S & M 2052

Anomaly Detection in Taxi Flow by a Projection Method

Myeong-Hun Jeong,[1] Seung-Bae Jeon,[1] Sangjun Park,[1*] and Sanggu Kang[2]

[1]Department of Civil Engineering, Chosun University, 309 Pilmun-daero, Dong-gu,
Gwangju 61452, Republic of Korea
[2]Spatial Information Research Institute, 163, Iseo-myeon, Wanju-gun,
Jeollabuk-do 55365, Republic of Korea

(Received August 25, 2019; accepted October 16, 2019)

Keywords: trajectory outliers, projection method, trajectory data mining, outlier detection

The prevalence of location-aware devices has propelled studies on the analysis of movement data. In this study, we investigated anomaly detection in taxi trajectories. A projection method was used to detect trajectory outliers. This is a robust statistical method and considers two dimensions of data simultaneously—distance and time. The experimental data included taxi movements in Seoul City and New York City. The results were compared with those of an alternative method, the Mahalanobis distance approach. The findings were observed to be similar. The proposed method can be used to improve the services of taxis and buses.

1. Introduction

The recent development of high-precision location tracking technology [e.g., global navigation satellite systems and radio frequency identification (RFID) tags] has made massive vehicle trajectory data increasingly available. Movement data, such as taxi trajectories, are commonly used to detect hidden patterns in urban flows and for transportation planning. Thus, trajectory data mining attracts researchers seeking knowledge from this rich source.(1) The aim of this study is to detect unusual taxi movements and flows by a projection method. Trajectory outliers that are significantly different from a typical trajectory pattern could then be identified.

In general, trajectory data mining uses one of four techniques—clustering, classification, pattern mining, and outlier detection.(2) In this study, we focused on the detection of trajectory outliers. Previously developed trajectory anomaly detection methods can be classified into two categories.(1) One category includes processes of identifying anomalous events by trajectories. An urban road network is partitioned, and each link is analyzed.(3,4) For example, a summary of traffic condition (e.g., the number of vehicles traveling a link), in a particular time period, is compared with standard patterns, leading to the identification of anomalous events based on trajectories.

The other category includes processes of detecting outlier trajectories. Most contemporary trajectory outlier detection methods can be performed by considering whole trajectories or parts of trajectories.(2) For example, a distance-based algorithm is utilized to detect a trajectory

[*]Corresponding author: e-mail: spark@chosun.ac.kr

https://doi.org/10.18494/SAM.2019.2565

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

outlier as a whole.(5) A partition-and-detect framework(6) and a relative distance(7) are used to detect outlying portions of trajectories. In this study, we focused on outlier trajectories as a whole. While a distance-based algorithm is based on assumed normal distributions in the underlying datasets, the proposed method is robust to a variety of distributions. In this study, we utilized a robust statistic, such as a projection method, which is not sensitive to non-normal data distributions, to detect outlier trajectories. Robust statistics, which performs well over a wide range of probability distributions,(8) has been used in a variety of applications, including trajectory data mining,(9) radioactive materials search,(10) and deformation analysis.(11)

Moreover, a distance-based algorithm typically focuses on one dimension, such as distance or time. Pan *et al.* used the Mahalanobis distance, which considers the correlations between two dimensions of a dataset and is scale-invariant.(4) In a similar vein, in this study, we used a projection method that also takes into account correlations in the datasets and is invariant under rotations of the data.

The remainder of this paper is organized as follows. In Sect. 2, we provide an overview of the experimental datasets and the proposed method. The results are discussed in Sect. 3. In Sect. 4, we conclude with a summary and perspectives on future work.

2. Materials and Methods

In this section, we present the experimental data and system architecture of this study. The proposed method for detecting anomalous taxi flows is then explained.

