Tourist Analyzer

Yash Maroo[1], Sairaj Marshetty[2], Aditi More[3], Pawan Gond[4], Hemalata Mote[5]

[1,2,3,4]B.E. Student, Dept. of Electronics and Telecommunication, Atharva College of Engineering, Mumbai, India
[5] Professor, Dept. of Electronics and Telecommunication, Atharva College of Engineering, Mumbai, India

---***---

Abstract - *Tourism is one of the massive and rapid growing economic sectors in the world. Since tourist stays are generally of short duration, it may be interesting to analyze daily behaviors of tourists. Based on geo-located information left by tourists on community websites of photo-sharing, we propose an original approach to analyze daily behaviors of tourists by analyzing sequences of places visited by tourists per day. Our work would be beneficial for the tourists as well as the people involved in the tourism industry by facilitating quicker services to the tourists. We experimented our approach with data from Tourpedia website about tourism in Paris.*

Key Words**: Tourism, geo-located information, daily tourism behaviour, Tourpedia.**

1. INTRODUCTION

The tourism industry in India generated about 16.91 Lakh Crore or US$240 billion in the year 2018 which was 9.2% of India's GDP. It was responsible for generating 42.673 million jobs or 8.1% of India's total employment. With such a large contribution, it is important that the government and the local authorities focus on improving this sector of the economy with the right research and the investments in the right places. The analysis of the tourist behaviour is the most important indicator or predictor of future tourist behaviour.

Taking under consideration the social role of the tourist, the behaviour of a tourist also can be an indicator of the behaviour of others. With their conduct, sightseers set the social standards of conduct within background of tourism. These standards are being followed by other consumers, those who have not yet engaged in or are engaged in travel or tourism activities. A tour operator must be able to assess whether the development, marketing and implementation of tourism activities are pertinent to their marketing and operational approaches, as it's essential to perceive the various types of behaviour at every point.

Only by knowing the fundamentals of tourist behaviour, as well as knowing how to observe and measure them, we can effectively plan offers and other sales activities in tourism. The theoretical basis is very important in empirical research or measurement of tourism behaviour, as it also shows the concepts to be measured and, in many cases, how to measure them.

This paper is based on a real life problem, faced by several tourists and businesses involved in tourism. If the problem is solved and their actions are clubbed collectively, it can produce phenomenal results and lead to enormous profits. Also after studying a few papers, we realised that major research has been done on tourist recommendation systems rather than tourist behaviour analysis. Also, research done on tourist behaviour analysis includes more information regarding the tourist attraction than the tourists themselves. So, we aim to provide a crystal clear analysis about the behaviour of the tourists. Due to the COVID-19 crisis, the tourism industry has been facing a lot of repercussions. The pandemic has a huge impact on the tourism industry due to the resulting travel restrictions as well as slump in demand among travellers.

We aim to analyse tourists behaviour based on the locations and places they have visited so far, to identify tourist interests, tourism demographics and to plan future tourism demands. It supports strategic decision-making in tourism destination management. We are going to make use of geo-tagged located information available on community websites like Tourpedia for datasets. Also we would structure the tourist demographic data for all the locations in the vicinity. Make geographical clusters to identify popular tourist locations from tourist interests. Construct a time series data to show the number of tourists at a particular spot throughout the year.

The paper is structured as - Section 2 presents the Literature Review, Section 3 presents the Data-set Description, Section 4 presents the Methodology and Section 5 presents the Conclusion.

2. LITERATURE REVIEW

In paper [1], a design science research (DSR) methodology was adopted, where the seven design guiding principles of Hevner et al. [2] are used to design, evaluate and communicate the solution. As defined by March and Smith [3], design artefact is specified as a method designed to process and analyse social media big data, such as geo-tagged photos and geo-located information, together with their joined personal and workflow, to support destination management organisations (DMO's) strategic analyzing within the context of Tourist destination management.

