prior assumption about the characteristics of the data (Haykin, 1999). When we fit and forecast with Back-Propagation (BP) neural network, various parameters should be settled, including the number of neurons, the training efficiency and the hidden-layer unit number and may affect the application of BP neural network. After repeated experiments, the study chooses three layers, 15 neurons in the hidden-layer and 0.01 of the training efficiency. The prediction results of CLSI-HHT model and three baselines in the first week of the test period are presented in Table 3. Also the monthly prediction results of the peak periods from July to October are presented in Table 4. The complete predicted values are shown in Appendix A.

In the first week of the test period, MAPE and RMSE of time series model is more than others. ARMA model with historical tourist volumes of Jiuzhaigou has the largest predicted error among all the models. Comparing search engine data model with BP neural network, the study found that M(2) with search engine data performs better than BP neural network in the next days. Yet the further 5 days of the first week, BP neural network predicts more accurately. This explains that search engine data have stronger predicted ability in the short terms. Tourists always search and confirm information in a few days before departure. From the complete predicted values of 184 days in the test period, BP neural network performs much better than search engine data model. Three reasons could help to explain. First, artificial intelligence methods are more self-adaptive and superiority than traditional statistical models. This result is agreed with most other research conclusions. Second, search engine data have strong timeliness, which make it difficult to forecast in the long run. This is the reason that Baidu Forecasting only reveal the predicted tourist traffic in the next three days. Third, without noise-processing, the noise constantly accumulates and leads to the impossibility of forecasting in the long term.

In Table 4, the monthly predicted values of each model are presented from July to October. These four months are at the peak season for Jiuzhaigou. From the predicted performance, CLSI-HHT obtains the closest predicted values with the original tourism volumes. The predicted error is below to 50 people and the MAPE is quite small. Especially in the forecasting of October, the monthly predicted error is only 0.06%. It implies that CLSI-HHT model could remarkably improve the forecasting error during the peak period of
As to the baselines, the web search data of Index7 almost outperform the other two baselines in the forecasting of the peak months. This proves the predicted ability of the search engine data again. The actual predictions of CLSI-HHT during the peak season could greatly benefit the authorities and the policy makers. This actual forecasting makes better distribution and management in advance.

In all four models, CLSI-HHT model significantly outperforms three baselines. Especially, after the noise-processing, the traditional econometric model with the efficient variables could obtain more accurate forecasting than artificial neural network. Figure 7 shows the original data of Jiuzhaigou and the predicted values of each model.

**COMPARISONS WAVELET TO HHT**

Wavelet transform

In the empirical study, Hilbert-Huang transform is utilized to denoise the series of Index7 and tourist volumes of Jiuzhaigou. CLSI-HHT model outperforms all three baselines. Nevertheless, in the signal science, there are other methods to deal with noise except for HHT. Fourier transform and Wavelet transform are both able to denoise the signals by spectrum analysis. As the multiple limitations of Fourier transform in social economics, Wavelet transform and high-pass filter method will be used to be compared with HHT.

Wavelet transform inherits and develops the ideas of the localization of the short-time Fourier Transform (STFT). Meanwhile, it overcomes the shortcomings that the window size cannot change with frequency. Wavelet can provide a variable “time-frequency” window with frequency and is an ideal tool in identifying the noise signals. The main characteristics is that it can multi-scale refine the signals through telescopic translation operations and self-adaptively satisfy the requirements of the signal analysis.

In this section, the study chooses the same data as in the empirical study. The frequency spectrum based on Wavelet transform for tourist volumes and Index7 are shown in Figure 8. The horizontal axis is the time range and the vertical axis is the period values. The inside of the U-shape curve represents 95% confidence interval. In Figure 8, the colors express the strength of the amplitude. The amplitude enhances from blue to red. In the Wavelet frequency spectrum of the original tourist volumes of Jiuzhaigou, the spectral band is significant in the period of seven days. This frequency band implies the 7 days period of tourist volumes. Above this band, there are no more obvious frequency bands, and the lower period is beyond recognition. Based on the frequency spectrum, the periods lower than 7 days are identified as the high-frequency noise. Similarly, the spectral band of 3 days cycles is significant and the lower period is unidentified. The periods lower than 3 days are treated as the noise signal.

**High-pass filter to denoise**

Based on the results of the Wavelet transform, the high-pass filter is conducted to extract the noise from the tourist volumes and Index7. Filtering extracts or removes some frequency signals from the original series while the high-pass filter decays or removes the low frequency and retains the high frequency and the sharp changes of signals. $M(x)$ represents the measurement mode and $H(x)$ is the filter function.

$$R(x) = M(x) * H(x)$$ (20)

The output function $R(x)$ is the convolution of the input functions $M(x)$ and $H(x)$. Some data will be consumed.
when we train the high-pass filter for tourist volumes and search index. In this study, the first 30 data and the last 30 data are consumed, and the remaining series span from 1 July, 2012 to 1 December, 2014. The training set starts on 1 July, 2012 and ends on 31 June, 2014 with 730 data (82.6% of the whole sample). The test set is from 1 July, 2014 to 1 December, 2014 with 154 data (17.4% of the sample).

After the high-pass filter, the study obtains the effective signals of tourist volumes and Index7 and their noise signal, respectively (Figure 9). In the sequence charts, the series with denoising maintain consistency with the original series while they are smoother. The black lines are the original series. The red lines are the effective
series after high-pass filter. The blue lines distribute around the vertical axis show the high-frequency noise.

