Outlier Detection and Comparison of Origin-Destination Flows using Data Depth

Myeong-Hun Jeong[1]
Department of Civil Engineering, Chosun University, Gwangju, Republic of Korea
mhjeong@chosun.ac.kr
[orcid]

Junjun Yin[2]
Social Science Research Institute; Institute for CyberScience, Penn State University, PA, USA
jyin@psu.edu
[0000-0002-4196-2439]

Shaowen Wang[3]
Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, IL, USA
shaowen@illinois.edu
[orcid]

Abstract

Advances in location-aware technology have resulted in the generation of a huge volume of trajectory data. Origin-destination (OD) trajectories provide rich information on urban flow and transport demand. This study presents a new methodology to detect OD flow outliers and conduct hypothesis testing between two OD flows in terms of the variations of spatial extent, namely, spread. The proposed method is based on data depth, which measures the centrality and outlyingness of a point with respect to a given dataset in $\mathbb{R}^d$. Based on the center-outward ordering property, the proposed method analyzes the underlying characteristics of OD flows, such as location, outlyingness, and spread. The ability of the proposed method to detect OD anomalies is compared with that of the Mahalanobis distance approach, and an F-test is used to verify the difference in scale. Empirical evaluation has demonstrated that the proposed method effectively identifies OD flows outliers in an interactive way. Furthermore, the proposed method can provide new perspectives by considering the overall structure of data when comparing two different OD flows in terms of scale.

2012 ACM Subject Classification Computing methodologies → Anomaly detection

Keywords and phrases OD Analysis, Trajectory Data Mining, Data Depth, Outliers Detection

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.6

Acknowledgements I want to thank . . .

1 Introduction

With the rapid rise in ubiquity of geolocation-aware sensors, knowledge discovery is greatly enhanced by extracting and mining interesting patterns from spatiotemporal big data in

[1] [funding]
[2] [This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562]
[3] [funding]

various domains. Location-acquisition technologies generate large volumes of movement data, which are used to track people, animals, vehicles, and even natural phenomena. Such data help us better model moving objects and reveal hidden patterns that are important to urban planning [17], urban human mobility [29, 11], the sustainability of urban systems [1, 3], the environment [4], and public security and safety [2].

This paper presents a new algorithm which estimates origination-destination (OD) flow anomalies and conducts hypothesis testing between two sets of different OD flows. In this study, the OD flow data is a subset of trajectory data, which records the origin and destination of each movement while ignoring the actual trajectory route [9]. The algorithm was applied to OD flows derived from extracted origin and destination data in a dataset describing New York City taxi trips, in which each record contained the origin and destination of each trip, without intermediate locations of the actual routes.

In recent years, researchers have investigated a variety of approaches to trajectory data mining. Most contemporary trajectory mining methods can be classified into four categories: clustering, classification, frequent/group pattern mining, and outlier detection [18, 32]. These techniques can be used independently or combinatorially for trajectory mining applications. This study focuses on outlier detection of OD flows. Outlier detection attempts to identify trajectories that do not follow the typical flows of trajectory datasets that characterizes the connectivity between regions [18]. Euclidean distance is employed by [7, 13] to find outlier patterns from trajectories. Studies by [20, 14] question the Euclidean distance approach because of the loss of local features and unavailability when external factors, such as topography, land cover or weather condition, affect the trajectories. In their research, [20, 14] addresses this by using robust distance measurements, i.e., Mahalanobis distance [20] and relative distance [14]. Instead of using distance or density, anomalous trajectories are detected by exploiting comparisons of the structural features of each trajectory segment [30] and an isolation tree of trajectories [31]. Most of the above methods are related to trajectory data analysis, and thus, it is reasonable to extend the application of these approaches to the identification of OD flow anomalies. To overcome the sensitivity of Euclidean distance-based approaches to non-normal distributed data and the difficulty of selecting parameters for anomaly detection techniques based on distance or density, this study employs robust statistics, such as data depth, to detect OD flow outliers.

