Detect Outliers from A Set of Trajectory During Hajj and Umrah

Hatim Al-Salmi
college of computer and Information Systems, Umm Al-Qura University, Saudi Arabia

The supervision of Dr Louai Alarabi
Department of Computer Science, Umm Al-Qura University, Makkah 24236, Saudi Arabia

Abstract
Trajectories are coordinates that indicate an object's movement during specific time periods. Detecting outliers of these Trajectories indicates Trajectories incompatible with the rest of the Trajectories. In this paper, outliers during the Hajj and Umrah seasons are explored. Outliers from buses using the DBSCAN algorithm, using the Seaborn library to find the lowest and most frequent buses as outliers, and also using Z-score algorithm to find the days when the bus traffic is more than the average.
Keywords
Spatial Outlier Detection, Spatial Cluster Detection, Detection Trajectory data

1 INTRODUCTION

Trajectories are a set of Trajectory created when objects movement during range time. these Trajectories extract from devices used GPS services as locations of objects in interval time.at the moment of moving objects create a big of data that represented locations and time of objects. this data can be analyzed and extract unexpected knowledge called outliers like objects Trajectory Inconsistent with the rest of the data. Detecting anomalies during Hajj and Umrah is important It gives indications such as identifying places of high traffic and irregularities that lead to congestion and unexpected behavior. Finding unfamiliar bus trajectories, as well as finding less functional buses, which leads to delays for pilgrims and Umrah performers. Also the most working days in the seasons of Hajj and Umrah.

Outlier detections is very important to detect values that are not homogeneous with the rest of the data. One of the main challenges in finding outlier detections is the volume of data that exceeds the human natural ability to analyze it, and finding outlier values, especially high-dimensional data, is also difficult to identify all possible normal behaviors. Data quality is one of the main challenges to discovering outliers. The data needs to be cleaned to improve its quality. Also, often one of the difficulties in finding outliers is that they are present in a very small percentage, especially in big data, and it is difficult to differentiate them from noise data. At [2], the authors collected the coordinates of 20,000 buses for 60 days and developed a platform to analyze data in terms of bus traffic, behavior, and speed of drivers. The data is then transferred to the platform for previous analyses. During pre-prossing, many data were removed as incorrect.

The authors in[4], improved the performance of a previous study to determine the most appropriate number of buses within each (zone) and to manipulate the data and to delete the abnormal and missing data by applying techniques to clean the data due to its effect on the quality and accuracy of the data. In previous studies, During pre-prossing, anomalies are ignored and deleted Outliers and were not taken into account as they may be one of the most important causes of the crowd or a link that may lead us to important information such as crowd or service delays, which affects the accuracy of the results.

In this paper, I will detect outliers trajectory from a set of trajectories for buses and analyze their data to create a more streamlined process for pilgrims and Umrah performers also to prevent conges- tion. dataset contains 25 millions records.in data, mining approaches should be clear data from noise and null values. dbscan algorithm is the most algorithm used in detected outliers. Therefore applied it in the dataset.

2 RELATED WORKS

2.1 SPATIAL OUTLIER DETECTION

The paper used bipartite methods to detect spatial outliers. the spatial statistical and spatial point estimation are the concepts. the paper defined three steps to detecting spatial outliers detection.firstly in the spatial dataset describe the neighbors to every object by using methods like fixed distance neighborhood and k nearest neighborhood. next step is used aggregation function to calculating the spatial points of every object based on neighbors' values the estimation is calculated as simple mean by using The local mean method or using the inverse distance method to estimates the values according to the distance between samples last step select witch statistical testing to determine outliers by merging the statistical z-

Manuscript received March 5, 2022
Manuscript revised March 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.3.99

score evaluation and point estimation[11].

the main idea in spatial outlier detection is to find outliers any objects with its neighbors. the limitation of some spatial outliers detection algorithms is considered the impact of objects depend on nonspatial attributes values, therefor proposed spatial outliers detection method calculated weights of neighbors based distance edge between objects in weighted z value algorithm after computing outliers of objects determined the highest outliers factors are outliers objects.in the Averaged Difference Algorithm gave every object and check it individually with neighbors first rather than computing the average than compares it.[5] Spatial Outlier Detection is considered an essential work in Spatial data mining. outliers in globally defined as some data inconsistent with other data in the same set. so can educe unexpected knowledge for example credit card fraud. Spatial Outlier detected by if the attribute values of nonspatial are different from it is neighbors. Spatial Outlier Detection extracts unexpected knowledge Related to spatial database applications or geographic information systems like traffic or services depending on us of locations.

