CHEMICAL ENGINEERING TRANSACTIONS

VOL. 46, 2015

A publication of AIDIC

The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Peiyu Ren, Yancang Li, Huiping Song

ISBN 978-88-95608-37-2; **ISSN** 2283-9216

DOI: 10.3303/CET1546071

Tourist Arrivals Real-time Prediction Based on IOWA-Gauss Method

Lin Chen, Maozhu Jin*, Yonghuan He

Business school, Sichuan University, Chengdu, 610000, China.
jinmaozhu@scu.edu.cn.

Currently, the forecasts research focuses on tourism in a tourist trends and tourists influencing factors. Although the forecast for the inter-annual and seasonal tourists has a wealth of research results, there is less study of everyday and real-time tourist arrivals. This paper analyzes the tourists' real-time arrival law of Jiuzhaigou, and then the use of hierarchical clustering and Gaussian fitting mathematical methods for data processing, study the changes of day arrivals of tourist by segment analysis, and proposed a new model for the prediction of tourists in real-time arrivals. We take Jiuzhaigou Valley as an example to analysis, and experimental results show that the forecast method is effective.

1. Introduction

The rapid development of the tourism industry has promoted local economic development, at the same time, over-exploitation and the too large tourism scale brings enormous pressure to ecological and environmental protection. Balance visitors scale, control of temporal and spatial distribution of tourists has become an important content of scenic area visitor management in peak travel period. However, tourists distribution are affected by season, time, weather, tourist type and other factors, so there is a big uncertainty. There are various methods about forecasting tourist arrival, for example, Cho (2003)find that the artificial neural networks to forecast tourist arrivals perform well in comparison to the exponential smoothing ;Gil-Alana (2005)forecast international monthly arrivals by using seasonal univariate long-memory processes ;Chu (2008) uses fractionally integrated ARMA models to forecast tourism arrivals; S Chen (2010) apply ANFIS model to forecast the tourist arrivals to Taiwan; H Song (2011) forecast quarterly tourist arrivals by using a new model ,the TVP-STSM. in addition, they also include ARIMA (Cho, 2003; Cang, 2011; Wan & Wang, 2013), GARCH (Bollerslev,1986; Kim & Wong, 2006; Coshall, 2009), SSA (Beneki & Eeckels, 2012), but we cannot find any paper adopting a model to forecast tourist arrivals of real-time dynamic in the scenic area. This is need to study real-time dynamic prediction of tourist number in the scenic area by Spatiotemporal analysis of tourism carrying capacity. So, the purpose of this paper is to fill this gap, this paper combine GAUSS algorithm with IOWA and build a scenic area real-time prediction model, segmentationly analyse the change of visitors on the scenic area, which helps to improve the prediction method of the visitor number.

2. Model construction

This paper collects data through field investigation. Firstly process the data, find a general law of tourist arrivals. That is, the total number is gradually increasing, and the increasing speed is gradually slowing down, finally the number tends to be stable. Next cluster data by using hierarchical clustering method according to size and predict by using the Gaussian fitting algorithm. Then modify prediction model using weight calculation based on the improved IOWA operator. Obtain the final prediction results. Finally, analyse the prediction results. Figure 1 is the research model of this paper.

Please cite this article as: Chen L., Jin M.Z., He Y.H., 2015, Tourist arrivals real-time prediction based on iowa-gauss method, Chemical Engineering Transactions, 46, 421-426 DOI:10.3303/CET1546071

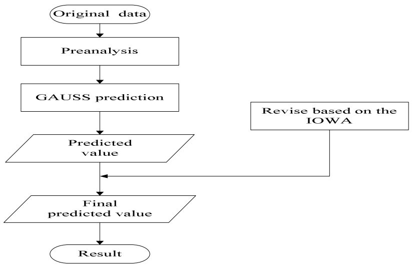


Figure 1: The research model

3. Empirical study

3.1 Data preprocessing

Observing the number of tourists who arrive scenic entrance at each time, we find it is a nonlinear non-stationary time series. However, if we integrally process the number of people of each moment, i.e accumulate the number of people at each time, Let t denote time (minute), N_t denote the number of tourists who arrive scenic entrance at each time, S_t denote the total number of tourists arrivals, then:

$$S_t = \sum_{i=1}^{m} N_t, t = 1, 2, \ldots m$$

3.2 Data clustering according to size

The basic principle of Hierarchical clustering is: firstly, classify a certain number of samples (or variables) into their own classes, then classify two classes which have the closest properties into a new class, calculate the distance between the different classes under the new class, combine two classes which have the closest properties, repeat this process until all the samples (or variables) are combined into one class. This paper conduct hierarchical clustering based on number scale, hoping that the scale distance of data in a class is as close as possible, make the average distance of all items in the combined class is smallest .So we define