Flow mapping, a type of visual analytics, is a common approach to analyzing OD flow data. Visual representations of massive movement data facilitate comprehensive exploration of data, in turn enabling perception and understanding complex flow trends. Aggregation and generalization of movement data are frequently utilized to resolve visual clutter [9]. While visual analytics can help to extract inherent patterns from massive data, it is difficult to quantitatively compare two sets of different OD flows based on a hypothesis testing. In other words, it is complicated to comprehend how two OD flows differ and, more importantly, the magnitude of the difference, using a test of statistical significance. Recently published articles employ multidimensional spatial scan statistics [8] and local Ripley's K-function [23] to identify clusters of flow data based on statistical significance tests. Thus, this paper applies bivariate hypothesis testing methods based on data depth to understand the difference between two OD flow datasets in terms of the amount of spatial extent.

It is worth noting that flow mapping approaches frequently suffer from the modifiable areal unit problem (MAUP). Essentially, MAUP is the influence of different aggregations determined by location on the presentation of coherent patterns. Kernel-based flow estimation and smoothing are used to overcome different spatial resolutions [9]. Instead of attempting to find the best areal unit by which to partition urban space and aggregates the OD flows,

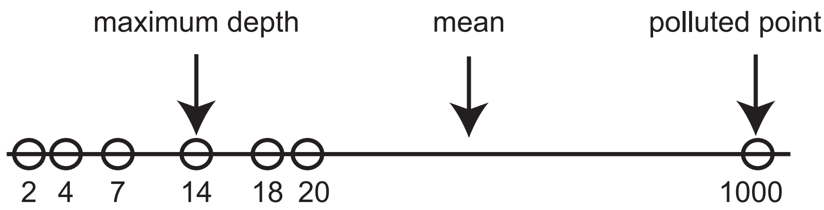


Figure 1 Robustness of halfspace depth for the univariate case

this study adopted the established traffic analysis zones of New York City as base unit. That said, the proposed method can be adapted to other implementation of areal units. In this study, New York City taxi trip data includes the origin and destination within traffic analysis zones, while ignoring the intermediate locations of the actual routes. It is not necessary to reconstruct individual movements for flow estimation (see [5]).

In summary, this paper presents a new algorithm which conducts outlier detection as well as hypothesis testing on OD flow data extracted from a trajectory dataset. Our approach investigates the central regions of OD flows, based on data depth, to detect OD flow anomalies and conduct hypothesis testing between two different OD flow datasets. We believe that our method for analyzing taxi trip data has the potential to aid administrative authorities in understanding crowd patterns, in turn improving urban planning activities such as determining transportation investments.

The remainder of this paper is organized as follows: Section 2 overviews how to detect OD flow outliers and conduct hypothesis testing between two different OD flow datasets using the concept of data depth. Experimental design and the evaluation of the proposed method are presented in Section 3. These results are discussed in Section 4. Section 5 concludes this paper with a summary and future perspectives.

2 Methods

2.1 Data Depth

Data depth measures the centrality of a point with regard to a given dataset in $\mathbb{R}^d$. Originally developed by [24], the notion of data depth (i.e., halfspace depth) generalizes the univariate concept of ranking to multivariate data. Halfspace depth represents how deeply a point is located within a given dataset by ordering all points according to their degree of centrality.

Generally, the halfspace depth (HD) of point x in $\mathbb{R}^d$ is defined as the minimum probability, P on $\mathbb{R}^d$, associated with any closed halfspace containing x [33].

$$HD(x; P) = inf\{P(H) : \text{H is a closed halfspace}, x \in H\},\ x \in \mathbb{R}^d.$$

For the univariate case, all values less than or equal (greater than or equal) to x form a closed halfspace. All values less (greater) than x are an open halfspace. The smallest probability associated with two closed halfspaces developed by x is the halfspace depth of point x. In Figure 1, the probability of values less than or equal to 4 is 2/7 and the probability of values greater than or equal to 4 is 6/7. Thus, the halfspace depth of 4 is 2/7, which is the minimum probability carried by any closed halfspace containing 4. Furthermore, as the sample median, 14 has the largest halfspace depth. Note that, the polluted point inflates the standard error of the sample mean, thereby distorting the view of the data.