It can detect outliers by use Graphical tests, such as variogram clouds and Moran scatterplots that are visualization of the spatial data and Indicate to outliers.[9]in this paper authors, propose a model look at weights as the criteria to defined the values of Neighboring objects in a spatial data set. there are 3 dimensions in this spatial data: cost, distance, and how many direct edges between neighboring objects. the model considers every neighbor object have differently effected from Neighboring objects It is measured by its weight Assuming the nearest neighbor has more weight .the distance determined by the user by radius every object inside the circle is neighbors to the specified object. all objects outside of the area discard it. the cost of connections can be calculated by defined lower cost after Counts direct and indirect connections between nodes.[10]

2.2 SPATIAL CLUSTER DETECTION

The benefits of using spatial cluster detection are to determine outliers of locations ,some regions Which not consistent with others in the set.also can be used to check if clusters objects are similar or not.steps to scanning spatial in framework: Acquire datasets of spatial locations, then select models in case data clusters used H0 or not clusters used H1.next step based on model selected Acquire score function the regions have the biggest values considered important values like calculating the score function F(S), finally by testing or calculating probabilities and statistics decided the regions It should be studied or ignored if the value not greater than the threshold it considered as spatial outliers.[7]

In this paper defined three steps to detecting spatial outliers. the first step is clustering the second step is checking spatial neighbor's third step is checking temporal neighbors .it can detecting based on spatial, nonspatial, and temporal values together. this algorithm improved the DBSCAN algorithm cluster to determine spatial neighbors and temporal neighbors of every object in the area. also moved the DBSCAN algorithm to de find spatial outliers if clusters have several densities. through allocate density factor to all clusters .the algorithm has four attributes: distance of spatial attribute, a distance of nonspatial attribute, minimum points required to spatial and nonspatial distance. the last attribute is to deny marge clusters while it has similar values of neighbor position. the objects are classified outliers if the numbers of points are lower than the numbers of points required in the area.[6] the paper proposes a new kind of spatial outlier robust spatial z test which supports detected multiple spatial outliers.rather of used spatial z test .for reason the spatial z test properties to determine single spatial outliers. but actually doesn't define outliers in case outliers share attributes values, found in a cluster or correlate together. spatial outliers' interest to find cultures are inconsistent with spatial neighbors and Extract unknown knowledge like traffic overcrowding and congestions. by using robust spatial z detecting different outlines Because of the influence of neighbors. than traditional spatial outliers detected. the firest advantage of robust spatial z is the distance calculated for all observations instead of n-1 observations. sacond advantage it can detect clusters of outliers even if humans try effected.[3]

2.3 SPATIAL OUTLIER DETECTION FOR TRAJECTORY DATA

The problem is defined trajectories that not consist of other trajectories .the concept is to detect anomaly trajectories it handling it by calculating density estimation of every point. the paper proposed aggregated anomaly detection with normalizing flows (GRADINGS) approach .it can from all trajectories initialize models to evaluate outliers specified trajectory. the GRADING method has split the set of trajectory into parts of trajectories, next step has estimated the distribution of all trajectories that have been divided previously, finally aggregate the anomaly scores of all parts Individually as anomaly score for original trajectory.[1]

The paper defined Trajectory spatial and temporal attributes data of object's content locations and movement towards at different intervals time.in the case, this outliers task is detecting objects movement out of boundary using the Trajectory Outlier Detection algorithm TODB.TODB

divide into three parts first pre-processing of trajectories to cleaning data, secondly boundary for the trajectories using the data after cleaning, and assigned it to Convex Hull Algorithm to determine boundary finally trajectories are classified to outliers on not based on points that are identified in the second step. also, the paper proposes a new algorithm classification integrate style movement with boundary determined by the convex hull algorithm.[8]

