Investigating the Time-varying Effect of Search Index in Predicting Tourism Volume Using Dynamic Model Averaging

Yuan Chen
School of Information Management, Wuhan University, Wuhan, 430072, China；Center for Ecommerce Research and Development, Wuhan University, Wuhan, 430072, China

Yiping Ma
School of Information Management, Wuhan University, Wuhan, 430072, China；Center for Ecommerce Research and Development, Wuhan University, Wuhan, 430072, China

Zhongyi Hu
School of Information Management, Wuhan University, Wuhan, 430072, China；Center for Ecommerce Research and Development, Wuhan University, Wuhan, 430072, China, Zhongyi.hu@whu.edu.cn

Tao Xiong
College of Economics and Management, Huazhong Agricultural University, Wuhan, 430070, China

Jiang Wu
School of Information Management, Wuhan University, Wuhan, 430072, China；Center for Ecommerce Research and Development, Wuhan University, Wuhan, 430072, China

See next page for additional authors
Follow this and additional works at: https://aisel.aisnet.org/whiceb2021

Authors

Yuan Chen, Yiping Ma, Zhongyi Hu, Tao Xiong, Jiang Wu, and Tingshun Xiong

Full Research Paper

Investigating the Time-varying Effect of Search Index in Predicting

Tourism Volume Using Dynamic Model Averaging

Yuan Chen[1,2], Yiping Ma[1,2], Zhongyi Hu[1,2,⁻], Tao Xiong[3], Jiang Wu[1,2], Tingshun Xiong[1]

[1]School of Information Management, Wuhan University, Wuhan, 430072, China

[2]Center for Ecommerce Research and Development, Wuhan University, Wuhan, 430072, China

[3]College of Economics and Management, Huazhong Agricultural University, Wuhan, 430070, China

Abstract: Search index has been gradually used for tourism volume forecasting. While most literature aimed to improve the accuracy of a tourism prediction model, rarely considered the influencing characteristics of the search index in the prediction model. In this study, a dynamic model averaging (DMA) approach is applied to build tourism volume model and to investigate the time-varying effect of search indexes in predicting tourism volume. The tourism volume of Jiuzhai Valley and related search indexes from Baidu Index are considered for experimental purpose. According to the results, search index present time-varying characteristic, specifically, the search indexes of *Jiuzhai Valley hotel*, *Jiuzhai Valley*, *Jiuzhai Valley map* and *Jiuzhai Valley guide* have higher probabilities in supporting the prediction of tourism volume.

Keywords: tourism forecasting, search index data, time-varying effect, dynamic model averaging

1.   INTRODUCTION

Tourism volume is an essential indicator of the tourism industry's stability and development. The accurate forecasting of tourism volume can help to allocate resources and formulate pricing strategies effectively [1]. Therefore, both tourism practitioners and government regulators pay much attention to the prediction of tourism volume.

Until now, many scholars have investigated tourism prediction and paid a lot of efforts to improve prediction accuracy. In terms of datasets, various online data is put forward to overcome the lag of simple historical tourist data, like search index and web traffic data [2,3]. Search index, which is produced by search engine companies, is a kind of index by integrating Internet users' searching query data in the search engine (e.g., Baidu, Google, and 360 search). Considering the widely use of search engines, the search index is useful in capturing Internet users' search behavior. In the area of tourism, many tourists would like to use search engines to search for related information about destinations to make travel plans. Therefore, the search index about a place to some extent reflects users' interests and intention to visit. Recently, it has been an emerging trend in using the search index for tourism forecasting [4].