Similarly, the halfspace depth of x for the bivariate case is defined by the minimal number of data points in any closed halfspace, which is determined by a hyperplane through x [21]. In

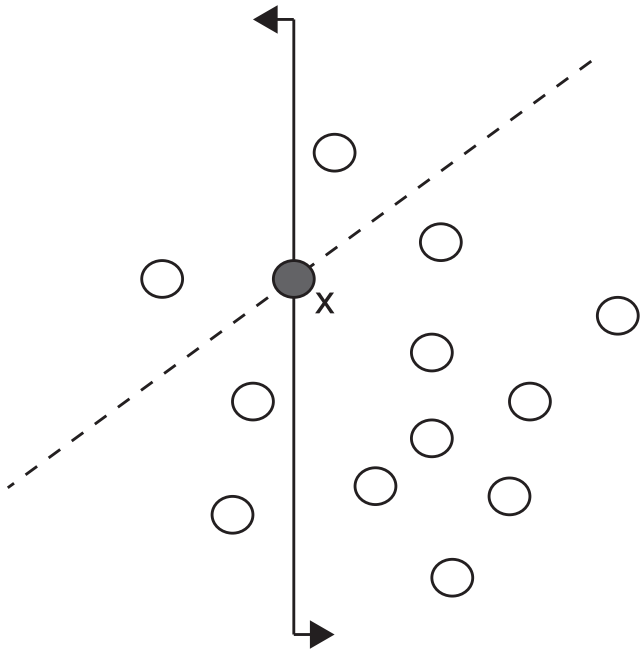


Figure 2 Halfspace depth for the bivariate case

Figure 2, the solid line through x is rotated by 180°. The halfspace depth of x is determined by the smallest portion of data separated by such a hyperplane. For example, the halfspace depth of x is $3/13$, as determined by the dotted line. However, the halfspace depth of x determined by the solid line is $4/13$. Thus, the halfspace depth of x is $3/13$, which is the minimal number of data points in any closed halfspace through x.

The property of halfspace depth is a center-outward ordering of points in $\mathbb{R}^d$ and is affine invariant [19]. These features make halfspace depth useful tool in nonparametric inference, which leads to various applications such as data classification and cluster analysis [12, 10]. There are multiple approaches to calculating data depth, including halfspace depth [21], projection depth [25], and simplicial depth [15]. While the computational complexity of the projection approach is $\mathcal{O}(n^2)$ (where n is the number of points), the computational complexity of simplicial depth is $\mathcal{O}(n^3)$. This can significantly increase execution time when n is large. Thus, this paper uses the more efficient method proposed by [21], in which the computation complexity for both approaches is $\mathcal{O}(n \log n)$.

2.2 OD Flow Outlier Detection Based on Depth

The center-outward ordering in data depth is closely related to the detection of outliers. The upper level sets of data depth in $\mathbb{R}^2$ form the central regions. The most central region can be regarded as a median. Conversely, the lower level sets of data depth, which coincide with larger distances from the center, can be regarded as outlyingness. This concept was utilized by [22, 28] to generate bag plots, which are analogous to one-dimensional box plots based on data depth. This paper uses the bag plot to identify the outliers of OD flows. Before explaining the method of outlier detection, we first introduce a basic definition of OD flow.

▶ **Definition 1.** Origin-destination (OD) flow. The OD flow $OD_i = (o_i, d_i, c_i, ts_i, te_i)$ is the number of trips from the origin ID to destination ID of traffic analysis zones between the start time (ts_i) and the end time (te_i), where $ts_i < te_i$.

