A Clustering Model for Taxi Customer to Avoid Covid-19 Potential Infection Using GPS Trajectory Data

An Do Phu, Quang Pham Xuan, Son Le Hai

***Abstract*—Advances in GPS tracking technology have enabled us to install GPS tracking devices in city taxis to collect a large amount of GPS traces under operational time constraints. These day, Covid-19 is a serious problem, Taxi's customer should avoid crowded place or avoid the place's rush hours to protect there self. These GPS traces provide unparallel opportunities for us to uncover crowded location and respective time. It help to give recommendation to avoid Covid-19 potential impact. In this paper, we develop an "analysis system to avoid Covid-19 potential infection for Taxi Passenger", which is able to systematically investigate taxi locations thence inferred passenger's crowed places and times. In this system, we first provide systems to find two parameters: location and time. To implement the system, we first identify interesting aspects from a large amount of taxi GPS logs. Then, we propose a clustering method to mine the location. Based on locations, we exploit the system to identify the time frame. Finally, we analysis the results and give recommendations to avoid crowed places and times, which helping to avoid Covid-19 potential infection.**

***Index Terms*—Covid-19, Clustering, big data, taxi, location, time, GPS.**

I. Introduction

ON march 11, the World Health Organization (WHO) officially declared COVID-19 a pandemic which was the reason for over one-tenth million cases of coronavirus illness spreading over 110 countries and territories around the world [1]. Compared to SARS or MERS-CoV, despite having a lower fatality cases, SAR-CoV-2, the virus that causes COVID-19 expands its serious effectiveness rapidly which severely impact public health system [2]. According to CDC of the United States, the main method for the transmission of SAR-CoV-2 is person-to-person transmission. People can be easily infected by directly communicating with COVID-19 patients or unwittingly contacting with respiratory droplets produced when an infected person coughs, sneezes or talks [3]. The situation becomes worse and more complicated due to the limitation of our knowledge about SARS-CoV-2 and diversified symptoms of the infected patients [4].

Furthermore, infected patients who are asymptomatic or with mild symptoms, could be highly contagious and these cases also contribute to the major part of the total with the proportion ranging from 18% to about 50% [5]. Consequently, national governments as well as international organizations need to early establish warning system about transmission paths of infected patients or build effective tools to prevent people from entering hot spots in the areas.

Taxis play an important role in offering a comfortable and flexible service within every where public transport system. However, customers and taxi drivers sometimes experience frustration while seeking taxis and passengers respectively. For example, taxis may be waiting at a vacant stand while customers may be queuing in vain elsewhere. This problem has baffled taxi service ever since it existed [6]. These day, taxis have been equipped with GPS receiver and some form of wireless communication device, in order to report its location to taxi monitoring control center [7].

The in-vehicle telematics device builds a location record, including vehicle ID, timestamp, speed, operation status, longitude and latitude. The status field indicates the current operation information of the taxi, specifying whether the taxi is carrying a passenger or empty. In the taxi monitoring control center, location records are not discarded but are being accumulated, since those data have much information on the movement history of each vehicle, taxi service time, dispatch performance, and so on [8]. Accordingly, it provide us with an unprecedented opportunity to automatically extract useful knowledge, which in turn deliver intelligence for real-time decision making in various fields, such as location recommendations, online taxi booking services and taxi business management.

Especially for new normal after Covid-19 pandemic,

Prof. Joonho Kwon with the School of Computer Science and Engineering, Pusan National University, Pusan, Korea.

Manuscript received July 02, 2020; revised July 03, 2020

Algorithm 1: k-means Algorithm

Result: List of clusters centroids
Initialization choose k points that are likely to be in different cluster;
Make these points the centroids of their clusters;
while *new points assignment or threshold not satisfied* **do**
    Find a centroid to which remaining p is closet;
    Add o to the cluster of that centroid;
    Adjust the centroid of that cluster to account for p;
end

the decisions of Taxi drivers and taxi customers are changing. In this paper, we propose to analyze and mining into the sample of T-Drive trajectory dataset contains a one-week trajectories of 10,357 taxis. By using one of the most prominent clustering method called k-means, we then show the local inhabitants about crowded areas which could be highly potential for COVID-19 pandemic spreading. When customers and drivers receive the recommendation, It will help them to make a better decision for there journey.