$\mathrm{d}(C_i, C_j) = \min_{Z_i \in C_i, Z_j \in C_j} \|Z_i - Z_j\|$ .That is Within-groups linkage method.

3.3 Prediction model based on Gaussian fitting algorithm

The days historical data is divided into a number of scales. This paper fits the curve according to different scales, gets a series of prediction models which have same structures, different parameters. In the actual fitting process, we do not require $y = f(x)$ strictly through all the point (x_i, y_i), but require the fitting error $D = f(x) - y_i$ in point x_i is the smallest according to certain criteria. Often use the least squares approximation searching the best fitting curve method.

This paper uses the gaussian function as a basic function of the curve. Namely, set $y = F(x)$ as a gaussian function system, where each gaussian function is determined by three parameters: peak height A, peak position B and peak width C. The entire gaussian function system is written as:

$$F(x) = \sum_{i=1}^{n} A_i * \exp\left[-\left(\frac{x - B_i}{C_i}\right)^2\right]$$

3.4 Weight calculation forecasting based on improved IOWA operator

In 2.3, we calculated the number of people prediction model under several scales. However, the size is not entirely consistent with the scale of the number of people arrival every day. When given the total arrival number on a day, we hope to be able to predict the number of each moment arrival. Therefore, appropriate weights can be given to the known sizes to obtain the prediction value of each actual size.

The concrete steps are given for solving attribute weights according to this thought as follows:

Step1: Calculate deviation distance d_{ij} between discrete scale i and target scale j . $d_{ij} = |v_i - v_j|, i = 1, 2, \ldots, m$, where v_i represents the size of scale i .

Step 2: Calculate the weight. $\omega_i = \frac{d_{ij}^2}{\sum_{i=1}^{m} d_{ij}^2}, i = 1, 2, \ldots, m$

Step 3: Sort order of the weight and the scale $\omega=(\omega_1,\omega_2,...,\omega_m)^T,\omega_1>\omega_2>...>\omega_m$ $d=(d_{1j},d_{2j},...,d_{mj})^T,d_{1j}<d_{2j}<...<d_{mj}$.

Step 4: Calculate the prediction equation $f_j=\sum_{i=1}^{m}\omega_i f_i, i=1,2,...,m$, where f_i represents the prediction equation for scale i.

4. Empirical analysis

In order to carry out performance evaluation on the proposed prediction model, we conduct a number of empirical studies on data based on real-life scenarios to predict the number of visitor arrival on each day and each time.

4.1 Data Sources

Taking jiuzhaigou scenic area as an example, we have collected data in real-time daily visitor arrivals from May 2012 to August 2012(Collected from RFID, in minutes).Data collected from 7:00 until 13:00, altogether 720 minutes. Tourists scale is seasonal and tourists scale is similar when the date is close. In order to take all scales into account as much as possible in the forecasting process, it is inappropriate to only let the previous data as the training data. So randomly select several days as the training data, assess model parameters, and the rest is used to test the accuracy of the model.

4.2 Scale hierarchical clustering

Conduct hierarchical clustering on the number of days of training data, stratified results are as the following table (20 layers).

