Advancing Infrastructure Development through a Novel Federated Learning Design for Visitor Privacy Preservation Solutions

Qiong Chen[1*]

[1]College of Home Economics, Hebei Normal University, Shijiazhuang, 050310, China

Corresponding author: Qiong Chen, 13755331114@163.com

Abstract: With the rise of tourism, competition in the tourism industry is becoming increasingly fierce. Combined with relevant technologies of AI that intelligent tourism analysis model is discussed. A multi-layer big data analysis model is proposed, which provides interfaces for query engines, semantic search, and analysis. This paper proposes a prediction method of tourist throughput based on federated learning. First, we preprocess and normalize the Hainan tourism data participating in the training, and then use the logical regression model for training. At the same time, we use homomorphic encryption to ensure the privacy and security of the data. Finally, we train a model with strong generalization ability through server aggregation. In addition, we also use real Hainan tourism throughput data to prove the feasibility and effectiveness of the model. The results demonstrate that this model has good recommendation accuracy.

Keywords: smart tourism; Big data analysis; federated learning; Logical regression; Tourism forecast; Advancing Infrastructure Development
DOI: https://doi.org/10.14733/cadaps.2023.S15.264-273

1 INTRODUCTION

The development of Hainan home stay tourism is based on rich tourism resources. Hainan is in the tropical area. The unique tropical climate and rich tropical crops make people yearn for it. Many tourists choose to escape from the cold in winter in Hainan; Home stay providers take advantage of the diversity of tropical crops in Hainan to develop more tourism projects, such as building tropical fruit gardens, tropical flower gardens, etc., to enrich tourists' travel life and enhance tourism competitiveness; Hainan is surrounded by the sea on three sides and has many islands. With the goal of building tourist attractions for leisure and winter vacation, Hainan encourages local farmers to develop home stay tourism, opens green entrepreneurship channels for entrepreneurs returning home to participate in home stay construction projects, and provides all-round support. While developing homestay tourism, Hainan focuses on building beautiful villages, greatly improving the rural living environment, and successfully creating various ecological tourism routes such as rural family fun.

In smart tourism services, as IoT, AI, big data, WSN and NFC have been used to provide tourists[16]. For example, visitors can download applications for leisure parks. This will shorten the waiting time for visitors and improve the overall utilization of the facility and visitor satisfaction[4]. To further improve the efficiency of tourism management and service and optimize the travel experience of tourists, smart tourism big data analysis model is discussed in combination with AI and big data technology, and efficient recommendation strategies are designed[19]. The passenger flow of Hainan tourism is an important production indicator of Hainan tourism of civil aviation, which is the basis for effective distribution of Hainan tourism resources, and also an important basis for investment decisions on Hainan tourism projects. In recent years, many scholars have proposed methods to predict traffic flow. However, due to the price policy and military aircraft drills of Hainan tourism, the data of a single Hainan tourism cannot be used to predict the tourist throughput. As the data related to civil Hainan tourism involves issues such as trade secrets and tourists' privacy security, it is difficult to summarize, and the phenomenon of data islands is serious[15],[17].

Google has put forward a new framework, called federated learning, which effectively solves the problem of privacy security[18]. In the training of federated learning mode, each client participates in the training of the model and can store its data locally without uploading. Therefore, each import the training model or gradient to the server for summary, and the summarized model or gradient information will be sent to the client by the server[7],[6].

These infrastructure components are essential for facilitating the smooth operation of federated learning, ensuring efficient data transmission, and maintaining the privacy and security of client data. By investing in infrastructure development, organizations can effectively implement federated learning and leverage its benefits in a secure and scalable manner.

In the process of federated learning, homomorphic encryption scheme is most often used. Homomorphic encryption is a special algorithm that can directly encrypt your own data[9]. Due to different transmission energies, these encryption schemes can generally be divided into partial homomorphic encryption schemes, finite homomorphic encryption schemes, and all homomorphic encryption schemes[8]. For example, Paillier scheme only supports addition between ciphertext, but does not support multiplication between ciphertext[24]. For example, the Bonh Goh Nissim scheme can support infinite homomorphic addition, but only one homomorphic multiplication can be supported at most.