Authors of the paper [4] propose P-DBSCAN, a new density-based clustering algorithm based on DBSCAN for analysis of places and events using a collection of geo-tagged photos. It is similar to DBSCAN for analysis of places and events using a collection of geo-tagged photos. Representative historic images were found in the city and country scales in combining coordinates of geo-tagged photos with content based and textual analysis using Mean-Shift algorithm based on kernel-density estimation for clustering.

The paper [5] presents a framework to identify the interests of tourists by integrating information carried by the geo-tagged photos shared on social media websites. Such strategy is expected to provide sustainable tracking on point of interest (POIs) updated by tourists and pick the best representative photos taken by them. The performance of the model was evaluated by conducting a case study using the geo-tagged photos taken in Hong Kong.

The authors of the paper [6] describe tourist behavior mining from analyzing photo content by using a computer deep learning model. 35356 Flickr tourist's photos are identified into 103 scenes and analyzed by ResNet-101 Deep learning model. Tourist's cognitive maps with diverse perceptual topics are visualized by the creators agreeing to photographic geological data. Statistical analysis and spatial analysis (by using hierarchical clustering analysis and ANOVA (analysis of variance)) are used for analyzing tourist behavior.

The authors of paper [7] presented a tourist behavior analysis system based on a digital pattern of life concept. The digital pattern of life extracts tourist behavior components in a convenient form for analysis and is based on an ontological approach, which allows to take tourist, city and POI context information into consideration. Digital pattern of life provides various convenient representation of the tourist regardless of source selection. Changes of tourist actions can be viewed in a specific time window, since the digital pattern of life information is stored with the time reference.

The paper [8] propose a mobile application, which will take the user's interest and recommend attractions, restaurants, and hotels. The system is trained using the dataset of Trip-Advisor. The clustering of the prepared dataset is done utilizing K-modes clustering which is an unsupervised learning calculation. Convolutional Neural Networks is used to reverse image search which is done for the dataset created by fragmenting images from Google. After this, the application received the data in the JSON format from the MySQL Database using Python Flask Server.

The work described in paper [9] propose a framework containing an improved cluster method and multiple neural network models to extract representative images of tourist attractions. A novel time- and user-constrained density-joinable cluster method (TU-DJ-Cluster) was proposed by the author which was specific to photos with similar geo-tags to detect place-relevant tags. Then there was merging and extending of the clusters according to the similarity between pairs of tag embedding's, as trained from Word2Vec.

Based on the clustering result, they filter noise images with Multilayer Perceptron and a single-shot multibox detector model, and further select representative images with the deep ranking model. The authors selected Beijing as the study area.

The authors of paper [10] presented an approach for exploring tourist's behavior based on the extracted dataset from geo-tagged photographs. The creators changed the dataset to a reasonable format and connected k-mean clustering to cluster the diverse tourists' behavior. After that, the tourist's behavior of each cluster, particularly behavior regarding interesting attractions, was analyzed by factual test.

The paper [11] presents a framework based on distributed Map / Reduce to carry out research and analysis of the flow behavior of tourists, with better efficiency and scalability. A Big Data Analytics platform was used for this paper to analyze the tracked data of tourists' mobile phone, find out the behavior patterns of tourists, and design an analysis of the tourist flows based on the traditional data warehouse, Hadoop cluster and the database of My SQL, which includes three modules.

The work described in paper [12] proposed an original approach to characterize daily behaviors of tourists by analyzing sequences of places visited during a day by each tourist based on geo-tagged and time-related information left by tourists by posting their photographs on photo sharing websites and twitter also. The authors of this had used R with TraMineR6 package that is based on the Needleman-Wunsch algorithm as optimal matching.

The paper [13] uses sequential patterns of tourist activities and locations from social media's as main source of behavior data. The Convolutional Long Short-Term Deep Learning method is used for prediction of the expected location. The proposed method combines Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM). The authors of this paper state that their solution outperforms other neural network models when evaluating with the accuracy and loss metrics.

3. DATASET DESCRIPTION

For user's profiles and places reviews database, the Tourpedia dataset has been used.