=

3 METHODOLOGY

3.1 DATASET

It contains 25 million records with a capacity of 2.4 GB. I got it from Dr. Louai Alarabi. There are 7368 buses and 23 zones for 10 consecutive days with 10 attributes (id, bus no, latitude, longitude, ate report, date time bus location report, altitude, speed, c id) 154430 19437466 61477 21.607751 39.198815'2016-09-01 02:51:50' '2016-09-01 02:28:58' " 0.0 0.0 1) is instance of data.. Each record consists of the bus number, the zones number, the time of obtaining the bus coordinates as well as the bus coordinates (longitude and latitude).

3.2 METHODS

3.2.1 FIND OUTLIERS OF POINTS

Using the jupyter notebook for Python, the data (locationHistory.txt) is read by the Panda Library Then data from duplicate rows and invalid and empty values were processed. The data contains 13.2 million records after cleaning and processing.The dataset was divided by zone to facilitate extraction of outliers for each zone separately. Then, the area data were read and the values of latitude and longitude were assigned to the variables X and Y and the data were plotted by matplotlib.pyplot.scatter where X represented latitude and Y represented longitude. The other method is to import the DBSCAN algorithm send dataset as an argument to extract outliers and plot them using matplotlib.pyplot.

3.2.2 FIND BUS OUTLIERS

Read the column containing the bus number with the Panda Library, then convert the values results to a list without duplicating the bus number. Also read the number of records for each bus separately. The list representing the bus number with the frequencies of each bus is plotted by matplotlib.pyplot to find the least frequent buses as anomalous values.

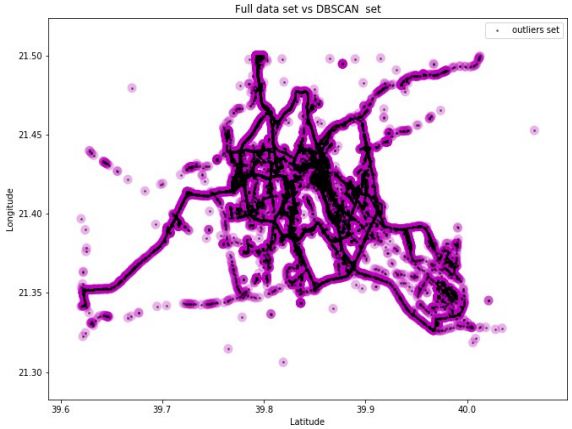
Algorithm 1: Detect outliers approach
SET Primary file = A
SET file number = n
SET fig number = f START
IMORT the libraries
determine the address of the file that we want to invoke
CALL file locationHistory. text
READ file
DROP columns we don't need from tables
separation of data according to cId values
if *col umns I d n* **then**
    save files columnsId n
    save files separately in excel file
end
START
call all file
initialize n to one
initialize f to zero
foreach *file n* $\in A$ **do**
IMORT dbscan
READ dataset
set DBSCAN(*dataset, eps, Min*)
PRINT result
return result
DRAW result = fig (f) SAVE f
PRINT f
END ;

4 EXPERIMENTS

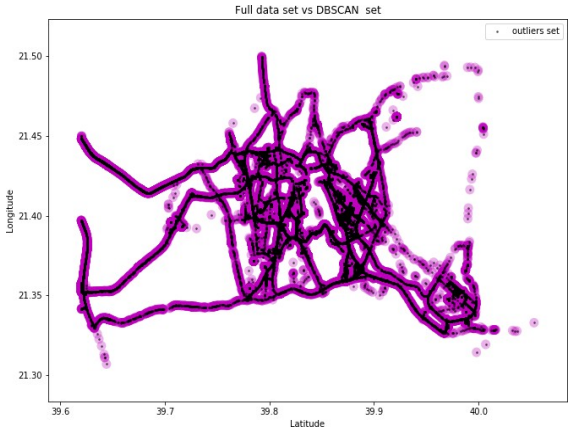
The dataset was tested from three different aspects using Python language First, the discovery of outliers for the different Trajectories Secondly, the discovery of the least frequent and most frequent buses Third, discover the days that are more and less frequent