To demonstrate the effectiveness of search engine data in forecasting tourist volume, many studies have applied search index in their prediction models. For example, Bangwayo-Skeete and Skeete[5] compared simple Autoregressive (AR) method and the Seasonal Autoregressive Integrated Moving Average (SARIMA) method of historical tourist arrivals data with AR method using high-frequency Google search queries data. The results showed that AR model with Google search queries data gave superior forecasts to AR and SARIMA time series models. Apart from Google Trends, Baidu Index is widely adopted when the study is about China, and it turns out that model with Baidu index performs well [6,7]. Sun et al. built a prediction model for tourist arrivals of popular destinations in China using search indexes from Google and Baidu [3]. The experimental results verified the Granger causality and co-integration relationship between search indexes and tourist arrivals. The

⁻ Corresponding author. Email: Zhongyi.hu@whu.edu.cn (Zhongyi Hu)

forecasting accuracy was also significantly improved while applying search indexes for predicting tourist arrivals. However, the existing studies mainly focused on the effectiveness of the search index in predicting tourism volume; the question about how the search index was involved in the prediction model remains unanswered. Especially, the coefficients of the search index in the prediction model can change over time [8,9].

To handle these issues, dynamic model averaging (DMA) approach was proposed in which both forecasting models and the coefficients of predictors can change over time. Raftery et al. [10] proposed a DMA model to forecast the output strip thickness for a cold rolling mill and pointed out that DMA performed better than single models. As a special case, if only the highest probability model at each point of time is selected to form the forecasting model, it turns to be a dynamic model selection (DMS) approach. DMA/DMS approach has been widely applied in different forecasting fields including housing market [11], stock market [12], and fossil fuels prices [13]. It turns out that the DMA/DMS approach not only has better forecasting performance but also identifies different factors' influence on prediction. As for tourism forecasting area, with the influencing factors being diverse, DMA method seems suitable to improve the forecasting performance and figure out factors' impact of future tourism arrivals. However, to our knowledge, among the existing studies on tourism prediction, while most literature focused more on improving forecasting accuracy, there has been no literature that considers the dynamically changing coefficients of predictors.

To fill the above research gap, this paper applies the DMA approach to forecast tourists' arrivals of Jiuzhai Valley with search index and historical tourist volumes. The predictors are selected from Jiuzhai Valley-related search indexes according to the Pearson correlation coefficient. The time-varying effect of search index is then demonstrated based on DMA. This paper contributes to the existing tourism prediction literature in three ways. First, we introduce the DMA approach for tourism forecasting, in which the forecasting model and coefficients of predictors are dynamically updated over time. To date, this is the first study that applies DMA in tourism forecasting. Second, this study investigates the time-varying effect of search index in predicting tourist volume. There are few, if any, studies have focused on the influencing of search index on tourist volume. Third, by taking Jiuzhai Valley as a case study, this study shows some interesting findings of the time-varying influences between search index and tourist volume.

The rest of this study is organized as follows. The methodology of the DMA approach is explained in Section 2. Section 3 describes datasets and some preliminary analysis. Section 4 presents the experimental results and discussions. Finally, we conclude this study with some future directions in Section 5.

2.   METHODOLOGY

Allowing the regression coefficients and forecasting models to change over time simultaneously, the DMA framework suits the situation well when the optimal forecasting model evolves over time.

To illustrate the DMA framework, we suppose a set of K forecasting models with the subsets of z_t being predictors. And y_t refer to tourism volume at time t. The set of forecasting models can be written as follows:

$$y^t = z_t^{(k)}\theta_t^{(k)} + \varepsilon_t^{(k)} \tag{1}$$
$$\theta_{t+1}^{(k)} = \theta_t^{(k)} + \eta_t^{(k)} \tag{2}$$

where $z_t^{(k)}$ denote the subset of predictors for $k = 1, 2, ..., K$, and errors $\varepsilon_t^{(k)} \sim N(0, H_t^{(k)})$ and $\eta_t^{(k)} \sim N(0, Q_t^{(k)})$

If z_t contains m predictors, the number of possible forecasting models comes to be $K = 2^m$, and at each point of time the models are different. At time $t - 1$, let $\pi_{t|t-1,k}$ denote the probability of model K being the optimal forecasting model. DMA method gets the final forecasting result by averaging forecasting values with $\pi_{t|t-1,k}$ as weights at each time point (see Equation 3). When the final forecasting value consists of the best performance model at each time point, it becomes DMS. Let $\pi_{t|t-1}^*$ denotes the maximum value of the probability of model selection at each time point, and the equation is shown as follows.