II. Methodology

A. Approach

In this paper, we focus on utilizing one of the most popular and effective clustering algorithms called k-means. In this algorithm, the Euclidean space and the number of k clusters are both assumed. By using the technical of trial and error, we then finally deduce the best value of k for our data. A k-means algorithm is described in the pseudo-code in the Algorithm.1. The significant point of the algorithm is the for-loop, in which we assign the points to their closet clusters by calculating the Euclidean distance between two points

$$D_{L_2} = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{1}$$

The new centroids are then adjusted, and the loop will stop if and only if there are no new point assignments or the k-means model satisfy a given threshold value. Indeed, in our research, we choose the number of iterations as the stopping criteria [9].

There are various methods for determining the optimal number of clusters (k value), such as the Elbow Method [10] or Silhouette method [11]. While the former is more of a decision rule, the latter is a metric used for validation while clustering. Therefore, the Elbow Method and the Silhouette Method are not alternatives to each other for finding the optimal K. Rather, and they are tools to be used together for a more confident decision. However, due to the time limitation and the low level of our data complexity, we choose to take the most naïve method (data visualization) for picking a suitable k for our research. After plotting and observing different data points' distribution according to different analyzed periods, we choose $k = 10$ to operate our k-means model. For example, Fig.1 indicates 10 clusters of data points after processing the algorithm.

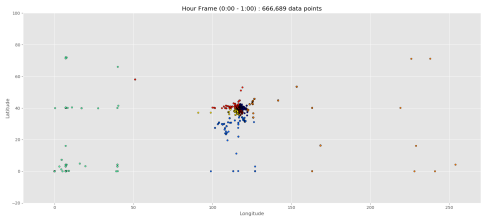


Fig. 1. The clustered GPS data points of taxi drivers in the hour frame from 0:00 to 1:00.

B. Program procedure

The Fig.2 illustrates insight into the working flow of our program. Firstly, we feed the .csv data file, k value, and iteration value into the model. In the next stage, the class of DataFileReader.java will preprocess the raw data, which converts the data into arrays before the processing phase. The system will use the ready-to-train data in our k-means algorithm. The K-means.java class is the heart of our program, which operates and organizes sub-classes functions.

To be more specific, the class mentioned above calls the mandatory sub-class, which is Cluster.java. This class will do the work of assigning data points and re-calculate the centroids based on the Euclidean distance method in Distance.java. Finally, when the number of iterations matches the stopping criteria, the program will print the centroid coordinates on the screen, and save the labeled data points into the format of .csv. We will introduce the detailed descriptions of our system in the following sections.

1) DataFileReader.java:

- Sub-classes: ReadCSVFile, ReadTSVFile.
- Main functions:
  Convert input data file into arrays.
  Record the number of data points and data dimension.

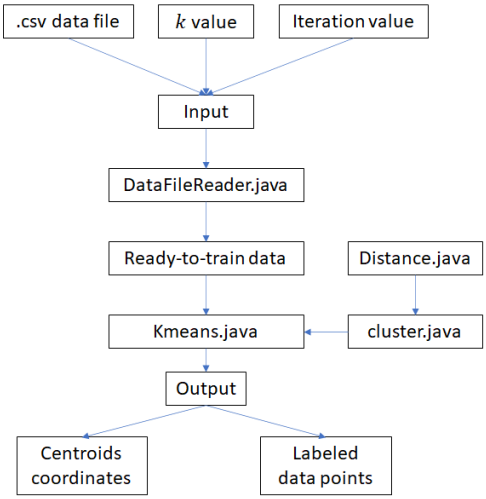


Fig. 2. Program work flow.

2) KMeans.java:

- Sub-classes: kmeans, ClusterPoints, FindNearestCentroid.
- Main functions:

  Performs k-means clustering on a dataset.

  Cluster all data points into their nearest centroids' cluster.

  Calculate the nearest centroid of the candidate point, returns the cluster's index.

3) Cluster.java:

- Sub-classes: Cluster, Centroid, ArrayList, SetCentroid, Insert, SumSquareError, ClearPoints, CalcCentroid.
- Main functions:

  Cluster instantiation.

  Retrieve centroid (average) of the cluster as an array of coordinate values.

  Return the list of points belong to a cluster.

  Set cluster centroid with coordinates of the given centroid.

  Insert given point into the cluster.

  Returns the sum of squared errors.

  Clears points in the cluster before each iteration.

  Calculates and stores the value of the new centroid from all points.

4) Distance.java:

- Sub-classes: Euclidean

  Main Functions:

  Calculates and returns Euclidean distance between two points of equal arbitrary dimensions

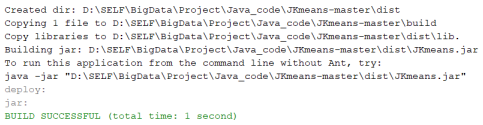


Fig. 3. Program demonstration step 1.