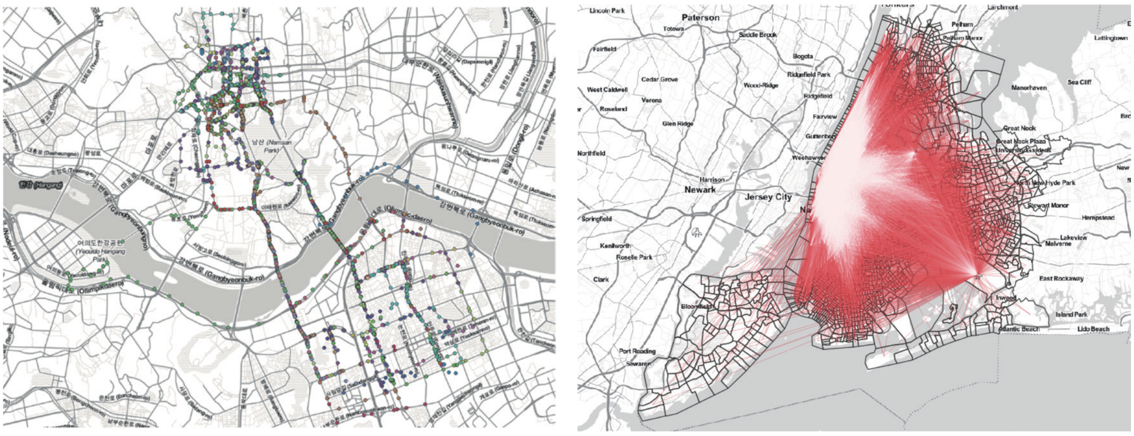
2.1 Materials

We used two data sets to evaluate the effectiveness of the proposed method. One is Seoul City taxi data, and the other is New York City taxi data. The Seoul City taxi movement data were collected from January 25 to 31, 2016. This trajectory data include car registration number, time stamps, driver ID, location information, and passenger information code (e.g., code 0 indicates there are no passengers in a taxi).

The New York City taxi data were collected in 2014. While the Seoul City taxi data included all trajectory information, New York City taxi data included only the IDs of the origin and destination (OD) traffic analysis zones (TAZs). TAZs are the basic geographic unit in transportation planning and are based on the underlying population distribution. TAZs were used in this study as the basic unit of taxi OD trips.

Figure 1(a) shows the Seoul taxi trajectories from the Gangnam district to the Seoul Station area on January 28, 2016, between 7 a.m. and 9 a.m., grouped according to driver. Figure 1(b) shows OD flows in New York City on July 1, 2014. White indicates the most significant OD flows. In particular, in this study, we looked at the OD flow from the Manhattan area (TAZ ID 1177) to LaGuardia Airport (TAZ ID 1597).

The Seoul City taxi data include 40 million trajectory points, and the New York City taxi data are composed of 173 million taxi trip records. In this study, we employed Hadoop with Pig to clean the extensive data and extract the experimental data. *R* was used to implement the



(a) (b)

Fig. 1. (Color online) (a) Seoul City taxi trajectories on Jan. 28, 2016. (b) OD flows of New York City on July 1, 2014.

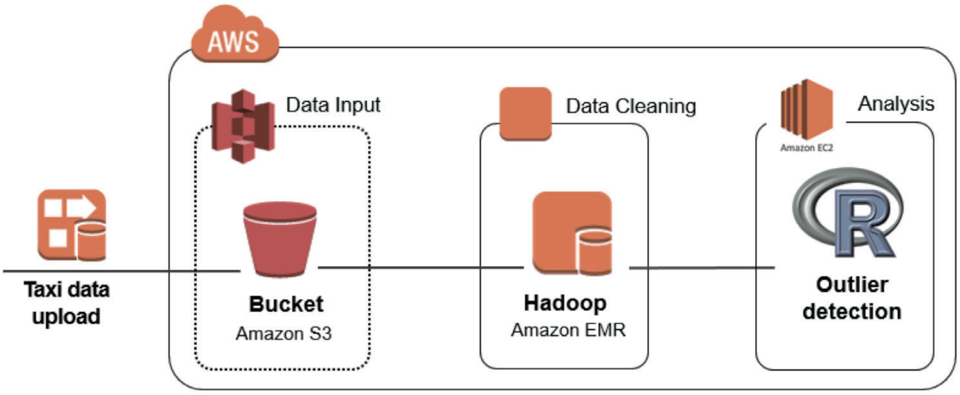


Fig. 2. (Color online) System architecture.