$$\hat{y}_t^{DMA} = \sum_{k=1}^{K} \pi_{(t|t-1,k)} \; \hat{y}_t^{(k)} \tag{3}$$

$$\hat{y}_t^{DMS} = \sum_{k=1}^{K} \pi_{(t|t-1)}^* \; \hat{y}_t^{(k)} \tag{4}$$

Therefore, the next step is to estimate the coefficients and probability of each forecasting model at each time point. By following Kalman filtering[14], we can get Equation (5), where $\Sigma_{t|t-1}^{(k)} = \Sigma_{t-1|t-1}^{(k)} + Q_t^{(k)}$. After updating the above equation according to Kalman filter, Equation (6) can be used for recursive forecasting.

$$\begin{aligned} \theta_{t-1}^{(k)}|y^{t-1} &\sim N(\hat{\theta}_{t-1}^{(k)}, \Sigma_{t-1|t-1}^{(k)}) \\ \theta_t^{(k)}|y^{t-1} &\sim N(\hat{\theta}_{t-1}^{(k)}, \Sigma_{t|t-1}^{(k)}) \end{aligned} \tag{5}$$

$$y_t|y^{t-1} \sim N(x_{t-1}\hat{\theta}_{t-1}^{(k)}, H_t^{(k)} + x_{t-1}^{(k)} \Sigma_{t|t-1}^{(k)} x_{t-1}^{(k)}{}') \tag{6}$$

To estimate $Q_t^{(k)}$, a forgetting factor approach [10] is applied, as shown in Equation (7):

$$\Sigma_{t|t-1}^{(k)} = \frac{\Sigma_{t-1|t-1}^{(k)}}{\lambda} \tag{7}$$

or $Q_t^{(k)} = (1 - \lambda^{-1}) \Sigma_{t-1|t-1}^{(k)}$, where $0 < \lambda \le 1$. The forgetting factor λ^m indicates the weight of the observation value in the past m period. A small forgetting factor value implies a rapid changing of the coefficient.

Due to the fact that Markov switching process has quite heavy calculation burden when estimating $\pi_{t|t-1,k}$, we apply a forgetting factors approach [10] and α is the forgetting factor:

$$\pi_{t|t-1,k} = \pi_{t|t-1,k}^{\alpha} / \sum_{l=1}^{K} \pi_{t|t-1,l}^{\alpha} \tag{8}$$

where $0 < \alpha \le 1$, and the interpretation can be similar to λ.

Finally, the estimate of $H_t^{(k)}$ uses the Exponentially Weighted Moving Average (EWMA) approach as follows[8]:

$$\hat{H}_t^{(k)} = \sqrt{(1 - \kappa) \sum_{j=1}^{t} \kappa^{j-1} (y_j - z_j^{(k)} \hat{\theta}_j^{(k)})^2} \tag{9}$$

where κ is a decay factor. The EWMA specification can get volatility forecasts approximation by a recursive form.

3. DATA PREPARATION

To verify the effectiveness of DMA approach with search indexes, this study takes Jiuzhai Valley's tourist volume and search index data as a case study. This section describes the dataset with preliminary analysis.

(1) Data description

Jiuzhai Valley, locally known as Jiuzhaigou, is a popular tourist spot in Sichuan province of China. Its outstanding view has attracted millions of tourists per year. Unfortunately, there was an earthquake in August 8, 2017, which resulted in shut down of the place until September of 2019. Therefore, the daily tourist arrival data, ranging from May 27, 2012 to August 7, 2017 (with 1899 observations), is considered as the explained variable. As for search index, according to the results of the CNZZ data center (www.cnzz.com), Baidu shares the largest market among the search engines in China. Thus, this paper collects search indexes about Jiuzhai Valley from Baidu Index (http://index.baidu.com/) with the same time period of tourist volume.

