Advances in Economics, Business and Management Research, volume 71

4th International Conference on Economics, Management, Law and Education (EMLE 2018)

Analysis on the Role of Daily Consumer Search Data in Forecasting Monthly Tourist Flow

A Mixed Data Sampling Approach*

Binru Zhang
School of Finance and Economics
Yangtze Normal University
Chongqing, China

Yulian Pu**
School of management
Yangtze Normal University
Chongqing, China
**Corresponding Author

Rong Hu
School of Mathematics
Sichuan University of Arts and Science
Dazhou, China

Runzhi Tang
School of Finance and Economics
Yangtze Normal University
Chongqing, China

***Abstract*—In order to evaluate the predictive ability of network search data of daily sampling frequency for monthly tourist flow, this paper predicts the monthly tourist flow of Chongqing, China. In consideration of the inconsistency of sampling frequency of network search data and tourist flow data, an autoregression mixed data sampling model (AR-MIDAS) is constructed for prediction to avoid the loss of information. This paper adopts factor analysis technology to extract the characteristic information contained in the consumer search data related to Chongqing tourism, and then puts the obtained comprehensive factor into the model for a prediction experiment. The research results show that AR-MIDAS model can improve the precision of monthly tourist flow prediction better than ARIMA and MIDAS prediction techniques. The research results can provide necessary reference for scientific decision-making of tourism related departments.**

Keywords—MIDAS model; tourist flows; consumer search data; forecasting precision

I. INTRODUCTION

Accurate prediction of tourist flow in scenic spots is a major issue to be addressed in risk prevention and control and emergency management of tourism public safety (Athiyaman & Robertson, 1992; Song & Li, 2008). However, the inconsistent frequency of available time series data in practical applications will lead to a dilemma in research work. For one thing, Baidu provides daily network search data, which imply potentially useful information and can effectively improve the prediction ability of the model (Li et al., 2017; Wei & Cui, 2018); for another, during the modeling procedure in reality, if the data sampling frequency of the predicted variable is lower, we often cannot directly make the utmost of the characteristic information of the high-frequency variable. This is because most regression models require consistent data sampling frequency for all variables. A common solution for this type of situation is to preprocess the data so that the frequency of all variables in the model is equal (Ghysels, Sinko, & Valkanov, 2007). However, this approach will result in the loss of high frequency variable information, and it is difficult for the model to detect the relationship between variables. This paper discusses how to predict tourism demand by using hybrid frequency data.

In the forecast of tourism demand, the predictive variables used in various regression models mainly include macrostatistics and network search data released by the official. The release of macrostatistics related to tourism demand has certain hysteresis quality, affecting the timeliness of forecasting (Zhang et al., 2018). With advances in information technology and all-round development of the Internet, the network information search becomes a major tool (Vila et al., 2018) for consumer tourism making decision and the search engines record a lot of consumer search data. Such data are easy to be obtained and can reflect tourists' potential tourism needs. In recent years, in a large number of literatures, network search data are used to predict tourism demand (See for example: Yang et al., 2015; Li et al., 2017; Zhang et al., 2017). Researches show that such data can effectively improve model's prediction performance. Even so, the network search data used in studies are daily or weekly data, and the predicted variables are mainly monthly data. To tackle the problem of frequency inconsistency, high frequency data will be processed by using the method of average weighted summation so that the frequency of the predicted variable is equal to that of the predictive variable. However, this will lose the characteristic information of high frequency data, resulting in that the model cannot fully reflect the real relationship between variables.

An efficient scheme for hybrid frequency data is to use the mixed data sampling regression model (The mixed data

*Project program: Fund Program of Chongqing Social Science of China (2017YBGL137); Scientific Research Fund of Sichuan Provincial Education Department (18ZA0416)

sampling, MIDAS) proposed by Ghysels, Santa-Clara and Valkanov (2004) to make predictions and researches. This method has a certain simplicity and flexibility, allowing the difference existing in data sampling frequency of the predictive variable and the predicted variable. The MIDAS model is mainly used in the fields of finance and macroeconomics (see a review by Andreou, Ghysels, & Kourtellos, 2010b). In terms of tourism demand forecast, only Bangwayo-Skeete and Skeete （2015） made a predictive study of tourist flow of the five tourist destinations in Caribbean by using Google trends data and Autoregressive Mixed-Data Sampling (AR-MIDAS) model. The results indicate that AR-MIDAS can effectively improve model's prediction ability compared with Seasonal Autoregressive Integrated Moving Average (SARIMA) and autoregression model. However, for Chinese consumers, they mainly make use of Baidu search engine for tourism-related information. Baidu search engine records daily and weekly search data. The predictive ability of such data for monthly tourism demand including tourist flow and hotel occupancy rate is worthwhile further discussing.