In this paper, a throughput prediction model based on federated learning is designed. First, each Hainan tourism company predicts and processes historical data, normalizes the throughput data after removing outliers, and eliminates the impact of dimensions; Secondly, through the combination of logical regression algorithm and homomorphic encryption, Hainan tourism can share and train a prediction model without disclosing relevant data.

2 PREPARATIONS

2.1 Federated Learning

Federated learning is a new distributed machine learning technology. The technical purpose is to ensure information security and legal compliance. Through efficient machine learning for each participating node, it can better conduct collaborative training, to obtain the overall model[11]. The basic algorithms in the research are not limited to statistical machine learning technology, but also include deep neural networks which are developing rapidly at present. The specific objective function of federated learning is shown in equation (1):

$$\min_w \sum_{k=1}^{m} p_k F_k(w) \tag{1}$$

Where m is the number of participants, $P_k \geqslant \mathbf{O}$ and $\sum_k P_k = 1$, F is the local optimization objective function of the kth participant. The local objective function is generally defined by empirical risk loss on data, as shown in equation (2):

$$F_k(w) = \frac{1}{n_k} \sum_{j_k=1}^{n_k} f_{j_k}\left(w; x_{j_k}, y_{j_k}\right) \tag{2}$$

Then the new parameters obtained in the t-round iteration are shown in equation (3):

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \tag{3}$$

Each participant is as follows (4):

$$w_{t+1}^k \leftarrow w_t^k - \eta \nabla F_k\left(w^k\right) \tag{4}$$

2.2 Logical Regression of Homomorphic Encryption

Paillier semi homomorphic encryption algorithm was proposed in 1999[3]. It can process encrypted data, and the calculation result is still encrypted. Users with keys can decrypt the encrypted result[20].

This training model is a logical regression model, so the activation function used is $g(\theta x) = \frac{1}{1+\mathrm{e}^{-\theta x}}$,At $g(z) \geqslant 0.5$, the label is 1, At $g(z) < 0.5$, the label is 0, Its objective function is shown in equation (5):

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(-y_i \boldsymbol{\theta} \mathbf{x}_i + \ln\left(1+\mathrm{e}^{\theta x_i}\right)\right) \tag{5}$$

Assume that the parameters of Hainan tourism participating in the training are respectively θ^A, θ^B, Then the objective function of the two together is shown in equation (6):

$$L = \sum_{i=1}^{N} \left(-y_i\left(u_i^A + u_i^B\right) + \ln\left(1+\mathrm{e}^{u_i^A + u_i^B}\right)\right) \tag{6}$$

Then the model parameters of Hainan Tourism A and Hainan Tourism B are updated as follows (7):

$$\boldsymbol{\theta}^A := \boldsymbol{\theta}^A - \eta \frac{\partial L}{\partial \boldsymbol{\theta}^A}$$
$$\boldsymbol{\theta}^B := \boldsymbol{\theta}^B - \eta \frac{\partial L}{\partial \boldsymbol{\theta}^B} \tag{7}$$

Since Paillier encryption algorithm only supports additive homomorphisms and scalar multiplication homomorphisms, the literature uses Taylor expansion to approximate the original logarithmic loss[14]. In this paper, we first expand the Taylor expansion of the logarithmic loss function log 1+e-z() at z=0, and the expression is equation (8):

$$\log\left(1+e^{-y\theta^{\mathrm{T}}\mathbf{x}}\right)\approx\log 2-\frac{1}{2}\mathbf{y}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}+\frac{1}{8}\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}\right)^{2} \tag{8}$$

The last item in the formula is directly removed by y because of $y^2=1$, as shown in equation (9):

$$L=\frac{1}{n}\sum_{i=1}^{n}\left\{\log 2-\frac{1}{2}\mathbf{y}_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i+\frac{1}{8}\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i\right)^{2}\right\} \tag{9}$$

Therefore, the corresponding encrypted gradient is as follows (10):

$$\left\|\frac{\partial L}{\partial\boldsymbol{\theta}}\right\|=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{4}\left[\left[\boldsymbol{\theta}^{\mathrm{T}}\right]\right]\mathbf{x}_i+\frac{1}{2}[[-1]]\mathbf{y}_i\right)\mathbf{x}_i \tag{10}$$