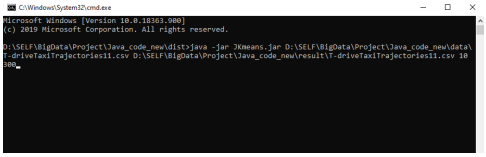


Fig. 4. Program demonstration step 2.

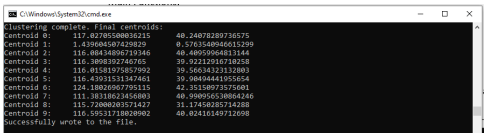


Fig. 5. Program demonstration step 3.

C. Program demonstration

For a brief demonstration of the program, a procedure was introduced as below:

- *Step 1*: Load the project folder, clean and build the project into the .jar file.
- *Step 2*: Open the command window, run the .jar file and give the input of data path, result data path, k value, and iteration value.
- *Step 3*: the program process the k-means algorithm, export the centroids coordinates and labeled data points coordinates.

III. Data

A. Data

Data in this paper was referenced from Jing Yuan et al. [12] via Microsoft publication homepage. This is a sample of T-Drive trajectory dataset that contains a one-week trajectories of 10,357 taxis. The total number of points in this dataset is about 15 million and the total distance of the trajectories reaches 9 million kilometers [13] [14].

Taxi Trajectories: the real trajectory dataset generated by over 33,000 taxis over a period of 3 months. The total distance of the data set is more than 400 million kilometers and the total number of GPS points reaches 790 million. The average sampling interval of the data set is 3.1 minutes per point and the average distance between two consecutive points is about 600 meters. After the

```
1,2008-02-02 15:36:08,116.51172,39.92123
1,2008-02-02 15:46:08,116.51135,39.93883
1,2008-02-02 15:46:08,116.51135,39.93883
1,2008-02-02 15:56:08,116.51627,39.91034
```

Fig. 6. Trajectory data of user ID 1.

preprocessing, they obtain a trajectory archive containing 4.96 million trajectories.

Real-User Trajectories: They use a 2-month driving history of 30 real drivers recorded by GPS trajectories to evaluate travel time estimation. This data is a part of the released GeoLife data-set [15] [16], and the average sampling interval is about 10s. That is, They can easily determine the exact road segments a driver traversed and corresponding travel times.

B. Preprocessing

The raw data includes properties: taxi id, date time, longitude, latitude. The fig.6 below indicates a piece of sample in a file. In this paper, we will clean the raw data and separate the data into 24 data files according to 24 time-ranges from 0:00 to 24:00. There are only 2 properties are remained: longitude and latitude.

IV. Evaluation and Results

A. Evaluation

Evaluating graphs: We build a set of histogram graphs with different values of time-range from 0:00 to 24:00. For intuitive visualization, we group all clustered data point into histograms as shown fig.7 for friendly reading. After project each real-user trajectory to histogram and compare the magnitude of each cluster of each individual time-range, we find that some clusters are more significant than the others. These clusters are the one need to investigate, so, we assume that it represents its time-ranges. Using mean as a reference value of every largest cluster in each time-frame, we easily see the trend of data via the fig.8. The figure show the significantly increasing of taxi service in time-frame 13-17 respect to time-range 13:00-18:00. This specific time-range have also known as rush hours.

Investigate the root cause for the high taxi demand in this time-range, we find the location ,via define on map with longitude and latitude, respect to this time range. We found that this is Xidan sub-district, a major traditional commercial area in Beijing, China. It is located in the Xicheng District. The Xidan commercial district incorporates the Xidan Culture Square, North Xidan Street, as well as many supermarkets and department stores. Understandably for the high taxi demand in this area,

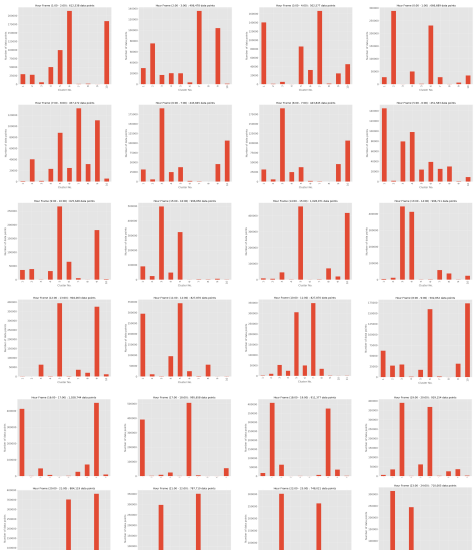


Fig. 7. Histograms of daily hour clustering .