proposed method for detecting anomalies in taxi flows. Figure 2 shows the system architecture. Data were uploaded into the S3 bucket in Amazon Web Service (AWS). The Elastic MapReduce system in AWS was used to resolve the large volume of data.

2.2 Projection method

A projection method is used to detect outliers of taxi flows. Taxi movement data are multivariate, including time, fare, and distance. Outlier detection methods (e.g., rules based on means and variances, and a method based on the interquartile range) for univariate data can miss outliers because they do not consider the overall structure of data. For example, these methods do not generate the same results after the data is rotated. Therefore, in this study, we use a projection method that is invariant under data rotations.[8]

The basic idea of the projection method is that a point among n points is an outlier when the point is an outlier under some projection of the n points.(8) That is, an outlier may be evident when the data is orthogonally projected onto any one of the n lines formed by joining the center of the whole data, as represented by θ, and each X_i. θ can be calculated using the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators.(12,13)

The formal explanation is as follows. For each point X_i, the remaining $n - 1$ points are orthogonally projected onto the line connecting θ and X_i. D_{ij} is the distance between θ and X_j based on this projection. Thus, the jth point may be declared an outlier if it satisfies Eq. (1).

$$D_{ij} > M_D + \sqrt{\chi^2_{.95,2}}\,(q_2 - q_1) \tag{1}$$

Here, M_D is the usual sample median based on D_{i1}, ..., D_{in} values and $\chi^2_{.95,2}$ is the 0.95 quantile of the chi-squared distribution with p degrees of freedom (p is 2 in this study). q_1 and q_2 are the upper and lower boundaries based on Carling's modification of the boxplot rule.(14) Therefore, X_j is declared an outlier if for any i, D_{ij} satisfies Eq. (1).

Figure 3 illustrates how the projection method detects an outlier. In this example, all points are projected onto the line between θ (i.e., a measure of location) and a point, i. Then an outlier is detected on the basis of Eq. (1).

3. Results

The proposed method was evaluated using two case studies: Seoul City taxi data and New York City taxi data. For benchmarking, the performance of the proposed method was compared with that of the Mahalanobis distance approach.(4)