3 TOURISM MODEL BASED ON FEDERATED LEARNING

In the model of this paper, firstly, the outliers of two Hainan tourism data are processed, and at the same time, the normalized processing is carried out to eliminate the impact of dimensions[26]. The method of logical regression is used for training, and the homomorphic encryption method is used for privacy protection. The federal server finally trains a model suitable for local Hainan tourism evaluation by aggregating the local models of the two. The federated learning framework of Hainan Tourism is shown in Figure 1.

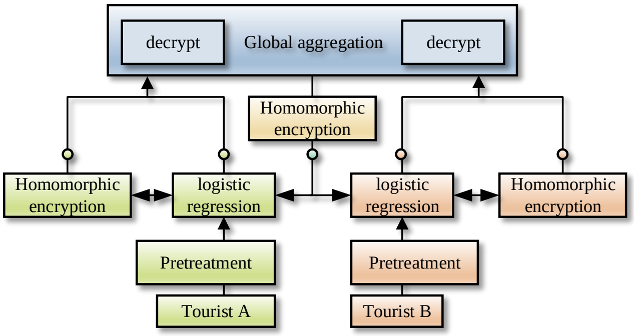

Figure 1: Hainan Tourism Federation Learning Framework.

3.1 Handling of Abnormal Values

Due to extreme weather and drilling events, Hainan tourism is prone to abnormal values. Therefore, when the tourist throughput value is not distributed in（μ— 3 σ，μ+ 3 σ) Will be judged as abnormal data. Remove abnormal values according to equation (11):

$$R_n=\left(x_n-\overline{x}\right)/\mu$$
$$R_n^*=\left(\overline{x}-x_n\right)/\mu \tag{11}$$

Among them, μ is the known overall standard deviation, $\overline{x}$ is the sample mean value, and the outlier is $R_n > R_{0.997}(n)$ or $R_n^* < R_{0.003}^*(n)$.

3.2 Training Process Under Homomorphic Encryption

In recent years, research has found that transmission through gradients will also lead to the risk of data privacy disclosure, so it is very important to carry out homomorphic encryption during the transmission of gradients[13]. The algorithm steps are as follows:

Step 1 :The throughput of Hainan Tourism A and Hainan Tourism B respectively generate a pair of public and private keys and send them to the server.

Step 2 :Hainan Tourism A and Hainan Tourism B calculate u_i^A and u_i^B of the local model respectively and encrypt them with public key, Send $\left[\left(u_i^A\right)^2\right]$ and $\left[\left(u_i^B\right)^2\right]$ to the server.

Step 3 :The server decrypts the model parameters of two Hainan tourism after receiving $\left[\left(u_i^A\right)^2\right]$ and $\left[\left(u_i^B\right)^2\right]$, At the same time, L_{AB} is calculated according to formula (6), and gradient is solved by formula (8), and then aggregation is performed.

Step 4 :The server encrypts in the same process and transmits it to Hainan Tourism A and Hainan Tourism B[22].

Step 5 :Hainan Tourism A and Hainan Tourism B get L_{AB} through decryption and calculate the gradient according to Formula (8), and then use the gradient descent method to update the parameters. After that, the homomorphic encryption is transmitted to the server again.

Repeat Step 1 to Step 5 until the model converges.

3.3 Prediction Process Under Homomorphic Encryption

When the server queries, the model is deployed in Hainan Tourism A and Hainan Tourism B, and the prediction process is like the above training process. This is to be explained as follows.

Step 1 The server divides the forecast data into x^A and x^B parts, Use the public key encryption x^A and x^B of the server to get $\left[x^A\right]_C$ and $\left[x^B\right]_C$ and send them to Hainan Tourism A and Hainan Tourism B respectively for calculation.

Step 2 Calculate $\left[u^A\right]_C = \theta^A \left[x^A\right]_C$ and $\left[u^B\right]_C = \theta^B \left[x^B\right]_C$ on Hainan Tourism A and Hainan Tourism B respectively and send them to the server [2].