Table 1: Hierarchical clustering results

layer	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
date	5.10 5.14 5.21	5.11 5.13 5.15 …	5.12 5.20 5.23 …	5.18 5.26 6.13 …	5.19 5.25 6.10 …	5.28 5.29 5.31 …	6.15 6.16 6.22 …	6.17 7.5	6.23 7.11 8.19 …	6.29 6.30
scale	11000	11800	12800	14187	13600	10000	16300	15600	21231	15000
layer	11	12	13	14	15	16	17	18	19	20
date	7.6 7.7 7.8 …	7.10 7.19 7.23 …	7.12 7.16 7.17 ….	7.13 7.27 8.6	7.14 7.15 7.20 …	7.18 7.30 7.31 …	7.29 8.14 8.18	8.3 8.7 8.8	8.4 8.17	8.5 8.10 8.11
scale	19000	20200	22000	25300	23500	22900	24200	26300	27700	28500

The scale of formation classification is arithmetic mean of the all day data in a class, arrival number at each time under the new scale is also arithmetic mean of the data at each time, changing from 1×10^4 to 2.9×10^4 spans in scale.

The line graph of different sizes is as the following figure. The abscissa represents time (in minutes), the ordinate represents the total number of the arrival at each time. From lower to upper curve represents scale is increasing in turn. We find there is a following regularity:

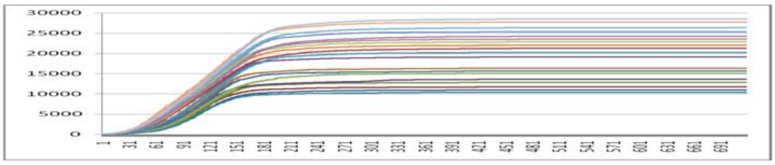


Figure 2: Visitor arrival curve

The graph curve of different sizes exist a high degree of similarity and consistent trend. And this curve shows that the growth rate of the number of visitor arrivals increases along with time increased at the beginning. After a growth period, value of the growth rate gradually slow down and close to a certain limit.

4.3 Segmented gaussian fitting

Of common fitting equation, gaussian fitting have the highest fitting accuracy to this study. Therefore, the following data of different sizes were gaussian fitted respectively. After testing, 8 gaussian fitting errors is the

smallest in the gaussian fitting process. So we choose eight formula gaussian fitting as fitting function. Equation is expressed as follows: $F(x) = A_1 * exp(-((x-B_1)/C_1)^2) + A_2 * exp(-((x-B_2)/C_2)^2)$
Describe it in detail by using 1.18×10^{5}scale as an example. Figure 4 are the fitting curve and the residual curve and the residual is large, so we consider fragmenting to reduce the fitting residuals.

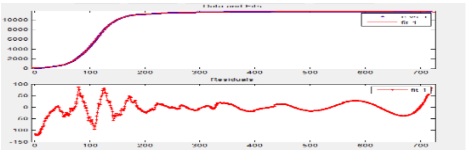


Figure 3: The fitting and residual curve

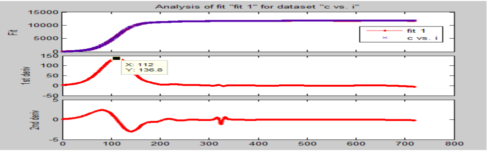


Figure 4: The segmented curve

In order to ensure the reasonableness of segmentation, we conduct first, second derivative on tourist arrivals growth curve. Shown in Figure 2 Introduce tourist destination tourists growth "speed" and "acceleration" concepts and its related variables. Divide it into three stages, and the segmentation points are 110 and 180.
Use the Matlab software to fit, we can obtain fit equation. The parameters of the three segments of gaussian fits are as follows:

Table 2: The three segments parameters of gauss equation

Coefficients (with 95% confidence bounds)

	The first segment **(0-110)**	The second segment **(110-180)**	The third segment **(180-720)**
A_1	667.3	155.4	1.17E+04
B_1	114.3	181.3	787.2
C_1	10.04	31.96	2056
A_2	-19.35	-7.944	20.7
B_2	105	163.2	227.9
C_2	0.6545	5.696	7.678
…..			
A8	189.6	1.29E+04	265.3
B8	105.3	190.8	374.5
C8	3.899	82.77	245.4

The first segment Goodness of fit: SSE: 2.215e+004; R-square: 0.9999; Adjusted R-square: 0.9999; RMSE: 16.24.
The second segment Goodness of fit: SSE: 4511; R-square: 1; Adjusted R-square: 1; RMSE: 10.13.
The third segment Goodness of fit: SSE: 2699; R-square: 0.9998; Adjusted R-square: 0.9998; RMSE: 2.291.
It shows that Gaussian fitting can pass the conformance testing and have high fitting precision to this study. Repeat the above experiment,20 parameters values of discrete scale equation can be obtained. Here is not to list.