Taking Chongqing as an example, this paper collects daily network search data related to Chongqing tourism. In order to keep the modeling as simple as possible and retain the original characteristic information of the data, this paper adopts the method of factor analysis to extract the principal factor in the network search data as the model input set, and constructs the AR-MIDAS model to predict the monthly tourist flow in Chongqing, China. It also introduces the MIDAS and ARIMAX models as the benchmark model to compare the predictive performance of the construction method. Predictions show that the predictive performance of AR-MIDAS outperforms those of other prediction methods in a short-term prediction.

II. METHODOLOGY

A. The AR-MIDAS Model

Ghysels et al. (2004) introduced MIDAS model based on the distributed lag model. Compared with simple weighted averaging scheme, MIDAS method adopts a more subtle weighting scheme that imposes a distribution constraint on the weights to keep the model simple. Based on the MIDAS principle (Ghysels et al., 2004), a simple AR-MIDAS prediction model can be expressed by Equation (1):

$$Y_{t+h} = \alpha + \sum_{i=1}^{p} \beta_i L^i Y_t + \gamma \sum_{k=1}^{m} \Phi(k,\theta) L_{HF}^k X_t + \varepsilon_t \tag{1}$$

Among it, Y is predicted variable, and X represents high frequency prediction variable, and L is lag operator, and β_i, γ are regression coefficient. $\Phi(k,\theta)$ is weighting function, of which k is lag order and θ is a hyperparameter vector that defines the shape of the weight function. By imposing constraints on $\Phi(k,\theta)$, that is, by standardizing weights, the coefficient γ can be identified.

The MIDAS method reduces the number of parameters estimated by optimizing the hyperparametric vector $\theta = (\theta_1, \theta_2, \cdots, \theta_j)$ of the weight function to keep the model simple. The MIDAS framework mainly uses the distribution functions of exponential Almon lag and exponential Almon lag. Because only a few parameters are required to describe various distribution shapes, these two distribution functions prevail in MIDAS literature. The exponential Almon lag defined by Ghysels et al. (2004) is taken as distribution function in this paper. The mathematical expression is as follows:

$$\Phi(k;\theta_1,\theta_2) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{j=1}^{m} \exp(\theta_1 j + \theta_2 j^2)} \tag{2}$$

When $\theta_1 = \theta_2$, we can obtain a simple time weighted average scheme. θ_1 determines the location of the function graph, and θ_2 determines the slope.

B. Forecast Evaluation

To compare the prediction ability of the constructed AR-MIDAS model, the MIDAS and ARMAX models are introduced as the benchmark model. ARIMAX is an ARIMA model that adds the network search data. The model transforms the network search data by using the method of average weighted summation to make it consistent with the frequency of the predicted variable. The introduction of the ARIMAX model aims to compare the predictive ability of the model under two weighting schemes. The MIDAS is brought to verify that the lagged variable of predicted variable has predictive ability.

We adopt two indicators of Mean Absolute Percentage Error (MAPE) and the correlation coefficient R to measure the predictive performance of the model. Their mathematical expressions are expressed as follows:

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100\% \tag{3}$$

$$R = \frac{\sum_{t=1}^{N} (y_t - \bar{y})(\hat{y}_t - \bar{Y})}{\sqrt{\sum_{t=1}^{N} (y_t - \bar{y})^2} \sqrt{\sum_{t=1}^{N} (\hat{y}_t - \bar{Y})^2}} \tag{4}$$

Among them, $y_t, \hat{y}_t$ respectively represent the observed value and predicted value, and N represents the number of forecast periods, and $\bar{y}, \bar{Y}$ respectively represent the average value of $y_t, \hat{y}_t$. Formula MAPE represents the deviation between the actual value and the predicted value. The smaller the value is, the more accurate the prediction will be. Formula

R represents the model fitting degree. The closer the value is to 1, the higher the model fitting degree is.