Step 3 After the server decrypts, u^A and u^B are obtained. $u = u^A + u^B$ Calculate the final output result $\frac{1}{1+e^{-u}}$.

4 EXPERIMENT

4.1. Experimental Set and Data Set

According to the air passenger throughput forecast in this paper, since the selected experimental data set has no missing values, but there are outliers, the data set is normalized, and outliers are eliminated[1,5]. The data set selected in this paper is two data sets of daily tourist throughput of Hainan tourism from January 1, 2019, to December 31, 2019.

The experimental environment is configured as Ubuntu 21.0 operating system, IntelCore i5-8300H, 8 GB memory, Python 3.6 framework, and RTX 3080 video card model [25]. Where, the learning rate of the model is 0.01, the momentum is 0.9, the number of iterations is 50, and 80% is the training value, and the remaining 20% is the prediction value.

4.2. Comparative Experimental Analysis

The comparison between the real data of Hainan Tourism A and the logical regression prediction is shown in Figure 2, the comparison between the real data of Hainan Tourism A and the federated learning prediction is shown in Figure 3, the comparison between the real data of Hainan Tourism B and the logical regression prediction is shown in Figure 4, and the comparison between the real data of Hainan Tourism B and the federated learning prediction is shown in Figure 5. It can be seen from Figure 2 that in the process of training for a single Hainan tourism, due to certain noise, the model has a problem of high prediction or large fluctuation in the fitting process. Both Figure 2 and 4 have a large jump or high prediction value. The training value after the federated study, whether Hainan Tourism A or Hainan Tourism B, has a very high fitting degree, especially the fitting of Hainan Tourism A is very close to the truth, whether in terms of time fluctuation or prediction error value.

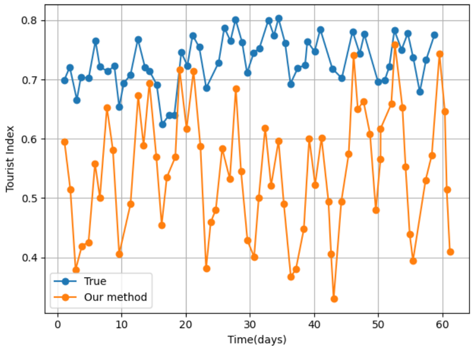

Figure 2: Comparison of Tourist's Real Data and Logistic Regression Forecast.

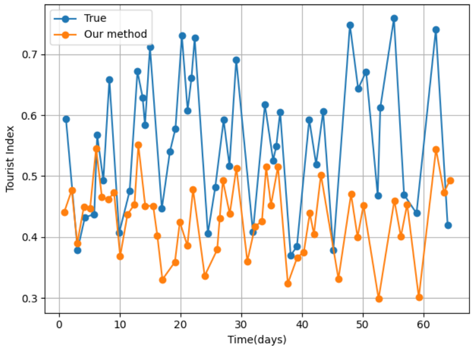

Figure 3: Comparison between real data of Tourist and federated learning prediction.

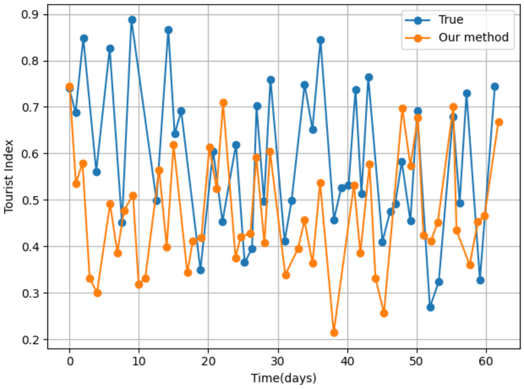

Figure 4: Comparison between the real data of Tourist B and the logistic regression prediction.