**(2) Keywords selection**

To select reasonable and comprehensive search indexes, it is vital to contain diverse and most related

searching keywords about Jiuzhai Valley. First, we set 12 initial keywords of six different aspects including eating, lodging, traffic, scenic spot, shopping and entertainment. Second, a keyword mining tool (https://www.chinaz.com/) was used to extend relevant keywords and finally resulted in 56 keywords. By collecting search indexes from Baidu Index based on each keyword, we obtained 35 keywords' search indexes, whereas the others were not tracked by Baidu Index. To filter out those search indexes that are more relevant to tourist volume, we further calculated the Pearson correlation coefficient between each keyword and tourist volume. Finally, eight keywords with the correlation coefficient above 0.65 were selected as final predictors, which are listed in Table 1. As shown in this table, the selected keywords are all highly relevant to travelling of Jiuzhai Valley. Figure 1 illustrates the trend of several search indexes and tourism volume. From this figure, we can find that these search indexes have a similar trend as tourism volume.

Table 1. Selected keywords (translated from Chinese)

Keywords	Correlation coefficients
Jiuzhai Valley map (九寨沟地图)	0.7794
Jiuzhai Valley Huanglong (九寨沟黄龙)	0.7521
Jiuzhai Valley guide (九寨沟攻略)	0.7408
Jiuzhai Valley scenic spot (九寨沟景点)	0.7164
Jiuzhai Valley hotel (九寨沟酒店)	0.7066
Jiuzhai Valley (九寨沟)	0.6743
Jiuzhai Valley scenic map (九寨沟景区地图)	0.6670
Jiuzhai Valley free travel (九寨沟自由行)	0.6541
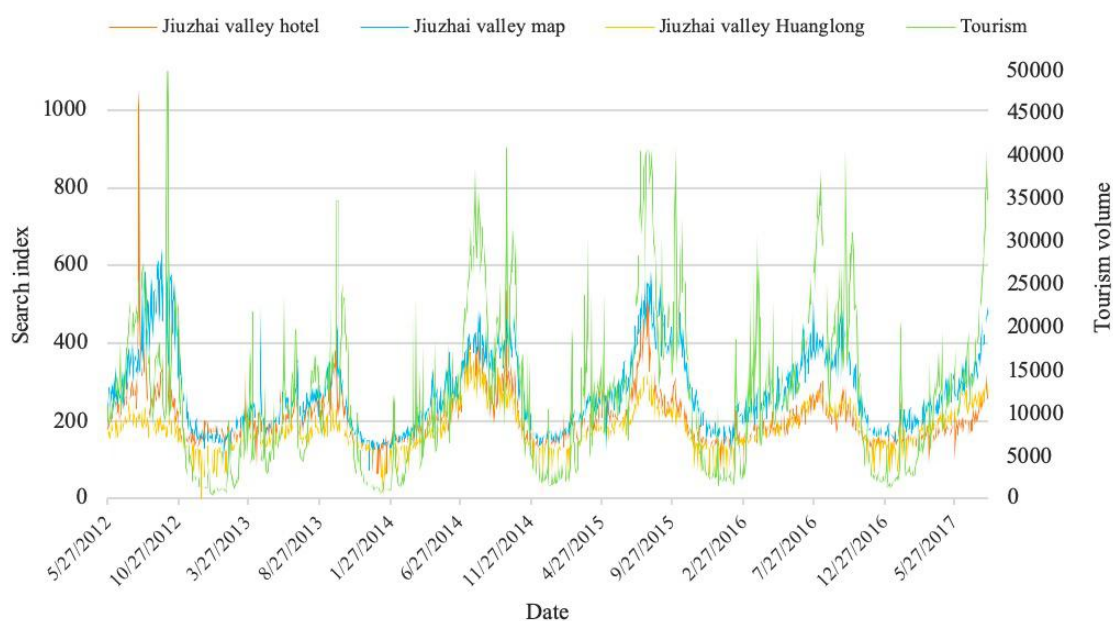
Figure 1. Visualization of the search index and tourism volume