4.1 TRAJECTORIES

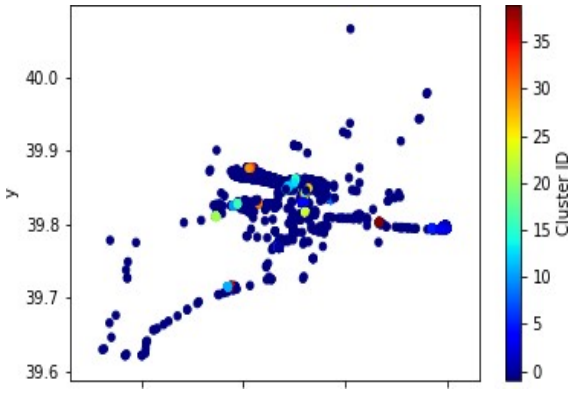
First, the discovery of outliers for the different Trajectories The first step is to read all data by using the Pandas library, and the second step is to classify it according to each zone. Using the matplotlib library, we show in Fig.1 (a) and (b) the coordinates of the buses are incompatible with the remainder as discrete single points of zone 1 and 2 .In Fig.1(c) and (d), the DBSCAN algorithm is used to represent the outliers of buses by different colors.



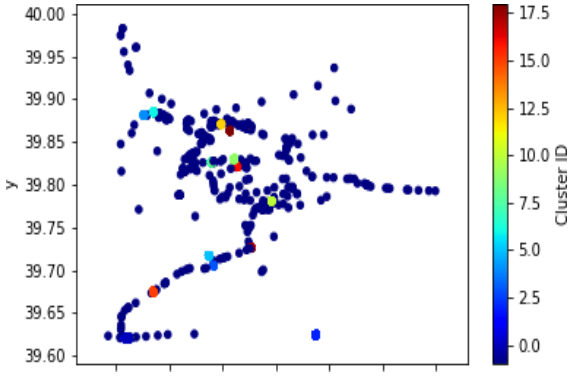
(a) outliers Trajectories of zone1 by using matplotlib



(b) outliers Trajectories of zone2 by using matplotlib



(c) outliers Trajectories of zone1 by using DBSCAN



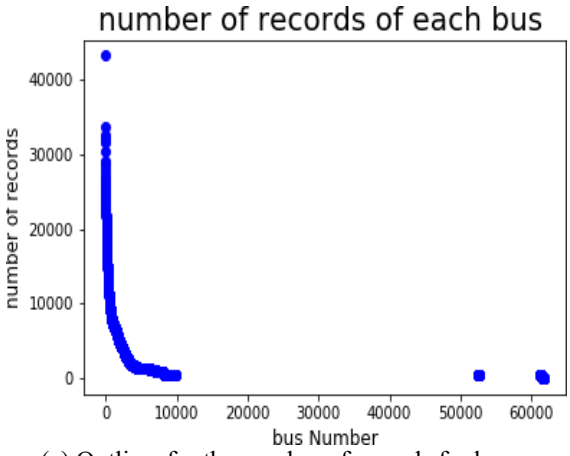
(d) outliers Trajectories of zone2 by using DBSCAN

Fig.1: Outliers Trajectories

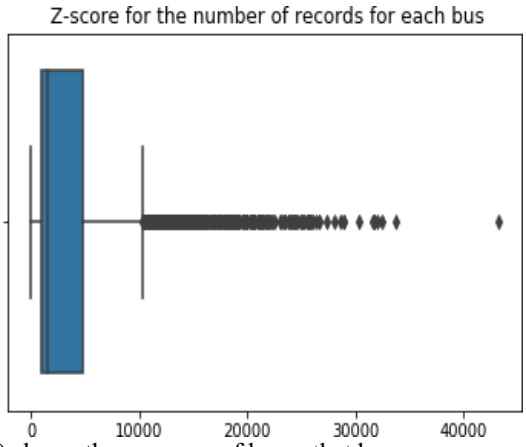
4.2 LEAST FREQUENT AND MOST FREQUENT BUSES

Secondly, the discovery of the least frequent and most frequent buses The first step is to read all the data by using the pandas library, and Secondly, the buses were divided and classified according to the number of each bus, and then the records for each bus were calculated Converting bus numbers to an array and calculating the number of frequencies for each bus

Plot the number of frequencies for all buses using the matplotlib.pyplot library as shown in Fig.2(a) x axis represent the bus id and y axis represent number of records of each bus. And also by using the Seaborn library to find outliers for the number of duplicates of buses. The Fig.2(b) shows the presence of buses that have more records than the normal range consider outliers buses.