Based on this basic definitions, Figure 3 depicts bag plots representing the OD flows of New York City taxi data collected on May 21, 2014 and July 1, 2014. We exploited taxi data on May 21, 2014 because the National September 11 Memorial Museum and Pavilion was opened to the public on this date. We randomly selected another data on July 1, 2014. In Figure 3a, the deepest depth of OD flows, depth median, is represented by a star symbol. This point is surrounded by a dark blue bag, which contains the half of OD flows. This region is regarded as a central region of OD flows. The OD flows in the bag are the dominant patterns. Magnifying the bag by a factor of three, relative to depth median, constructs a fence, as indicated by the light-blue area. The fence is comparable to the whiskers of a one-dimensional boxplot. The OD flows outside the fence, represented by red circles, are outliers. Every OD pair is represented by a point in Figure 3. The x-axis indicates the counts of forward OD flows (e.g., the number of OD flows from origin ID 2 to destination ID 10), and the y-axis indicates the counts of reverse OD flows (e.g., the number of OD flows from origin ID 10 to destination ID 2) in Figure 3a.

The bag plot presents the data using the following attributes:location is represented by the depth median; spread or the spatial extent of bag; correlation or the orientation of the bag; and skewness, as represented by the shape of the bag and the fence [22]. In Figure 3a, we observe that some forward OD flows have higher counts than their paired reverse OD flows. We also note the relatively linear correlation between forward OD flows and reverse OD flows and the skewness of forward (reverse) OD flows.

It is also possible to detect the outliers of OD flows of two different time stamps. In Figure 3c, we visualize the OD flows recorded on two separated days. Comparing the two sets of OD flows not only indicates the central region of OD flows, it also distinguishes the significantly different OD flows.

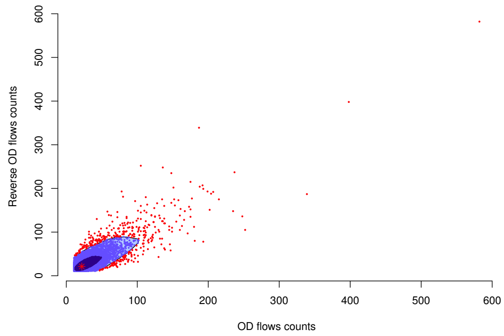
The OD flows in high activity areas of a city are more likely to have large trip volumes. We use set operations to detect such outliers. We regard OD flows on July 1 as the control dataset (*control*); OD flows on May 21 as test dataset (*test*); and the combination of two OD flows as combination dataset (*combination*) in Figure 3. Then we can calculate the intersection of three outliers sets ($control \cap test \cap combination$), which are represented as rectangle symbols in Figure 3d.

In addition, it is interesting to detect the outliers of OD flows which are typical patterns at time t_1 and atypical behavior at time t_2. We define the union of points in the bag, the central region, at time t_1 and t_2. Then we calculate the intersection of two sets, the outliers of the combination set and the previous union set. These outliers are represented as triangle symbols in Figure 3e. These outliers are typical OD flows at time t_1, located in the central regions in the bag plot. When we consider two OD flows together, they become unusual OD flows, some have more trips and some have less trips, relative to the control dataset. Thus, we can detect and treat outliers interactively based on data depth.

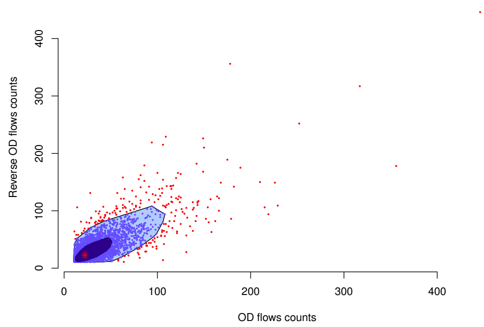
2.3 OD Flow Comparisons Based on Depth

Data depth can compare bivariate data from two independent groups. A t-test can be used to compare means from two independent groups. For example, the t-test reveals whether the means of two OD flows are different at two different temporal ranges. However, it is also worth examining how groups differ in terms of scale, which is also referred to as spread. Comparisons of central regions in data depth evaluate the marginal distribution, thereby considering the overall structure of the data [26].