There are mainly 4 categories of places:

A] Accommodation

B] Attraction

C] Point of Interest

D] Restaurant

Each tuple has the following entries- address, id, latitude, longitude, location, name, original id and reviews. For our model, we are using 26928 datasets of Paris.

4. METHODOLOGY

4.1 GEOGRAPHICAL DATA CLUSTERING

Clustering is the task of grouping a set of objects in such a way that observations in the same group are more similar to each other than to those in other groups. It is one of the most popular applications of the Unsupervised Learning (Machine Learning when there is no target variable). Geospatial analysis is the field of Data Science that processes satellite images, GPS coordinates, and street addresses to apply to geographic models.

First of all, we will be importing libraries- For data (pandas, numpy), For plotting (matplotlib and seaborn), For geospatial (folium and geopy), For machine learning (scipy), For deep learning (minisom). Then we will read the data into a pandas Dataframe.

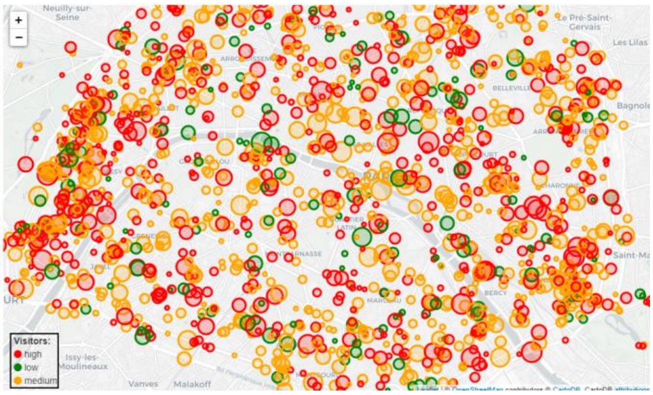


Fig -1: Clustering

K-Means algorithm [14] aims to partition the observations into a predefined number of clusters (k) in which each point belongs to the cluster with the nearest mean. It starts by randomly selecting k centroids and assigning the points to the closest cluster, then it updates each centroid with the mean of all points in the cluster. This algorithm is convenient when you need the get a precise number of groups, and it's more appropriate for a small number of even clusters.

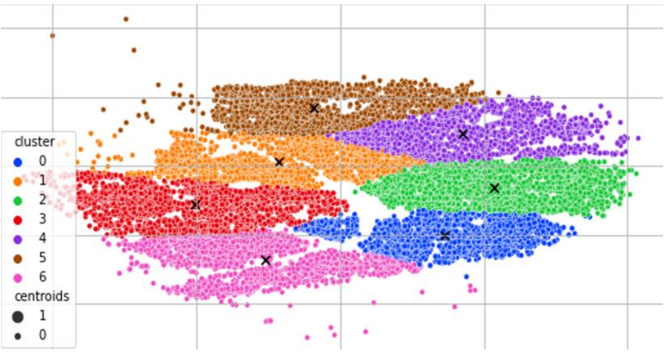


Fig -2: K-means algorithm

Here, in order to define the right k, we use the Elbow Method [15]: plotting the variance as a function of the number of clusters and picking the k that flats the curve. The elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.