(a) Outliers for the number of records for buses



(b) shows the presence of buses that have more records than the normal range consider outliers buses

Fig.2: Outliers buses

4.3 LEAST FREQUENT AND MOST FREQUENT DAYS

Third, discover the least frequent and most frequent days The first step is to read all the data by using the Pandas library, and the second step is to classify it according to the days Plot the number of occurrences for all days using the matplotlib.pyplot library To find outliers for the number of occurrences of days as shown in Fig.3 The second day (2/9/2016) contains a very large number of records compared to the rest of the days, which is considered as outliers.

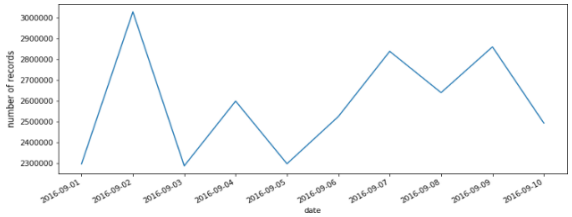


Fig.3: Outliers for the number of records by days

5 CONCLUSION

detect outliers from a set of trajectories during the Hajj and Umrah seasons. concept of spatial outliers and algorithms used to detect spatial outliers from a set of trajectories. The dataset contains 10 attributes and 25.9 million rows of bus locations. I got it from Dr. Louai Alarabi. I analyzed this data to detect outliers trajectories. and detect outliers count records of buses by using DBSCAN and Z-score algorithms.

REFERENCES

[1] Madson LD Dias, César Lincoln C Mattos, Ticiana LC da Silva, José Antônio F de Macedo, and Wellington CP Silva. Anomaly detection in trajectory data with normalizing flows. *arXiv preprint arXiv:2004.05958*, 2020.

[2] E Felemban, FU Rehman, AA Biabani, A Naseer, O Hussain, and EU Warriach. An interactive system for analyzing movement of buses in hajj. *J. Theor. Appl. Inf. Technol*, 98:3468–3481, 2020.

[3] Ali S Hadi and AHM Rahmatullah Imon. Identification of multiple outliers in spatial data. 2018.

[4] Omar Hussain, Emad Felemban, and Faizan Ur Rehman. Optimization of the mashaer shuttle-bus service in hajj: Arafat-muzdalifah case study. *Information*, 12(12):496, 2021.

[5] Yufeng Kou, Chang-Tien Lu, and Dechang Chen. Spatial weighted outlier detection. In *Proceedings of the 2006 SIAM international conference on data mining*, pages 614–618. SIAM, 2006.

[6] Alp Kut and Derya Birant. Spatio-temporal outlier detection in large databases. *Journal of computing and information technology*, 14(4):291–297, 2006.

[7] Daniel B Neill and Andrew W Moore. Anomalous spatial cluster detection. In *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, 2005.

[8] GM Roopa, Arun Kumar GH, Naveen Kumar KR, and CR Nirmala. Optimized data mining tech- niques for outlier detection, removal, and management zone delineation for yield prediction. In *Modern Techniques for Agricultural Disease Management and Crop Yield Prediction*, pages 222–258. IGI Global, 2020.

[9] Shashi Shekhar, Michael R Evans, James M Kang, and Pradeep Mohan. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214, 2011.

[10] Ayman Taha, Hoda M Onsi, Osman M Hegazy, et al. A model for spatial outlier detection based on weighted neighborhood relationship. *arXiv preprint arXiv:1911.01867*, 2019.

[11] Mingzhen Wei, Andrew H Sung, and Martha E Cather. Detecting spatial outliers using bipartite outlier detection methods. In *IKE*, pages 236–244, 2004.