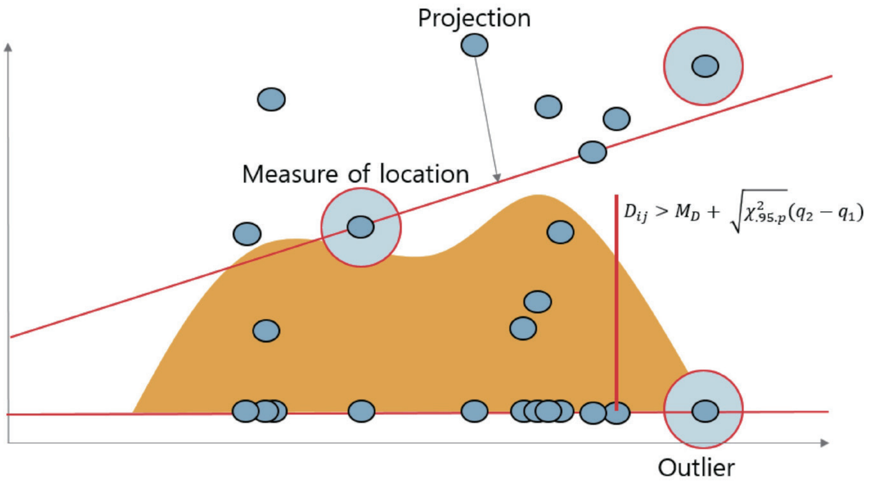
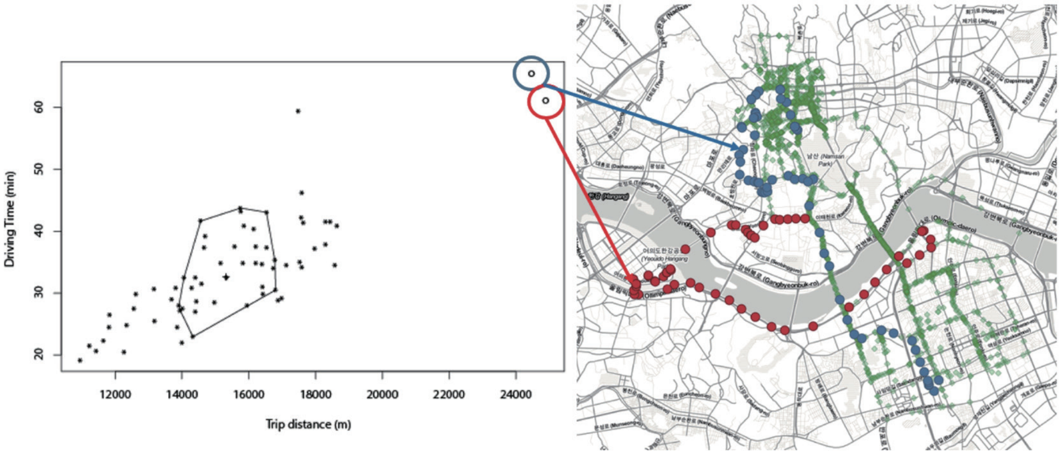


Fig. 3. (Color online) Illustration of a projection method.

3.1 Case study: Seoul City taxi data

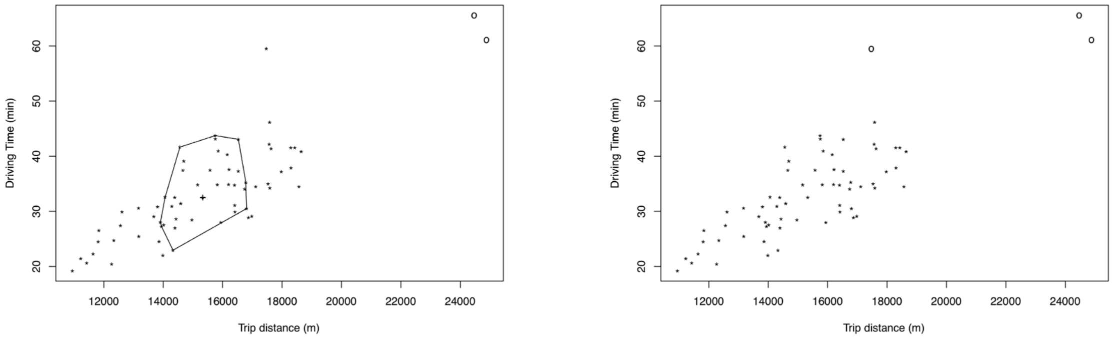
Figure 4 presents the outlying trajectories of the Seoul taxi data from the Gangnam district to the Seoul Station area on January 28, 2016, between 7 a.m. and 9 a.m. Each point in Fig. 4(a) represents the trajectory of a driver. The solid line in Fig. 4(a) indicates an area within which half of the points are located. In this study, we considered not only trip distance but also driving time. The proposed method takes into account the two dimensions simultaneously, as shown in Fig. 4(a). Two trajectories are identified as outliers, which are linked to real trajectories in Fig. 4(b). Typical trajectories are shown in green in Fig. 4(b).