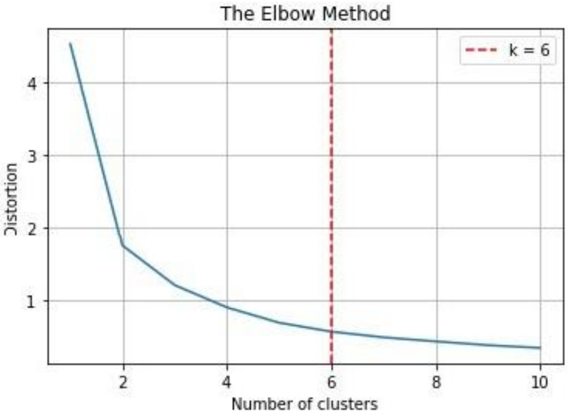


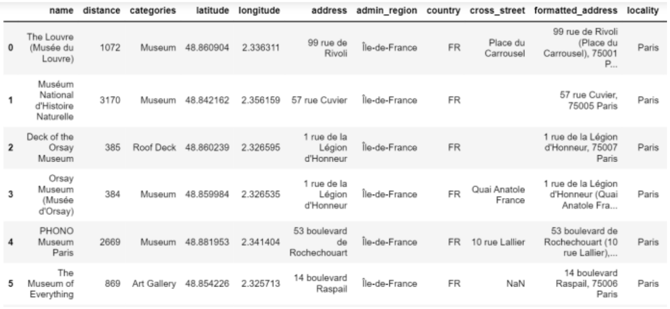
Fig -3: Elbow Method

4.2 POPULAR LOCATION IDENTIFICATION

People often tend to travel from one location to another location and explore new things, and in their journey, they need to know all the popular places around them so that they don't miss out on any. So for those people who need to know all the densely populated areas around them. input to this model is the current location and the radius of the search. We use FourSquare API, that gives all the popular places around a given location and python to visualize this stuff.

Model Development: First of all, we will be importing libraries- For data (pandas, numpy), For creating maps (folium), For retrieving information (requests). Then we will be converting address to coordinates and also converting JSON to Dataframe. Then we will be reading the current location from the user and coverting it to the coordinates followed by fetching data from the FourSquareApi, the result is a JSON data.

Then we will be cleaning the data and converting it to dataframe. Our data will be visualized as follows:



	name	distance	categories	latitude	longitude	address	admin_region	country	cross_street	formatted_address	locality
0	The Louvre (Musée du Louvre)	1072	Museum	48.860904	2.336311	99 rue de Rivoli	Île-de-France	FR	Place du Carrousel	99 rue de Rivoli (Place du Carrousel), 75001 P...	Paris
1	Muséum National d'Histoire Naturelle	3170	Museum	48.842162	2.356159	57 rue Cuvier	Île-de-France	FR		57 rue Cuvier, 75005 Paris	Paris
2	Deck of the Orsay Museum	385	Roof Deck	48.860239	2.326595	1 rue de la Légion d'Honneur	Île-de-France	FR		1 rue de la Légion d'Honneur, 75007 Paris	Paris
3	Orsay Museum (Musée d'Orsay)	384	Museum	48.859984	2.326535	1 rue de la Légion d'Honneur	Île-de-France	FR	Quai Anatole France	1 rue de la Légion d'Honneur (Quai Anatole Fra...	Paris
4	PHONO Museum Paris	2669	Museum	48.881953	2.341404	53 boulevard de Rochechouart	Île-de-France	FR	10 rue Lallier	53 boulevard de Rochechouart (10 rue Lallier),...	Paris
5	The Museum of Everything	869	Art Gallery	48.854226	2.325713	14 boulevard Raspail	Île-de-France	FR	NaN	14 boulevard Raspail, 75006 Paris	Paris

Fig -4: Data visualization

Below output depicts- a] Name of the location; b] What is the category, it is famous for; c] Address of the location; d] Distance from current location. Also the red marker shows the current location, whereas blue markers show the popular locations nearby.