4.4 Calculation of weights

The closer the scale is, the closer the curve graph is. To reduce the interference by the too large distance of the scales to the scale to be predicted, we conduct a secondary clustering, and cluster different scales into four types. The result is as the following table.

Table 3: Secondary clustering results

Classification	Scale
1	10000 11000 11800
2	12800 13600 14187 15000 15600 16300
3	19000 20200 21231 22000 22900 23500 24200
4	25300 26300 27700 28500

This paper choose data from May 10, June 13, July 22, August 8 to predict the above four classifications respectively to test the validity of model prediction.

from the historical data can be known, the tourist number on May 10, June 13, July 22 and August 8 are respectively 1.1044×104, 1.4301×104, 2.2126×104 and 2.6288×104 (unit: person) the four-day Predicting scales are respectively recorded as F_1、F_2、F_3、F_4 , according to improved IOWA operator, we can get:

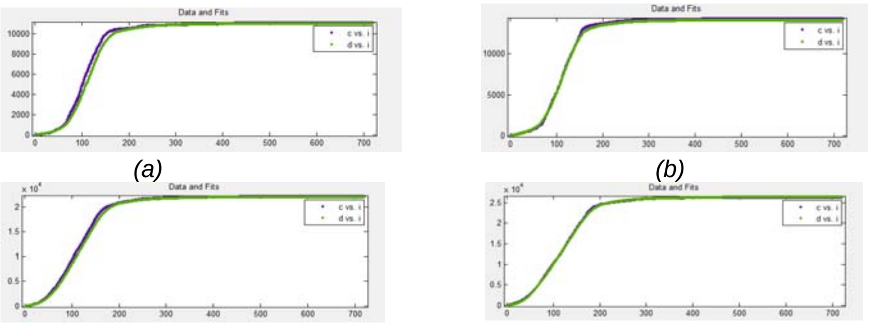
(1) F_1=0.00116f_1+0.65524f_2+0.34359f_3 (f_1、 f_2 and f_3 respectively represent the gaussian fitting function formulas of 1.0000×104, 1.1000×104 and 1.1800×104 in scale)

(2) $F_2=0.00021f_4+0.31264f_5+0.66492f_6+0.00021f_7+0.31264f_8+0.0002f_9$

(3) $F_3=0.00077f_{10}+0.03932f_{11}+0.18215f_{12}+0.47983f_{13}+0.21122f_{14}+0.09270f_{15}+0.02942f_{16}$ (4) $F_4=0.25356f_{17}+0.62228f_{18}+0.12414f_{19}+0.00002f_{20}$

4.5 Analysis of experimental results

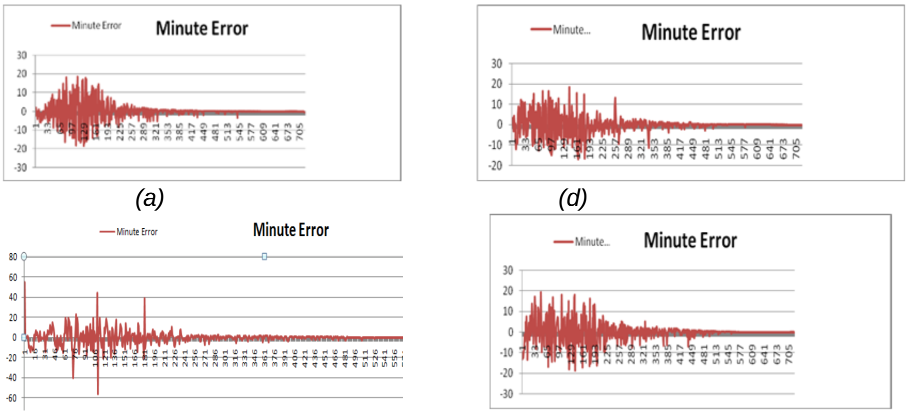
The following is the experimental results, the blue line is the accumulated value each time of the actual data, the green line is the prediction data. (a), (b), (c), (d) in the figure 5 respectively represent the real and predicted data on May 10, June 13, July 22, August 8;(a), (b), (c), (d) in the figure 6 respectively represent every minute error curve on May 10, June 13, July 22, August 8.