(3) Stationary test

The augmented Dickey-Fuller (ADF) test method is first applied to examine the stationarity of both explanatory and explained variables, and the test results are shown in Table 2. It can be concluded that the original series of some variables like *Tourist volume*, *Jiuzhai Valley*, *Jiuzhai Valley hotel* are stationary , which do not need further transformation. On the other hand, the original non-stationary series like *Jiuzhai Valley map* and *Jiuzhai Valley scenic spot,* are stationary after transformation by taking their first order difference.

Table 2.   Results of variables' stationary test

|	Variables	ADF (original series)	ADF (first order difference series)	T code
Explained Variables	*Tourist volume*	0.003***	--	0
Explanatory Variables (predictors)	*Jiuzhai Valley hotel*	0.001***	--	0
	Jiuzhai Valley map	0.013	0.000***	1
	Jiuzhai Valley free travel	0.056	0.000***	1
	Jiuzhai Valley guide	0.655	0.000***	1
	Jiuzhai Valley scenic spot	0.091	0.000***	1
	Jiuzhai Valley scenic map	0.090	0.000***	1
	Jiuzhai Valley Huanglong	0.101	0.000***	1
	Jiuzhai Valley	0.000***	--	0

Note: ADF: the p-value statistics from the augmented Dickey-Fuller unit root tests. The asterisks ***: rejections of the null hypothesis at 1% significance level. T code: transformation code, 0 means no transformation; 1 means transformation with the first-order difference.

4.   EXPERIMENTAL RESULTS

This section presents experimental results using DMA model with the aforementioned eight selected explanatory variables.

4.1 The average number of predictors in DMA

With the DMA approach allowing explanatory variables to change over time, the numbers of explanatory variables may be different at each time point according to the forecasting performance. The average number of predictors at each time point is presented in **Figure 2**. From this figure, it is clear that the average number of explanatory variables varies over time. At the beginning of 2012, less than four explanatory variables are used, which are under half of the whole eight alternative predictors. Then the number rises over four by the end of 2012. And from then on, the fluctuation about the average number of predictors at each time point is relatively stable, mainly between 3 to 6. Overall, the average number of predictors is 4.17, which is less than eight alternative predictors. The results indicate that not all the search indexes are useful in predicting tourism volume, and the DMA can select good predictors effectively.
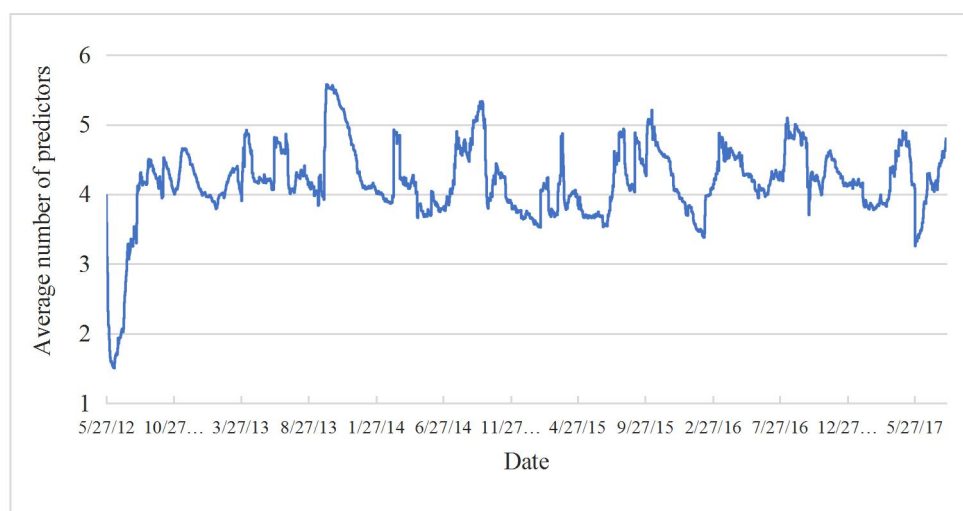
Figure 2. The average number of predictors in DMA