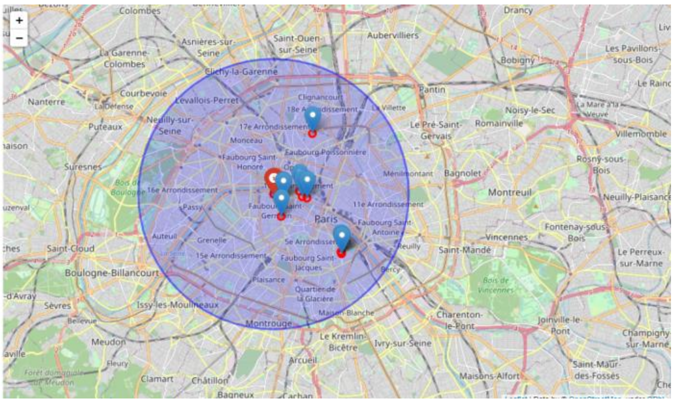


Fig -5: Popular locations nearby

4.3 TIME SERIES MODELLING

Time series is defined as a set of random variables ordered with respect to time. Time series are studied both to interpret a phenomenon, identifying the components of a trend, cyclicity, seasonality and to predict its future values. For time series, we have trend analysis which determine whether it is linear or not as most models require this information as input, outliers detection and seasonality analysis.

First of all, we will be importing libraries- For data (pandas, numpy), For plotting (matplotlib), For outliers detection (sklearn). Then we will read the data into a pandas Dataframe.

Trend Analysis: The trend of the time series can be estimated using a parametric approach because it produces smooth trend curves representing the overall tendency, and allowing for future trends to be computed for prediction purposes. Popular fitting functions include linear, exponential and quadratic types [16]. The trend is the component of a time series that represents variations of low frequency in a time series, the high and medium frequency fluctuations having been filtered out. The objective of this analysis is to understand if there is a trend in the data and whether this pattern is linear or not. The best tool for this job is visualization.

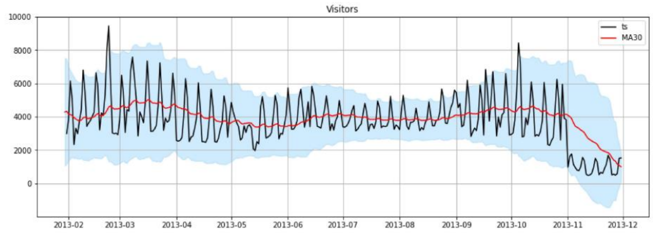


Fig -6: Visitors

Next we want to see within the plot some rolling statistics such as: Moving Average: the unweighted mean of the previous n data (also called "rolling mean") and Bollinger Bands: an upper band at k times an n-period standard deviation above the moving average, and a lower band at k times an N-period standard deviation below the moving average.

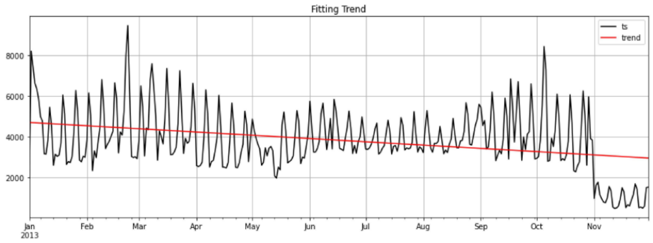


Fig -7: Trend

This is useful in model design as most of the models require that you specify whether the trend component exists and whether it is linear (also said "additive") or non-linear (also said "multiplicative").

Outliers Detection: An outlier [17] is a data value that lies in the tail of the statistical distribution of a set of data values.

The objective of this section is to spot outliers and decide how to handle them. In practice, traditional deterministic methods are often used, like plotting the distribution and label as an outlier every observations higher or lower than a chosen threshold.

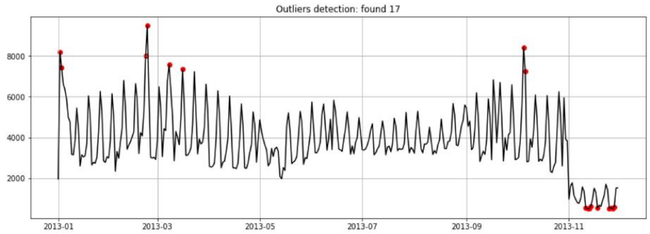


Fig -8: Outliers detection

Firstly, we write a function to automatically detect outliers in a time series using a clustering algorithm from the scikit-learn library: one-class support vector machine, it learns the boundaries of the distribution (called "support") and is therefore able to classify any points that lie outside the boundary as outliers. With this function we will be able to spot outliers.