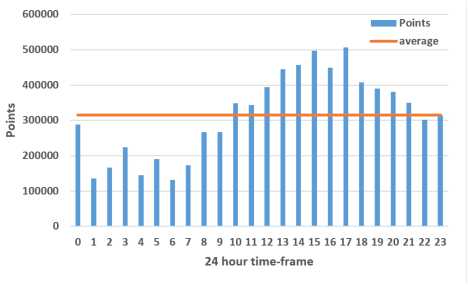


Fig. 8. All largest cluster in every time-frame and its average

where people go for shopping, hanging out and have dinner. This also explain for the significantly decrease of demand after 18:00, when people are finish the activities and go home. We all know that Covid-19 start from a Wuhan's traditional market and rapidly spread out to the worlds. With high density of population, supermarkets, department stores, and restaurants, Xidan sub-district area absolutely is high risk place for Covid-19 infection.

In addition, in different regions, the temporal characteristics of taxi demand are very different. For example, the city hall has many pick-ups, especially in the night time, namely, from 21:00 to 5:00 the next day. During this period, there is no public bus transportation, but city hall areas still gather. Oppositely, the airport area has a high pick-up ratio during the airplane arrival and departure time [7] [17].

A gathering of all clustered GPS data points in the daily hour clustering was shown is fig.9. This is an intuitive way to track the visual change of location

Fig. 9. The clustered GPS data points of taxi drivers in the daily hour clustering.

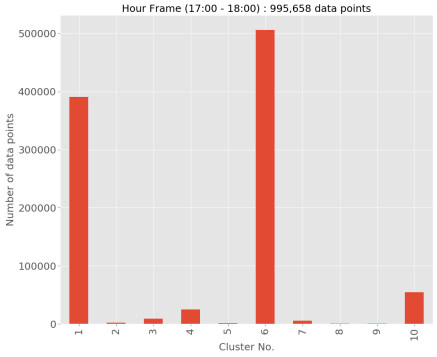


Fig. 10. Histograms for clustered hour frame 17:00-18:00.

distribution and see the distribution of each clustered time-frame.

B. Results

Via latitude and longitude, we can find the exact location, the Sanlihe Road, 100032 Beijing, China , as shown in fig.12. Vary coordinate can integrate with map to show exact location. The Sanlihe Road is part of Xicheng District, Beijing, China. The Xidan commercial district, Beijing Financial Street (Jinrongjie), Beihai Park, Jingshan Park, Shichahai and Zhongnanhai are within its jurisdiction. The popular Houhai bar area is also in Xicheng Precinct. This district is most active area in Beijing, which include 1,259,000 resident and poluation density of 25000 inhabitants per km^2. Others locations of each time-frame from 13 to 17 is also Xicheng District and its surrounding area with trivial distance.

As evaluation in the subsection.IV-A above, the peak value of the trend is time-frame 17 (which from 17:00 to 18:00). The specific histogram of clustered hour frame 17:00-18:00 was shown in the fig.10 to point out the data distribution in each clustering. Respectively, data distribution over longitude and latitude as shown in fig.11. We can visually see the dense clusters and the sparse clusters.

Due to the research, we recommend taxi users and taxi drivers avoid to enter this area from 13:00 to 18:00 or they may come in another time. If there are unavoidable affair in this area, wearing marks, bringing hand sanitizer as precautions are suggested.

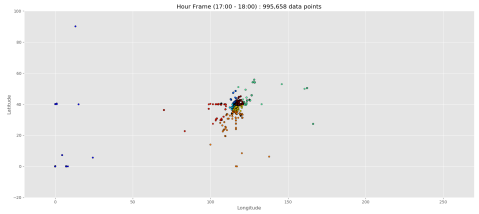


Fig. 11. The clustered GPS data points of taxi drivers in the hour frame from 17:00 to 18:00.

Fig. 12. Define location on map with latitude and longitude of time-frame 13.

V. Conclusion

This paper presents an approach that finds out the practically crowded area at a specific time-range in terms of taxi drivers' intelligence learned from a large number of historical taxi trajectories. We first group data points with clustering method to crowded area. We evaluate the results to find the specific time-range. The results show that our method can give a brief recommendation in the aspects of effectiveness and efficiency. We agree that a recommended crowded area would become less dangerous if many people avoid it. The given recommendation is taxi users and taxi drivers avoid to enter this area from 13:00 to 18:00 or they may come in another time. This is recommendation may change taxi user decision and lead to the changing of taxi driver operation. The system will adapt and give the updated recommendation. In the future, integrate map to find moving route, moving frequency, and combining real-time traffic information with our approach to help taxi driver avoid potential

infection area.