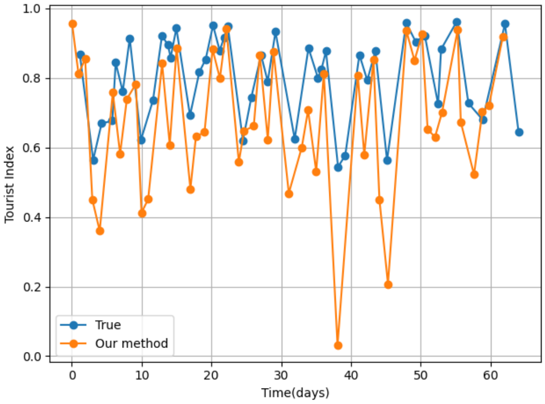

Figure 5: Comparison of Tourist B Real Data and federated learning Forecast.

See Table 1 for the comparison results of MPAE of logical regression and federated learning regression. The comparison between the MPAE values of the two methods of logical regression and federated learning shows that if logical regression is directly used for prediction, the accurate values obtained are 0.220 4 and 0.191 4 respectively, and the model after aggregation of the two has a significant improvement effect. This can also prove the feasibility and effectiveness of this scheme.

	Tourist	*Tourist B*
logical regression	*0.2205*	*0.1915*
federated learning	*0.1548*	*0.1493*

Table 1: MPAE of logistic regression and federated learning.

4.3. Spatial Equilibrium Characteristics

Through the factor detector, we can identify the impact of each factor on the high-level tourist attractions and obtain the q value of each impact factor (Table 2). Then the explanatory power of each factor on the spatial differentiation of high-level tourist attractions is as follows: city (0.45)>traffic (0.37)>NDVI (0.32)>DEM (0.29)>population (0.20)>GDP (0.12)>water system (0.04). This reveals that urban distribution, traffic conditions, vegetation coverage and terrain fluctuation (greater than the average q score of 0.26), while population, social and economic development level (GDP) and river system distribution have less influence on high-level scenic spots.

Statistical value	*traffic*	*river system*	*population*	*city*	*DEM*	*NDVI*	*GDP*
q	*0.38*	*0.05*	*0.21*	*0.46*	*0.29*	*0.33*	*0.13*
p	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*

Table 2: Action intensity of each factor (q value of geographical detector).

The suitability of high-level tourist attractions is simulated and predicted by applying deep learning technology. Squeeze Net model in convolutional neural network, as a representative network in deep learning, has simple structure, high calculation efficiency and high accuracy. After complete sampling, obtain the sample data set that simulates the spatial differentiation characteristics of scenic spots, and input it into the Squeeze Net model for repeated iterations to obtain the training model; Finally, the model is used to predict the study area. The higher the value, the more suitable for the development tourist attractions, and vice versa (Figure 6).

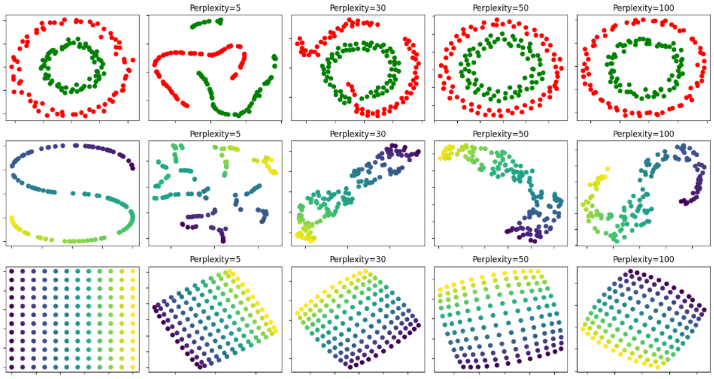

Figure 6: Deep learning prediction of the spatial pattern of the scenic spot.

As a whole, the southwest region is very suitable for developing high-level scenic spots, followed by coastal areas and the three northeastern provinces, and the northwest region is less suitable; From the perspective of urban agglomerations, the urban agglomerations in central Yunnan, Lanxi, Guanzhong Plain and central and southern, followed by Beibu Gulf urban agglomerations, Pearl River Delta urban agglomerations, cross-strait urban agglomerations, Harbin Great Wall urban agglomerations, central and southern Liaoning urban agglomerations and Tianshan northern slope urban agglomerations(Figure 6). In terms of DEM, the average altitude of 2849 m is very suitable for tourist, while the average altitude of 1566 m is not suitable for tourist attractions. Overall, the higher the altitude is, the more suitable it is for scenic spots, and the change law does not show a linear growth trend.