4.2 The number of predictors for the optimal forecasting model

According to the selection probability of the prediction model at each time point, the prediction model with the highest selection probability can be referred to as the optimal prediction model at each time point. The number of predictors for the optimal prediction model at each time point is presented in Figure 3. As indicated in the figure, most optimal models use only 2.8 predictors on average from 8 potential predictors. Besides, only few optimal models have more than six explanatory variables. The results further provide evidence that, at different time points, the search index variables play different roles. By applying the DMA approach, the time-varying effect is likely to be considered, resulting in less explanatory variables in the prediction modelling and less computation cost.
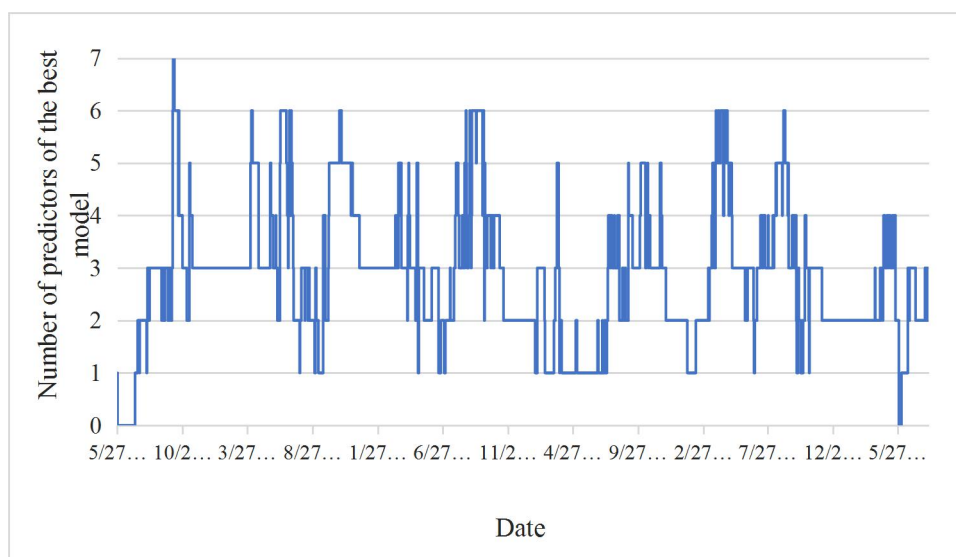

Figure 3. The number of predictors for the optimal forecasting model

4.3 Inclusion probability of predictors in DMA modelling

The inclusion probability of a predictor indicates the sum probabilities of the predictor employed among all potential predictors in the forecasting model. A predictor with high probability presents its strong prediction power for tourism volume. The probabilities of the eight predictors that were selected in the forecasting model are shown in Figure 4.

As Figure 4 presents, the predictors for tourism volume forecasting vary over the whole time period. At the year of 2012, *Jiuzhai Valley hotel* and *Jiuzhai Valley* present higher forecasting power than the other predictors with the inclusion probability over 80%. Then the inclusion probability of *Jiuzhai Valley hotel* drops rapidly. The forecasting power of *Jiuzhai Valley free travel* is at a low level from 2012. But around 2014 the inclusion probability of *Jiuzhai Valley free travel* rises to over 70% while at the same time *Jiuzhai Valley scenic spot, Jiuzhai valley guide, Jiuzhai Valley map, Jiuzhai Valley Huanglong* are at a low forecasting level. Before the year 2014, *Jiuzhai Valley map* appears an upsurge but gradually lose its forecasting power over time. After that, *Jiuzhai Valley scenic spot, Jiuzhai Valley guide, Jiuzhai Valley map* and *Jiuzhai Valley Huanglong* have similar fluctuations. At the middle of the year 2014, the inclusion probability of *Jiuzhai Valley scenic spot* increases significantly (up to 90%). Thus, *Jiuzhai Valley scenic spot* overtakes other predictors to become one of the best predictors. Generally speaking, *Jiuzhai Valley Huanglong, Jiuzhai Valley scenic spot, Jiuzhai Valley scenic map* do not show strong prediction power during the whole time period with the average inclusion probability around 42%. *Jiuzhai Valley hotel* and *Jiuzhai Valley* have the highest average inclusion probability of 78.4% and 76.1%, which overtakes the others. Therefore, *Jiuzhai Valley hotel* and *Jiuzhai Valley* show superior forecasting power to tourism volume.