APPENDIX A
FOR JAVA CODE AND DATA SAMPLE

Code and data can be found and downloaded via github: https://github.com/dophuan/t-drive.

ACKNOWLEDGMENT

The authors would like to thank Prof. Joonho Kwon with the School of Computer Science and Engineering, Pusan National University, Pusan, Korea for the Big data course. The authors also thank everyone in the team for the endless efforts to finish this final project.

REFERENCES

[1] J. Ducharme, "The who just declared coronavirus covid-19 a pandemic," *Time*, 2020. [Online]. Available: https://time.com/5791661/who-coronavirus-pandemic-declaration/

[2] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, B. Du, L.-j. Li, G. Zeng, K.-Y. Yuen, R.-c. Chen, C.-l. Tang, T. Wang, P.-y. Chen, J. Xiang, S.-y. Li, J.-l. Wang, Z.-j. Liang, Y.-x. Peng, L. Wei, Y. Liu, Y.-h. Hu, P. Peng, J.-m. Wang, J.-y. Liu, Z. Chen, G. Li, Z.-j. Zheng, S.-q. Qiu, J. Luo, C.-j. Ye, S.-y. Zhu, and N.-s. Zhong, "Clinical characteristics of coronavirus disease 2019 in china," *New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, 2020. [Online]. Available: https://doi.org/10.1056/NEJMoa2002032

[3] C. for Desease Control and Prevention, "How coronavirus spreads," *Time*, 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fprepare%2Ftransmission.html

[4] X. Wu, Y. Cai, X. Huang, X. Yu, L. Zhao, F. Wang, Q. Li, S. Gu, T. Xu, Y. Li *et al.*, "Co-infection with sars-cov-2 and influenza a virus in patient with pneumonia, china," *Emerging infectious diseases*, vol. 26, no. 6, p. 1324, 2020.

[5] J. Qiu, "Covert coronavirus infections could be seeding new outbreaks." *Nature*, 2020.

[6] Q. Meng, S. Mabu, L. Yu, and K. Hirasawa, "A novel taxi dispatch system integrating a multi-customer strategy and genetic network programming," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 14, no. 5, pp. 442–452, 2010.

[7] D. Santani, R. K. Balan, and C. J. Woodard, "Spatio-temporal efficiency in a taxi dispatch system," in *6th international conference on mobile systems, applications, and services, mobisys*, 2008.

[8] J. Lee, I. Shin, and G.-L. Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *2008 Fourth International Conference on Networked Computing and Advanced Information Management*, vol. 1. IEEE, 2008, pp. 199–204.

[9] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.

[10] M. Syakur, B. Khotimah, E. Rochman, and B. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1. IOP Publishing, 2018, p. 012017.

[11] P. Perner, *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings*. Springer, 2017, vol. 10358.

[12] Y. Zheng, "T-drive trajectory data sample," August 2011, t-Drive sample dataset. [Online]. Available: https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/

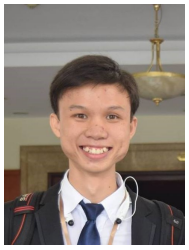
[13] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving directions based on taxi trajectories." New York, NY, USA: Association for Computing Machinery, 2010.

[14] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 316–324. [Online]. Available: https://doi.org/10.1145/2020408.2020462

[15] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 247–256. [Online]. Available: https://doi.org/10.1145/1367497.1367532

[16] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, *Understanding Mobility Based on GPS Data*. New York, NY, USA: Association for Computing Machinery, 2008, p. 312–321. [Online]. Available: https://doi.org/10.1145/1409635.1409677

[17] A. Agarwal, "A comparison of weekend and weekday travel behavior characteristics in urban areas," 2004.



Do Phu An is a Ph.D. candidate in Computer Engineering at the Pusan National University. He completed undergraduate study at University of Information Technology VNU-HCM. He works on the big data analysis, AI, and digital image processing.

Pham Xuan Quang is a Ph.D. candidate in Aerospace Engineering at the Pusan National University. He completed undergraduate study at Ho Chi Minh City University of Technology and his Master of Engineering at Institut Teknologi Bandung. He works on the finite element method, composite behavior, and reduced basis method.

Le Hai Son is a Master student in Aerospace Engineering at the Pusan National University. He completed his undergraduate program at Ho Chi Minh City University of Technology. His research focuses on machine learning, deep learning applications in computer vision.