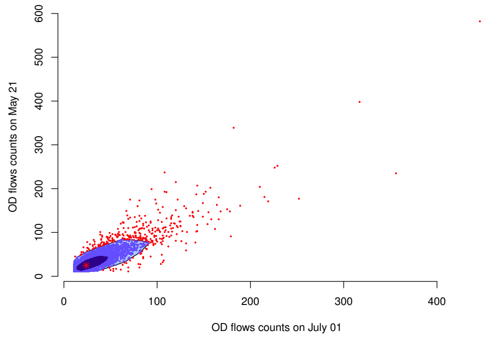
Let X and Y be the random variables having distributions F and G for two independent groups. The quality index proposed by [16] is the probability that the depth of Y is greater



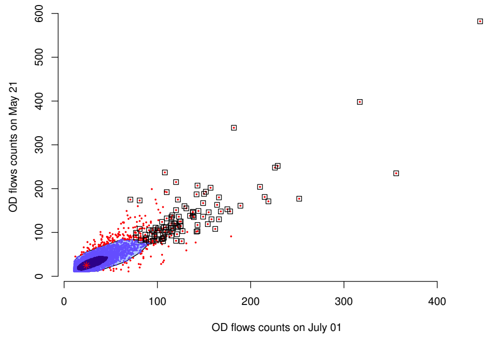
(a) May 21 2014



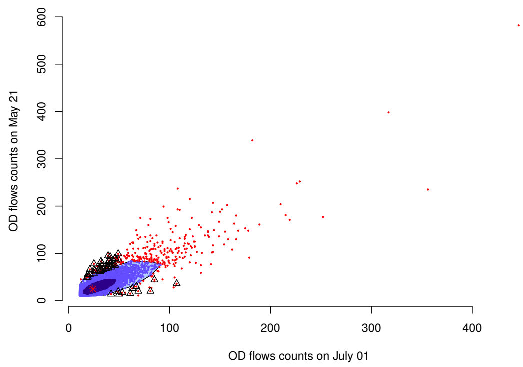
(b) July 1 2014



(c) Combination of May 21 and July 1



(d) Detection of outliers with high volume of trips



(e) Relative outliers detection

Figure 3 Outliers detection of OD flows using a bag plot

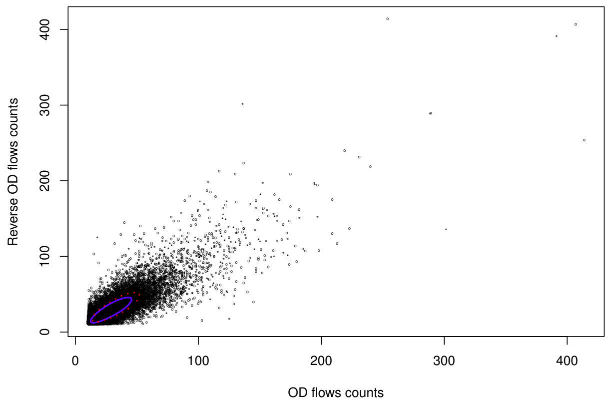


Figure 4 Central regions of two OD flows: ∘ indicates the OD flows for Saturday, March 29 2014 and * indicates the OD flows for a list of Saturdays; blue line presents the central region of the OD flows for the list of Saturdays and red dotted line presents the central region of the OD flows on March 29.

than or equal to depth of X.

$$Q(F, G) = P[D(X; F) \leq D(Y; F)],$$

where P is the probability and $D(X; F)$ is the depth of randomly sampled observations according from distribution F. The range of Q, as presented by [16], is $[0, 1]$ and $Q(F, G) = 0.5$ if and only if $F = G$. If $Q < 0.5$ or if $Q > 0.5$, the scale increases or decreases from F to G. Therefore, it is possible to detect differences in scale using a bootstrap method.

Let $X_1, ..., X_a$ be a random sample from F, and $Y_1, ..., Y_b$ be a random sample from G. The estimate of $Q(F, G)$ is calculated as shown below.