5 CONCLUSION

To develop Hainan's tourism, this paper will analyze the advantages and current situation of Hainan's home stay tourism, grasp the tourists' travel trends and psychology, actively explore the direction of tourism development in the new era, and provide suggestions for promoting the development of Hainan's home stay tourism. Through the prediction of the tourist throughput, Hainan tourism can choose to carry out maintenance or other construction activities in the period of less passenger flow, to minimize the impact on the operation and management of Hainan tourism. In this paper, the federated learning is applied to enable two Hainan tourists to train a common model when privacy is involved, and homomorphic encryption is used in the process of transferring the model, thus ensuring that the data will not be disclosed.

In the future, Hainan tourist image segmentation will be combined with data from other sensors and extended to the depth learning model.

Qiong Chen, https://orcid.org/0009-0001-6076-2192

REFERENCES

[1] Alzahrani, H. A.: The direct cost of diabetic foot management in some of private hospitals in jeddah, saudi arabia, International Journal of Diabetes in Developing Countries, 33(1), 2013,34-39. https://doi.org/10.1007/s13410-012-0107-x

[2] Arnold, J. M.; Hussinger, K.; Exports versus fdi in german manufacturing: firm performance and participation in international markets, Review of International Economics, 18(4), 2010, 595-606. https://doi.org/10.1111/j.1467-9396.2010.00888.x

[3] Atter, N.V.; Moyer, R. A.; Beck, C. J.; Mclane, A.: Advocacy services for survivors of intimate partner violence: pivots and lessons learned during the covid-19 quarantine in tacoma, Washington, Family Court Review, 60(2), 2022, 288-302. https://doi.org/10.1111/fcre.12642

[4] Buonincontri, P.; Micera, R.: The experience co-creation in smart tourism destinations: a multiple case analysis of European destinations, Information Technology & Tourism, 16(3), 2016, 285-315. https://doi.org/10.1007/s40558-016-0060-5

[5] Catherine, F.: Cinema that stays at home: the inexportable films of belgium's gaston schoukens, edith kiel and jan vanderheyden, Screen(3), 2010, 256-271. https://doi.org/10.1093/screen/hjq015

[6] Cheah, W. H.; Singaravelu, H.: The coming-out process of gay and lesbian individuals from Islamic Malaysia: Communication strategies and motivations, Journal of Intercultural Communication Research, 46(5), 2017, 401-423. https://doi.org/10.1080/17475759.2017.1362460

[7] Chen, C.; Liu, C.; Lee, J.: Corruption and the quality of transportation infrastructure: evidence from the US states, International Review of Administrative Sciences, 88(2), 2022, 552-569. https://doi.org/10.1177/0020852320953184

[8] Hack, K.: Detention, Deportation and Resettlement: British Counterinsurgency and Malaya's Rural Chinese, 1948-60, The Journal of Imperial and Commonwealth History, 43(4), 2015, 611-640. https://doi.org/10.1080/03086534.2015.1083218

[9] Hernández-Romero, I. M.; Nápoles-Rivera, F.; Mukherjee, R.; Serna-González, M.; El-Halwagi, M. M.: Optimal design of air-conditioning systems using deep seawater, Clean Technologies and Env. Policy, 20(3), 2018, 639-654. https://doi.org/10.1007/s10098-018-1493-7

[10] Javed, M.; Tučková, Z.: The role of government in tourism competitiveness and tourism area life cycle model, Asia Pacific Journal of Tourism Research, 25(9), 2020, 997-1011. https://doi.org/10.1080/10941665.2020.1819836

[11] Jingchun Zhou; Dehuan Zhang; Wenqi Ren; Zhang Weishi.: Auto Color Correction of Underwater Images Utilizing Depth Information, 19, 2022, 1-5, IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1109/LGRS.2022.3170702

[12] Lau, C. K.; Huang, J.; Feng, S. Y.; Qiu, H.:. Profiling trusted information sources for Chinese tourists traveling to Pacific SIDS, Journal of Global Scholars of Marketing Science, 32(1), 2022, 77-96. https://doi.org/10.1080/21639159.2020.1808834