We then remove them because time series forecasting is easier without data points that differ significantly from other observations, but by removing these points can deeply change the distribution of the data. So to exclude the outliers, the most convenient way to remove them is by interpolation. So we write a function to remove outliers after they are detected by interpolating the values before and after the outlier.

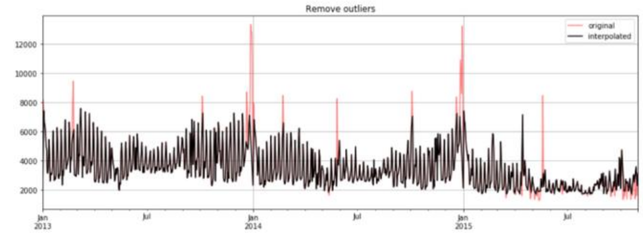


Fig -9: Outliers removed

Seasonality Analysis: The seasonal component is that part of the variations in a time series representing intra-year fluctuations that are more or less stable year after year with respect to timing, direction and magnitude.

The objective of this last is to understand what kind of seasonality is affecting the data (weekly seasonality if it presents fluctuations every 7 days, monthly seasonality if it presents fluctuations every 30 days, and so on).

In particular, when working with seasonal autoregressive models we must specify the number of observations per season. There is a super useful function into the stats-model library that allows us to decompose the time series. This function splits the data into 3 components: trend, seasonality and residuals.

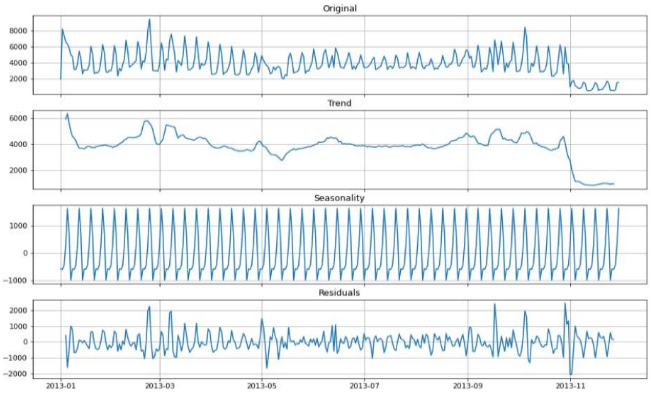


Fig -10: Original, Trend, Seasonality, Residuals

Next we did Time-Series forecasting. Time-Series Forecasting uses past data to predict the future. Meanwhile, causal variable forecasting tries to find the relationship between the desired variable and other external variables.

Time series forecasting is a flexible technique. Its implementation does not require much data, and it is capable of catching fluctuations. The main limitation of using time series forecasting is the lack of empirical justifications.

However, time series forecasting does minimize the need for future estimations and data collection [18].

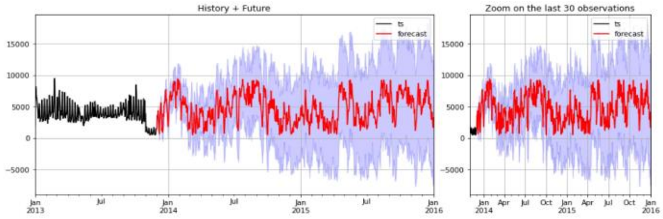


Fig -11: Time-Series Forecasting

5. CONCLUSION

In this paper we have demonstrated our efforts to build a tourist analyzer model using Paris dataset. We propose to cluster the dataset intro clusters and then compute the popular location for each geo-tagged clusters.

We have presented the three techniques for tourist analyzer- geographical data clustering, popular location identification and time series modelling, and have gone through the related works of the authors. We have presented a method to extract, rank, locate and identify meaningful tourist information from unstructured big data sets for supporting the DMO strategic decision-making. The results shows that such a system is helpful for users to find tourism destinations of interests.

In our future work we will concentrate on the evaluation approaches, runtime optimization, database integration and different analytical tasks.