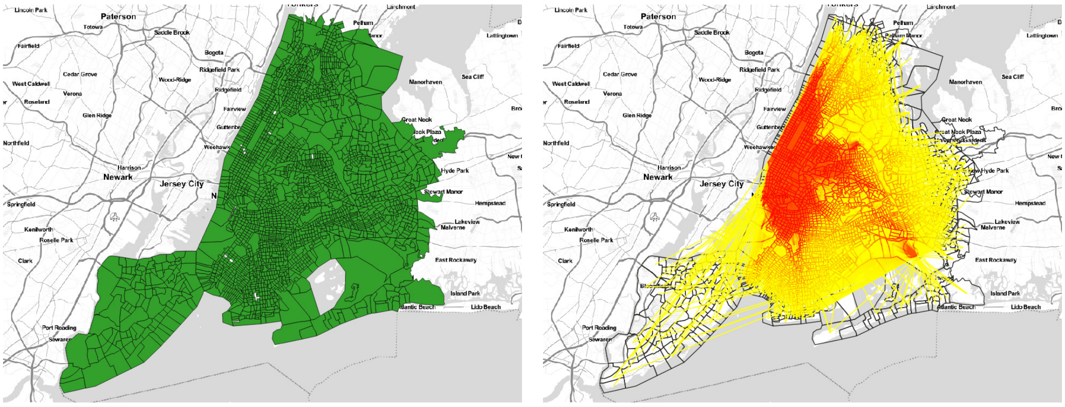
$$\hat{Q}(F, G) = \frac{1}{b} \sum_{i=1}^{b} R(Y_i; F_a),$$

where $R(Y_i; F_a)$ indicates the proportion of X_j which has $D(X_j; F_a) \leq D(Y_i; F_a)$. Similarly, the estimate of $Q(G, F)$ can be defined as follows:

$$\hat{Q}(G, F) = \frac{1}{a} \sum_{i=1}^{a} R(X_i; G_b).$$

Bootstrap samples are obtained by resampling from the two groups (F and G). Under the null hypothesis ($H_0 : Q(F, G) = Q(G, F)$), the difference of the resulting bootstrap estimates is $Q^*(F, G) - Q^*(G, F)$. Thus, if the confidence interval of $Q(F, G) - Q(G, F)$ does not contain zero, we can reject the null hypothesis, H_0 [16, 26].

For ease of understanding, Figure 4 presents the central regions of two OD flows. One dataset is OD flows for Saturday, March 29, 2014, and the other dataset includes multiple Saturdays, those of March 1, 8, 15, 22, and April 5. At 552,064 taxi trips, the day of March 29 had the highest number of taxi trips for the year of 2014. The dataset for the other five



(a) 2,250 traffic analysis zones in New York City **(b)** OD flows on July 1 2014

Figure 5 Experimental data: New York City taxi data

Saturdays comprised 2,621,703 taxi trips. The bootstrap method reveals that the confidence interval is 0.0247 and 0.0596. This confidence interval does not include zero, thus rejecting the H_0 null hypothesis. This indicates that the amount of scale is significantly changed between two OD flow datasets. Furthermore, the OD flows from the group of Saturdays is nested within the OD flows corresponding to March 29. This additional perspective was based on data depth comparisons.

The bootstrap method is a time consuming process. For this study, we generate $2,000$ bootstrap samples. To improve the execution efficiency of the bootstrap computation, we distributed the work across multiple nodes and cores by implementing an embarrassingly parallel R code.

3 Experiments

3.1 Data

This study uses New York City taxi data collected in 2014 to evaluate the effectiveness of the proposed approach. Figure 5a presents traffic analysis zones in New York City which indicate the origin and the destination IDs of the OD flows. A traffic analysis zone (TAZ) is the most commonly adopted basic geographic unit in transportation planning models. The geographic areas of TAZ are delineated by transportation officials for tabulating traffic-related data. The size of TAZ varies because it accounts the underlying population in each zone, which consists of one or more census blocks, block groups, or census tracts. The shapes of the TAZs in this study are derived from the cartographic boundary shapefiles developed by the U.S. Census Bureau in conjunction with the 2010 census (https://www2.census.gov/geo/tiger/TIGER2010/TAZ/2010/). Considering the TAZs are particularly useful for journey-to-work and place-of-work statistics, we employed them as the basic units for accounting the taxi trips. Figure 5b shows OD flows on July 1. Red lines indicate the dominant OD flows.