[13] Liao, Z. X.; Peng, G.; Ren, P. Y.; Luo, Y. Y.; Zhang, X. P.; Feng, G.: Research on prediction of tourists' quantity in jiuzhai valley based on ab@g integration model, Tourism Tribune, 28(4), 2013, 88-93. https://doi.org/10.1504/IJEP.2013.054028

[14] Liu, K.; Dong, X.; Zhou, Y.; Duan, P.; Cheng, J.: Three-proof optimization of gas turbine generator set of offshore platform, Chemistry and Technology of Fuels and Oils, 58(5), 2022, 820-827. https://doi.org/10.1007/s10553-022-01457-6

[15] Ma, X.; Wu, W.: Deficiencies in China's island development processes compared with other countries, Emerging Markets Finance and Trade, 56(13), 2020, 2963-2976. https://doi.org/10.1080/1540496X.2019.1644498

[16] Palau-Saumell, R.; Forgas-Coll, S.; Amaya-Molinar, C. M.; Sánchez-García, J.: Examining how country image influ-ences destination image in a behavioral intentions model: The cases of Lloret De Mar (Spain) and Cancun (Mexico), Journal of Travel & Tourism Marketing, 33(7), 2016, 949-965. https://doi.org/10.1080/10548408.2015.1075456

[17] Pegas, F. D. V.; Weaver, D.; Castley, G.: Domestic tourism and sustainability in an emerging economy: Brazil's littoral pleasure periphery, Journal of Sustainable Tourism, 23(5), 2015, 748-769. https://doi.org/10.1080/09669582.2014.998677

[18] Qin, Y.; Qin, J.; Liu, C.: Spatial-temporal evolution patterns of hotels in China: 1978-2018, International Journal of Contemporary Hospitality Management, 33(6), 2021, 2194-2218.

[19] Sarky, S.; Wright, J.; Edwards, M.: Evaluating consistency of stakeholder input into participatory GIS-based multiple criteria evaluation: a case study of ecotourism development in Kurdistan, Journal of Environmental Planning and Man-agement, 60(9), 2017, 1529-1553. https://doi.org/10.1080/09640568.2016.1236013

[20] Su, M.; Pan, T.; Chen, Q. Z.; Zhou, W. W.; Gong, Y.; Xu, G.: Data analysis guidelines for single-cell rna-seq in biomedical studies and clinical applications, Military Medical Research, 9(1), 2022, 1-24. https://doi.org/10.1186/s40779-022-00434-8

[21] Weaver, D.: Creative periphery syndrome? Opportunities for sustainable tourism innovation in Timor-Leste, an early stage destination, Tourism Recreation Research, 43(1), 2018, 118-128. https://doi.org/10.1080/02508281.2017.1397838

[22] Xiangyang, Y. U.; Shanfeng, H. U.; Zhu, G.; Deming, L. I.: Research on medium-term prediction of tourist arrivals in scenic areas based on least squares support vector machines, Tourism Tribune, 28(4), 2013, 75-82.

[23] Zhang, B.; Li, N.; Law, R.; Liu, H.: A hybrid MIDAS approach for forecasting hotel demand using large panels of search data, Tourism Economics, 28(7), 2022, 1823-1847. https://doi.org/10.1177/13548166211015515

[24] Zhang, C.; Li, M.; Wu, D.: Federated Multidomain Learning With Graph Ensemble Autoencoder GMM for Emotion Recognition, IEEE Transactions on Intelligent Transportation Systems (2022). https://doi.org/10.1109/TITS.2022.3203800

[25] Zhou, H.; Wu, T.; Sun, K.; Zhang, C.: Towards high accuracy pedestrian detection on edge gpus, Sensors, 22(16), 2022, 5980. https://doi.org/10.3390/s22165980

[26] Zhu, T. X.; Ren, J. X.;Zhang, X. Y.: Research on the tourists demand of high-speed rail traffic in beijing-tianjin-hebei region based on utility function, Journal of Railway Engineering Society, 35(3), 2018, 102-108.