REFERENCES

[1] Finsterwalder, Jörg & Laesser, Christian. (2013). Segmenting outbound tourists based on their activities: Toward experiential consumption spheres in tourism services?. Tourism Review. 68. 21-43. 10.1108/TR-05-2013-0023.

[2] Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research, MIS Quarterly 28 (1), 75-105

[3] March, S. & Smith, G. (1995). Design and Natural Science Research on Information Technology. Decision Support Systems, 15, 251-266

[4] Kisilevich, S., Mansmann, F., & Keim, D. (2010). P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos BT - Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research &#38; Applicat. 1–4.

[5] Zhong, L., Yang, L., Rong, J., & Kong, H. (2020). A Big Data Framework to Identify Tourist Interests Based on Geotagged Travel Photos. *IEEE Access*, *8*, 85294–85308. https://doi.org/10.1109/ACCESS.2020.2990949

[6] Zhang, K., Chen, Y., & Li, C. (2019). Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of Beijing. *Tourism Management*, *75*(November), 595–608. https://doi.org/10.1016/j.tourman.2019.07.002

[7] S. Mikhailov, A. Kashevnik and A. Smirnov, "Tourist Behaviour Analysis Based on Digital Pattern of Life," 2020 7th International Conference on Control, Decision and Information Technologies (CoDIT), 2020, pp. 622-627, doi: 10.1109/CoDIT49905.2020.9263945.

[8] Parikh, V., Keskar, M., Dharia, D., & Gotmare, P. (2018). A Tourist Place Recommendation and Recognition System. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, *Icicct*, 218–222. https://doi.org/10.1109/ICICCT.2018.8473077

[9] Han, S., Ren, F., Du, Q., & Gui, D. (2020). Extracting representative images of tourist attractions from flickr by combining an improved cluster method and multiple deep learning models. *ISPRS International Journal of Geo-Information*, *9*(2), 1–22. https://doi.org/10.3390/ijgi9020081

[10] Arthan, S., Jandum, K., & Tamee, K. (2021). Exploring Tourist Behavior from Social Media Using Geotagged Photographs. 2021 Joint 6th International Conference on Digital Arts, Media and Technology with 4th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, ECTI DAMT and NCON 2021, 285–288. https://doi.org/10.1109/ECTIDAMTNCON51128.2021.9425761

[11] Lu, D. D., & Zhong, Y. De. (2016). A tourist flows analysis system based on phone big data. *Proceedings of 2016 IEEE International Conference on Big Data Analysis, ICBDA 2016*. https://doi.org/10.1109/ICBDA.2016.7509822

[12] Loiseau, T. J., Djebali, S., Raimbault, T., Branchet, B., & Chareyron, G. (2017). Characterization of daily tourism behaviors based on place sequence analysis from photo sharing websites. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 2760–2765. https://doi.org/10.1109/BigData.2017.8258241

[13] Kanjanasupawan, J., Chen, Y. C., Thaipisutikul, T., Shih, T. K., & Srivihok, A. (2019). Prediction of tourist behaviour: Tourist visiting places by adapting convolutional long short-Term deep learning. *Proceedings of 2019 International Conference on System Science and Engineering, ICSSE 2019*, 12–17. https://doi.org/10.1109/ICSSE.2019.8823542

[14] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[15] D. Marutho, S. Hendra Handaka, E. Wijaya and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," 2018 International Seminar on Application for Technology of Information and Communication, 2018, pp. 533-538, doi: 10.1109/ISEMANTIC.2018.8549751.

[16] Cooray, T. M. J. A. (2008). Applied Time Series: Analysis and Forecasting. Alpha Science Intl Ltd. Oxford, UK.

[17] H. C. Mandhare and S. R. Idate, "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 2017, pp. 931-935, doi: 10.1109/ICCONS.2017.8250601.

[18] El-Shafie A, Jaafer O, Seyed A. Adaptive neuro-fuzzy inference system based model for rainfall forecasting in Klang River, Malaysia. Int J Phys Sci 2011;6:2875-2888.