III. EXPERIMENTAL DATA

In order to verify the prediction performance of the constructed prediction method, this paper takes Chongqing as an example to predict the monthly tourist flow of Chongqing. Chongqing is famous as a mountain city with more than 300 natural and cultural attractions. In 2017, Chongqing received 3.5835 million inbound tourists, and its foreign exchange earnings from tourism was USD 1.948 billion, with an increase of 13.2% and 15.5% respectively. However, the sharp increase in consumer demand for tourism has put some pressures on the environmental carrying capacity of scenic spots. In consequence, it is necessary to forecast the tourist flow of Chongqing in order to provide necessary reference for the scientific decision-making of tourism-related departments. The data of domestic tourist flow received by Chongqing derive from wind database, which is the leading financial data sharing platform in China. Data were collected from January 2011 to April 2018. Due to the data missing at some time points, this paper fills in the missing values according to the trend smoothing method.

The network search data comes from Baidu. Characterized by easy access and strong real-time, these data can predict the dynamic characteristics of tourist flow in scenic area with USD 1.948 billion in advance (Zhang et al., 2018). The network search data were collected from January 1, 2011 to April 30, 2018. This paper makes use of the method constructed by Zhang et al. (2018) to get network search data related to Chongqing tourism, weighs and averages keyword variables to monthly data, and obtains 6 keyword variables with predictive ability through Pearson correlation analysis. The Pearson correlation coefficients of these keyword variables and the predicted variables are all above 0.60. The descriptive statistical analysis of the final selected keyword predictive variables and the predicted variables is shown in "Table I".

TABLE I. DESCRIPTIVE STATISTICS OF EACH VARIABLE

Variables	Sample Number	Mean	Standard Deviation	Min. Value	Max. Value
Tourist flow	88	3191.59	1552.52	1094.40	1094.4
Chongqing weather forecast	2677	6872.32	3536.01	918	25926
Chongqing foreigner's Street	2677	641.71	275.76	140	2723
Chongqing jiangfangbei	2677	528.18	205.20	170	1964
Chongqing Chaotianmen	2677	354.34	124.40	134	2097
Chongqing zoo	2677	353.32	170.80	82	1919
Chongqing night view	2677	480.70	188.63	208	1591
	2677	367.08	127.04	148	1093

IV. EMPIRICAL RESULTS AND DISCUSSIONS

To reduce the complexity of prediction, this paper uses the method of factor analysis to extract the predictive variables (Kim & Swanson, 2017), which can not only reduce dimensionality but also extract the feature information of each keyword variable. Finally, one comprehensive factor F is obtained, which can explain 84.7% of the variance. Augmented Dickey Fuller (ADF) test indicates that all variables are trend stationary time series. To conduct prediction test, we divide the experimental data set into training set and testing set, and take the period from January 2011 to April 2017 as the training set to carry out model training and takes the last 12 months as the testing set used for model testing. Through the training, the main parameter of AR-MIDAS model is $\beta_1 = 0.69, \gamma = 0.43, \theta_1 = 0.001, \theta_2 = -2e-04$, and the predicted results of models are shown in "Table II".

TABLE II. THE PREDICTED RESULTS OF MODELS ON THE TESTING SET

Time	Tourist Flow	AR-MIDAS	ARIMAX	MIDAS
2017.05	3554.99	3485.22	3651.1	3599.31
2017.06	4815.08	4867.36	4700.21	4890.1
2017.07	4189.91	4066.17	3995	4380.34
2017.08	5164.89	4910.21	5299.37	4889.02
2017.09	4355.11	4199.52	4106.33	4286.3
2017.10	10417.52	9021.08	8900.33	12011.65
2017.11	3218.92	3002.01	3126.44	3187.2
2017.12	3430.77	3562.2	3789.1	3654.33
2018.01	3185.23	2987.09	3055.12	3299.7
2018.02	5461.43	5388.9	5490.22	5521.48
2018.03	4245.89	4366.2	4512.33	4427.3
2018.04	3558.91	3645.31	3498.22	3367.01

"Table II" shows that in the prediction in 12 months, the best prediction (boldface letter) of AR-MIDAS accounts for 5 months, and the best prediction of ARIMAX and MIDAS respectively accounts 3 months and 4 months. However, from the last 6 months, all the three models are 2 months. This results show that with the increasing of number of forecast periods, the predictive ability of AR-MIDAS is gradually reducing. But overall, the prediction model constructed by this paper has the best prediction effect. The statistical performance

index values based on the prediction results are shown in "Table III".