By comparison, an alternative method, the Mahalanobis distance, detected three outlier trajectories, as shown in Fig. 5(b). The Mahalanobis distance method in Fig. 5(b) identified one more trajectory as an outlier than the proposed method in Fig. 5(a). It is significant to select an appropriate outlier detection method. In this case study, the Mahalanobis distance method may have more chances of generating a false-positive error (i.e., Type I error).



(a) (b)

Fig. 4. (Color online) (a) Illustration of a projection method. (b) Taxi trajectories on a map.



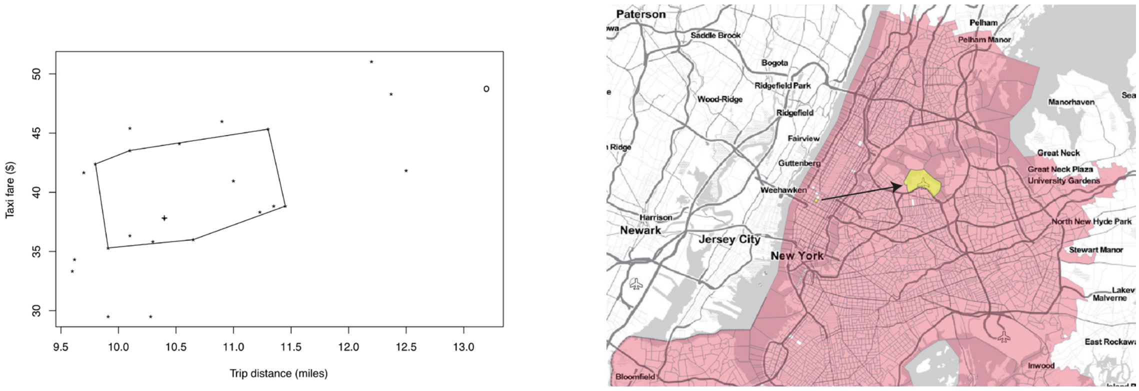
(a) (b)

Fig. 5. (a) Outlier detection using the projection method. (b) Outlier detection using the Mahalanobis distance.

3.2 Case study: New York City taxi data

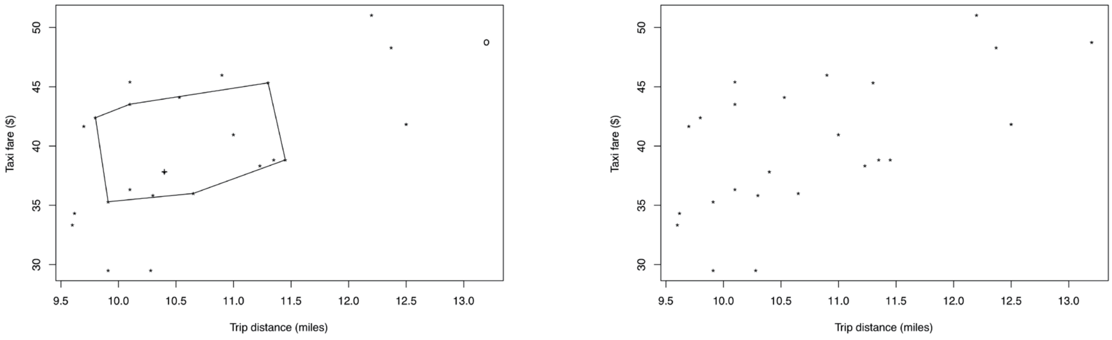
The trajectory outliers of OD flows, derived from the New York City taxi trips, are evident in Fig. 6(a) based on trip distance and taxi fare. New York City taxi trip records only contain OD TAZ IDs, without intermediate locations. In this study, we focused on the OD flow between TAZ IDs 1177 and 1597: the block in midtown Manhattan between 46th and 50th Streets and Seventh and Eighth Avenues and LaGuardia Airport. This OD flow, which is among the highest in the dataset, is shown in Fig. 6(b).