**Predictions and comparisons**

M(1)-(3) are trained with the effective signals based on the high-pass filter. \textit{Index7}, denotes the composite searchindex, \textit{Index7\_hht}, denotes the effective series of Index7 with HHT and \textit{Index7\_wf}, denotes the effective series of Index7 with Wavelet transform and high-pass filter. Stationary test with ADF are performed for all variables. \textit{Index7}, \textit{visitor}, \textit{Index7\_wf}, and \textit{visitor\_wf}, are stationary. \textit{Index7\_hht}, and \textit{visitor\_hht}, are non-stationary. The pair of these series are co-integration series with the same order. The regression results of CLSI-HHT model and Wavelet-filtering model are presented in Table 5. All the variables are significant and the residual of all three models are significant on the 0.01 level. CLSI-HHT model fits better than Wavelet-filtering model and the search engine model without denoising.

Furthermore, the next 154 days tourist volumes are forecast based on these models to compare the performance of Wavelet-filtering model with CLSI-HHT model. The values of MAPE for the test set are illustrated in Figure 10. It is obvious that CLSI-HHT model outperforms Wavelet-filtering model in the predicted set except for the last week of November. Compared with Wavelet denoising, HHT has higher signal to noise ratio (SNR) for non-linear and non-stationary signals. Specially, Wavelet-filtering model performs worse than search engine model without denoising in the short term. This proves the strong timeliness of search engine data and ability of short-term prediction. Furthermore, during the peak period from July to October, CLSI-HHT has the smallest predicted error among three models. This implies that HHT is more effective than Wavelet in the peak season of the tourism volumes in denoising.

**CONCLUSION**

Since big data has boomed, many studies focus on the correlationhip analysis and prediction with web search
Table 5. Regression comparison of three models.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Index7</th>
<th>CLSI-HHT</th>
<th>Wavelet-filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index7</td>
<td>1.300***</td>
<td>1.229***</td>
<td>1.555***</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.74***</td>
<td>2.003***</td>
<td>2.534***</td>
</tr>
<tr>
<td>MA(7)</td>
<td>0.183***</td>
<td>-1.696***</td>
<td>-2.446***</td>
</tr>
<tr>
<td>AR(2)</td>
<td>-4.998</td>
<td>-29.426</td>
<td>-82.227</td>
</tr>
<tr>
<td>MA(1)</td>
<td>-</td>
<td>16.044</td>
<td></td>
</tr>
</tbody>
</table>

R² 0.848 - 0.994 - 0.983 -
AIC 9.629 - 6.326 - 6.514 -
SC 9.655 - 6.364 - 6.535 -
DW 1.95 - 1.90 - 1.87 -
Residue 1% -3.439 -3.849 -3.495
stationary 5% -2.865 -2.887 -2.889
10% -2.569 -2.581 -2.582

( ) shows the t-value; *, ** and *** means that it's remarkable on the level of 10, 5 and 1% respectively.

Figure 10. The MAPE of each predicted models.

data. The methods of how to composite the search data and how to capture the web data were constantly proposed. The web data reflects human behavior and intention which could be used to forecast the future behavior and decision-making. However, these data always contains much noise which may mislead the forecast and the data further analysts, the policy makers and the managers. It is necessary to pre-process the noise before data mining. There are many methods of denoising, and HHT has its own improved algorithm and covers the lacks of Fourier and Wavelet transform. Taking Jiuzhaigou as an example, this study analyzed the predicted performance of Baidu search engine data with denoising by HHT, particularly during the peak period. We further compared Wavelet-filtering with HHT on the predicted performance of tourist volumes.
Compared to previous studies on using search engine data to forecast tourist volumes, this study makes two main contributions in theory and application. First, HHT is widely utilized in physics, engineering and geophysics. For social science, especially for tourism management, this study enriched the application. Second, this study analyzed the methods of denoising for web search data and tourist volumes, including HHT and Wavelet-filtering. If we ignore the noise-processing, the results of data mining may mislead the decisions and analysis. Especially, during the peak seasons there are higher risk of congestions for the tourism destinations. The deflection prediction may mislead the policy makers to distribute the resources of the scenic spot incorrectly. The managers may implement the inappropriate marketing. Noise-processing with HHT improves the efficiency of policy makers and managers when they use the web search data. This is particularly crucial in the peak seasons. In this study, CLSI-HHT forecasting model was able to improve the prediction accuracy and controlled the number of prediction error below one hundred people. The study composited the search engine data by CLSI to omit the irrelevant queries and noise preliminarily. Then the noise was extracted from the composite index and tourist volumes by HHT. The forecast results demonstrated that the search engine model of noise-processing with HHT improved the predicted accuracy remarkably. Besides, the search engine data without denoising performs almost the same compared with time series model. BP neural network performed better than the untreated search engine model but much worse than CLSI-HHT model. The study then applied Wavelet transform and high-pass filter to denoise the series and compared with HHT. The results turned out that HHT performed better than Wavelet transform in dealing with non-linear and non-stationary signals. However, this study had a number of limitations. In the empirical test, the study only takes Jiuzhaigou as the example. Whether CLSI-HHT method could be efficient in other forecast objects needs further researches. Besides, HHT could deal with the noise in forecasting the tourist volumes with search engine data. The micro blog and online news are the web data, either. They have the different features and data structure compared to search engine data. Apply HHT to deal with these data source would also be a useful research direction.

ACKNOWLEDGEMENTS

This paper is supported by National Natural Science Foundation of China (Grant No.71172199 and 71202115).

Conflicts of interests

The authors have not declared any conflict of interests.

REFERENCES