TABLE III. THE STATISTICAL INDEX VALUE ON THE TESTING SET BASED ON PREDICTION RESULTS

Model	MAPE (%)	R
AR-MIDAS	4.274	0.993
ARIMAX	4.878	0.988
MIDAS	4.286	0.994

It can be seen from "Table III" that on the overall, MAPE values of the three models are all below 5%, and AR-MIDAS performs better, which means that the model has the highest prediction accuracy; R values of the three models are close to 1, but R of MIDAS has a higher score, which implies that there is a better relevance between the predicted value and the actual value, and the fitting degree of model is higher.

Because MIDAS fails to use the historical information of the predicted variables, its prediction accuracy is lower than that of AR-MIDAS, which fully indicates that the 1 phase lag variable of tourist flow in scenic area has certain predictive ability. Because ARIMAX method uses the method of average weighting to convert network search data into monthly data and loses the characteristic information of the daily data, AR-MIDAS has higher predictive ability than ARIMAX.

V. CONCLUSION

In view of the inconsistency between the predicted variable and the frequency of predictive variable in the tourism demand forecast, this paper takes Chongqing, China case as an example and constructs an AR-MIDAS mixed model to predict and analyze the domestic tourist flow in Chongqing. This paper collects the daily data of network search and monthly data of tourist flow from January 2011 to April 2018, constructs experimental data sets and adopts factor analysis to conduct feature extraction for the Internet search data. The predicted results show that the constructed prediction models can improve the short-term predictive ability of models, but with the increasing of number of forecast periods, the predictive ability has a downward trend.

The excellent predictive ability of the constructed prediction models benefit from the following reasons. First, the addition of autoregression improves the prediction accuracy, which is related to the features of periodic fluctuation of customer flow. There is a great relevance between the observation of customer flow in the last time and in the current time, therefore, the customer flow information in a month in advance has predictive ability. Second, different from MA-MIDAS, ARIMAX model conducts average weighted summation and processes the predictive variables, loses the characteristic information of network search data, reducing the prediction accuracy of the models.

This paper conducts factor analysis for the network data with predictive ability and extracts common factor. This method can reduce the complexity of model prediction, but also loses part of the information of the keyword variables with predictive ability. In future research, building a more predictive MIDAS model and conducting prediction and early warning research is the direction of further efforts.

REFERENCES

[1] Athiyaman, A., & Robertson, R.W. (1992).Time series forecasting techniques: Short-term planning in tourism. International Journal of Contemporary Hospitality Management, 4(4), 8-11.

[2] Bangwayo-Skeete P F, Skeete R W. Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach.[J]. Tourism Management, 2015, 46:454-464.

[3] Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models . (IDEAS Working Paper Series from RePEc No. 2004s-20).

[4] Ghysels,E., Sinko,A., & Valkanov, R. MIDAS Regressions: Further Results and New Directions[J]. Econometric Reviews, 2007, 26(1):53-90.

[5] Kim, H. H., & Swanson, N. R. (2017). Methods for backcasting, nowcasting and forecasting using factor-midas: with an application to korean gdp. Journal of Forecasting, 37(1).

[6] Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. Tourism Management, 59, 57-66.

[7] Sampling Regression Models . (IDEAS Working Paper Series from RePEc No. 2004s-20).

[8] Song, H., & Li, G. (2008). Tourism demand modelling and forecasting — a review of recent research. Tourism Management, 29(2), 203-220.

[9] Vila, T. D., Vila, N. A., Gonzã¡Lez, E. A., Brea, J. A. F., & Zhang, Z. (2018). The role of the internet as a tool to search for tourist information. Journal of Global Information Management, 26(1), 58-84.

[10] Wei,J.R., & Cui,H.M.(2018).The Construction of Regional Tourism Index and Its Micro-Dynamic Characteristics: A Case Study of Xi'an.Journal of Systems Science and Complexity, 2018, 38(2): 177-194.

[11] Yang, X., Pan, B., James, A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. Tourism Management, 46, 386-397.

[12] Zhang, B.R., Huang, X.K., Li, N., & Law, R. (2017). A novel hybrid model for tourist volume forecasting incorporating search engine data. Asia Pacific Journal of Tourism Research, (3), 245-254.

[13] Zhang, B.R., Liu,S.L., Zhang, C.F., & Pu,Y.L. (2018). Forecasting hotel occupancy rate based on consumer search within network environment. Statistics & Information Forum, 33(3), 93-99.