As a case study, this paper examined OD flows recorded on weekdays and weekends in June 2014. The weekday dataset includes taxi trajectories collected on June 3, 10, 17, and 24, and represents 1,721,655 taxi trips. The weekend dataset includes taxi trajectories collected on June 8, 15, 22, and 29, and describes 1,593,480.

3.2 Procedure

The performance of the proposed method was compared with alternative methods. Trajectory anomaly detection based on Mahalanobis distance [20] was used to evaluate the performance of outlier detection by the proposed method. The Mahalanobis distance is distinguished from Euclidean distance by its consideration of the correlations of the data, in this case, the two OD flow datasets. According to [20], the anomaly detection threshold can be defined as follows:

$$d_M(OD_{t_1}, \mu_{[t_0,t_1)}) \geq 3 \cdot \sqrt{\frac{1}{N} \sum_{t \in [t_0,t_1)} (OD_t - \mu_{[t_0,t_1)})^2}.$$

where OD_{t_1} is the current OD flow, and $\mu_{[t_0,t_1)}$ is the median of all OD flows during $[t_0, t_1)$. In addition, we visualized the results in order to compare them and make the difference easier to understand. The difference of scale was evaluated using standard statistics, such as F-test, to compare the variance of two datasets.

For data cleaning process, this study used Hadoop with Pig. We developed a Hadoop program to handle the large data volume, which was composed of 173 million taxi trip records, remove trips with invalid OD coordinates, and assign each OD locations into the corresponding traffic analysis zone. To implement the OD flow outlier detection, this study used R. The computing environment used Amazon Web Service and the Bridges supercomputer at the Pittsburgh Supercomputing Center. This study only evaluated OD flows more than 10 trips, as the low trip number OD flows could have distorted the view of the data. All the code will be released as open source (the link to the code will added later and currently is available upon request).

3.3 Case study: weekdays vs weekends

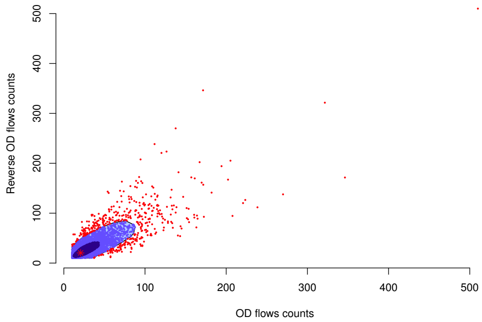
3.3.1 Outlier Detection

The bag plots presented OD flow outliers on weekdays and weekends in Figures 6a and 6b, respectively. The outliers are detected by considering forward OD flows and reverse OD flows together.

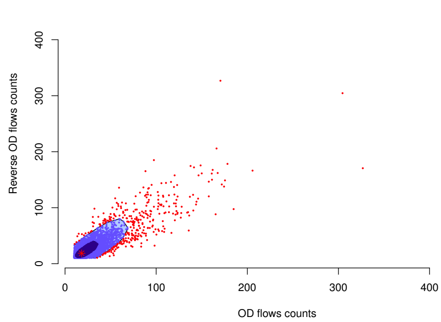
To find the difference between two datasets, we considered two forward OD flows together with the bag plot. Then, we identified the outliers OD flows in Figure 7a. The outliers with rectangle symbols indicate OD flows with large volumes of taxi trips during weekdays and weekends. Figure 7b depicts these outliers are superimposed on a map with red lines. The yellow lines represent the other OD flows, excluding the large volume OD flows on weekdays and weekends.

This case clearly demonstrates that most OD flows occurred in three broad areas: within Manhattan, between the center of Manhattan and the two major airports (J.F.K International Airport and LaGuardia Airport), and between the two airports.