(a) (b)

(c) (d)

Figure 5: Real data and predicted data



(a) (d)

(c) (d)

Figure 6: Every minute error curve

From Figure 6 (a), (b), (d) can be seen: every minute errors on May 10, July 22 and August 8 are respectively in 20 or less; in the (c) ,despite three errors per minute value on June 13 are other than 20, but the overall trend is in 20 or less. It shows the validity and reliability of the model our proposed.

5. Conclusion and prospect

This paper proposes a new research question about the tourists arrival prediction, and enunciates the necessity and importance of this research question from a more detailed time scale. i.e studies the number of tourists who reach the scenic area at each time per day. We findThe graph curve of different sizes exists a high degree of similarity and consistent trend and this curve shows that the growth rate of the number of visitors arrival increase along with time increment at the beginning. After a growth period, value of the growth rate gradually slow down and close to a certain limit. In order to study this problem deeply, on the base of stage prediction theoretical model based on tourism theory and clustering theory, this paper build a phased Gaussian fitting and weights combined forecasting model to predict tourists arrival, and analyse and test the model using Jiuzhaigou scenic area as an example. Example verification shows that the proposed new model has a good predictive accuracy.

However, we do not take the errors after fitting into account in the research process. The error can be used to correct in the future to improve accuracy. Then due that we solve a new problem to forecast real-time tourist number in the scenic area. The proposed prediction model for this new research question does not conduct compared prediction with other model in tourist arrivals filed. This prediction model and prediction method will be improved in future studies.

Acknowledgements

This work was supported by the Major International Joint Research Program of the National Natural Science Foundation of China (Grant no. 71020107027), and the National Natural Science Foundation of China (Grant no. 71001075) and Humanities and Social Sciences project of The Ministry of education of China (Grant no.12YJC630023).

References

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327. DOI: 10.1016/0304-4076(86)90063-1

Broomhead, D. S., & King, G. P., 1986, Extracting qualitative dynamics from experimental data. Physica D: Nonlinear Phenomena, 20(2-3), 217-236. DOI: 10.1016/0167-2789(86)90031-X

Cang, S., 2011. A non-linear tourism demand forecast combination model. Tourism Economics, 17(1), 5e20. DOI: 10.5367/te.2011.0031

Chen M. S., Ying L. C., Pan M. C., 2010, Forecasting tourist arrivals by using the adaptive network-based fuzzy inference system [J]. Expert Systems with Applications, 37(2): 1185-1191. DOI: 10.1016/j.eswa.2009.06.032

Chu, F. L., 2008, A fractionally integrated autoregressive moving average approach to forecasting tourism demand. Tourism Management, 29(1), 79-88. DOI: 10.1016/j.tourman.2007.04.003

Coshall, J. T., 2009, Combining volatility and smoothing forecasts of UK demand for international tourism. Tourism Management, 30(4), 495-511. DOI: 10.1016/j.tourman.2008.10.010

De Gooijer, J., Hyndman, R., 2006. 25 years of time series forecasting. International Journal of Forecasting, 22 (3), 443–473. DOI: 10.1016/j.ijforecast.2006.01.001

Gil-Alana L. A., 2005, Modelling international monthly arrivals using seasonal univariate long-memory processes[J]. Tourism Management, 26(6): 867-878. DOI: 10.1016/j.tourman.2004.05.003

Kim, S. S., & Wong, K. K., 2006, Effects of news shock on inbound tourist demand volatility in Korea. Journal of Travel Research, 44(4), 457-466. DOI: 10.1177/0047287505282946

Song H., Li G., 2008, Tourism demand modelling and forecasting—A review of recent research [J]. Tourism Management, 29(2): 203-220. DOI: 10.1016/j.tourman.2007.07.016

Wan S. K., Wang S. H., Woo C. K., 2013, Aggregate vs. disaggregate forecast: case of Hong Kong. Annals of Tourism Research, 42, 434-438. DOI: 10.1016/j.annals.2013.03.002