In conclusion, search indexes do have time-varying effect in predicting tourist volume of Jiuzhai Valley.

Figure 4. Inclusion probability of predictors

5. CONCLUSIONS

Although previous studies have provided sufficient evidence that the search index can improve the prediction of tourism volume, few studies investigated the time-varying effect of search index in building prediction models of tourism volume. In this study, we fill this gap by proposing a DMA-based model and exploring the variation of search indexes in predicting tourism volume of Jiuzhai Valley. Based on the analysis of the average number of predictors in DMA and the optimal prediction models, and the inclusion probability of

predictors at different time period, we find that these predictors have time-varying effect in predicting tourism volume of Jiuzhai Valley. In particular, some keywords-related search index such as *Jiuzhai Valley hotel*, *Jiuzhai Valley*, *Jiuzhai Valley map* and *Jiuzhai Valley guide* show higher probabilities in supporting the prediction of tourism volume. This study provides both scholars and industry practitioners a clear understanding of relations between search indexes and tourism volume.

In the future, this study will include more relevant search indexes and explore the forecasting power of more predictors. The prediction performance of DMA-based model and other counterparts will also be considered for further comparisons.

ACKNOWLEDGEMENT

REFERENCES

[1] Song H, Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. Tourism Management, 29 (2): 203–220.

[2] Emili S, Figini P, Guizzardi A. (2020). Modelling international monthly tourism demand at the micro destination level with climate indicators and web-traffic data. Tourism Economics, 26 (7): 1129-1151.

[3] Sun S, Wei Y, Tsui, K.-L, Wang S. (2019). Forecasting tourist arrivals with machine learning and internet search index. Tourism Management, 70: 1–10.

[4] Liu H, Liu Y, Wang Y, Pan C. (2019). Hot topics and emerging trends in tourism forecasting research: A scientometric review. Tourism Economics, 25 (3): 448–468.

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

[6] Huang X, Zhang L, Ding Y. (2017). The Baidu Index: Uses in predicting tourism flows -A case study of the Forbidden City. Tourism Management, 58: 301–306.

[7] Wen L, Liu C, Song H. (2019). Forecasting tourism demand using search query data: A hybrid modelling approach. Tourism Economics, 25 (3): 309–329.

[8] Koop G, Korobilis D. (2012). FORECASTING INFLATION USING DYNAMIC MODEL AVERAGING. International Economic Review, 53 (3): 867–886.

[9] Wei Y, Cao Y. (2017). Forecasting house prices using dynamic model averaging approach: Evidence from China. Economic Modelling, 61: 147–155.

[10] Raftery A.E, Kárný M, Ettler P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. Technometrics, 52 (1): 52–66.

[11] Bork L, Møller S.V. (2015). Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection. International Journal of Forecasting, 31 (1): 63–78.

[12] Risse M, Ohl L. (2017). Using dynamic model averaging in state space representation with dynamic Occam's window and applications to the stock and gold market. Journal of Empirical Finance, 44: 158–176.

[13] Naser H, Alaali F. (2018). Can oil prices help predict US stock market returns? Evidence using a dynamic model averaging (DMA) approach. Empirical Economics, 55 (4): 1757–1777.

[14] Kalman R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering, 82 (1): 35–45.