The results of the projection method are compared with those of the Mahalanobis distance method in Fig. 7. While the proposed method detected one trajectory outlier, the Mahalanobis distance method did not detect any trajectory outliers. The reason behind this is that if the distribution of data is ellipsoidal, the Mahalanobis distance method fails to detect outliers.



(a) (b)

Fig. 6. (Color online) (a) Illustration of a projection method. (b) OD flow on a map.



(a) (b)

Fig. 7. (a) Outlier detection by the projection method. (b) Outlier detection using the Mahalanobis distance.

4. Discussion and Conclusions

The overall goal of this study is to identify trajectory outliers as part of understanding urban vehicle flow using taxi movement data. A projection method is used to identify anomalous trajectories.

The projection method detects trajectory outliers effectively. The proposed method considers two dimensions (e.g., distance and time) together. Previous methods (e.g., rules based on means and variances, and a method based on the interquartile range) considered a single dimension, which cannot explain the overall structure of data.

Moveover, the performance of the proposed method was compared with that of the Mahalanobis distance method. Both methods gave similar results. It is nontrivial to choose an outlier detection method.(15) However, it is well known that the Mahalanobis distance method is not robust and unsatisfactory for specific purposes. For instance, the distribution of data in Fig. 7(b) is ellipsoidal. The Mahalanobis distance method suffers as a consequence of detecting outliers based on this distribution.

In this study, we made no attempt to consider subsections of whole trajectories. Further investigations should focus on cases involving the whole and parts of trajectories together. Such research will lead to a better understanding of movement data. Ultimately, movement data analysis provides a window to the world in which we live.

Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1C1B5043892).

References

1 Y. Zheng: ACM Trans. Intell. Syst. Technol. **6** (2015) 29. https://doi.org/10.1145/2743025

2 J. D. Mazimpaka and S. Timpf: J. Spatial Inf. Sci. **2016** (2016) 61. https://doi.org/10.5311

3 W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing: Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min. (2011) 1010. https://doi.org/10.1145/2020408.2020571

4 B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi: Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geog. Inf. Syst. (2013) 344. https://doi.org/10.1145/2525314.2525343

5 E. M. Knorr, R. T. Ng, and V. Tucakov: The Int. J. Very Large Data Bases **8** (2000) 237. https://doi.org/10.1007/s007780050006

6 J.-G. Lee, J. Han, and X. Li: 2008 IEEE 24th Int. Conf. Data Eng. (2008) 140. https://doi.org/10.1109/ICDE.2008.4497422

7 L. Liu, S. Qiao, Y. Zhang, and J. Hu: Int. J. Geog. Inf. Sci. **26** (2012) 1789. https://doi.org/10.1080/13658816.2012.654792

8 R. R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing (Academic Press, San Diego, 2011) https://doi.org/10.1016/C2010-0-67044-1

9 M.-H. Jeong, J. Yin, and S. Wang: 10th Int. Conf. Geog. Inf. Sci. (2018). https://doi.org/10.4230/LIPIcs.GISCIENCE.2018.6

10 M.-H. Jeong, C. J. Sullivan, Y. Gao, and S. Wang: Trans. GIS **23** (2019) 860. https://doi.org/10.1111/tgis.12533

11 K. M. Z. Hassan: Spatial Inf. Res. **24** (2016) 485. https://doi.org/10.1007/s41324-016-0047-5

12 P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection (Wiley, New York, 2005) p. 589. https://doi.org/10.1002/0471725382

13 M. Schyns, G. Haesbroeck, and F. Critchley: Comput. Stat. Data Ana. **54** (2010) 843. https://doi.org/10.1016/j.csda.2009.11.005

14 K. Carling: Comput. Stat. Data Ana. **33** (2000) 249. https://doi.org/10.1016/S0167-9473(99)00057-2

15 R. R. Wilcox: J. Stat. Comput. Simul. **78** (2008) 701. https://doi.org/10.1080/00949650701245041