In addition, we investigated abnormal weekend OD flows that are typical weekday OD flows. These abnormal weekend OD flows exhibited substantial variance in number of taxi trips relative to their weekday counterparts. Figure 8a presents these OD flows outliers with triangle symbols. In Figure 8b, red lines indicate the substantial increases in weekend trip volumes. Conversely, blue lines indicate the decreases in trip volume. Figure 8b reveals that OD flows between the center of Manhattan and the two airports or between the two airports were not significantly different during weekdays and weekends. However, we did observe

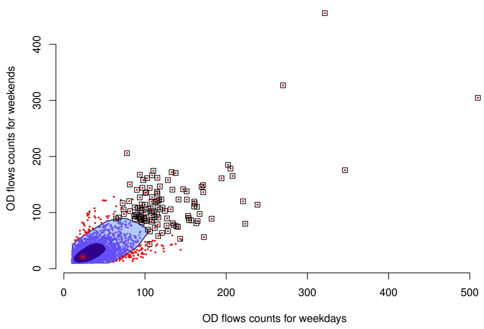


(a) Bag plot on weekdays

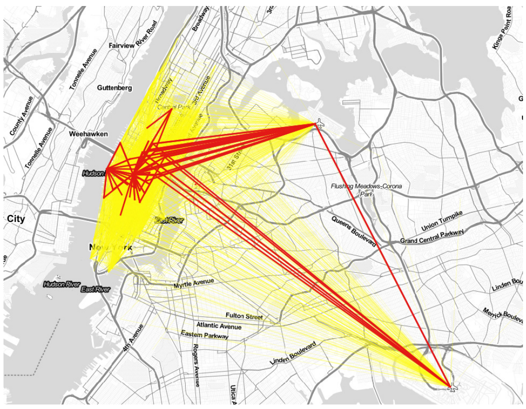


(b) Bag plot on weekends

Figure 6 Outliers detection of OD flows: X-axis indicates forward OD flows counts and Y-axis indicates reverse OD flows counts.

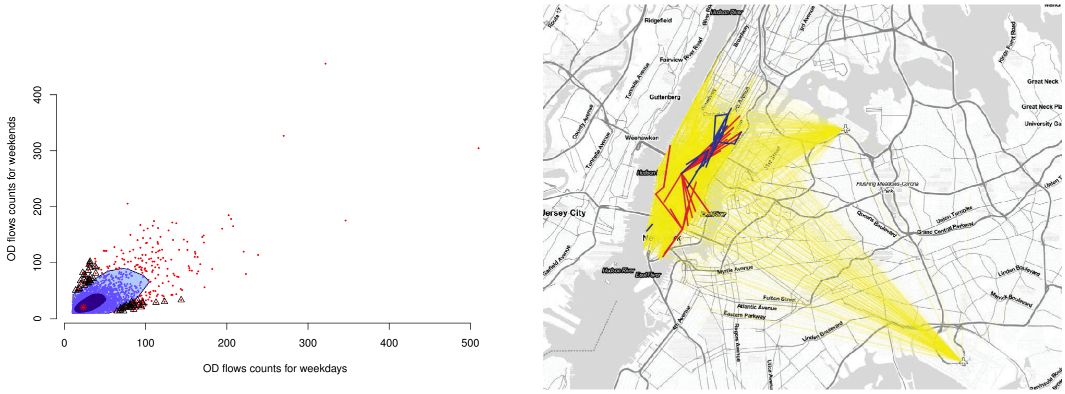


(a) OD flows with high volume of trips



(b) OD flow map with high volume of trips

Figure 7 Outliers with high volume of trips on weekdays and weekends: Rectangles in Figure 7a coincide with red lines in Figure 7b.



(a) Relative OD flows outliers **(b)** OD flow map for relative OD flows outliers

Figure 8 Relative OD flows outliers on weekdays and weekends: Triangles in Figure 8a coincide with red and blue lines in Figure 8b.

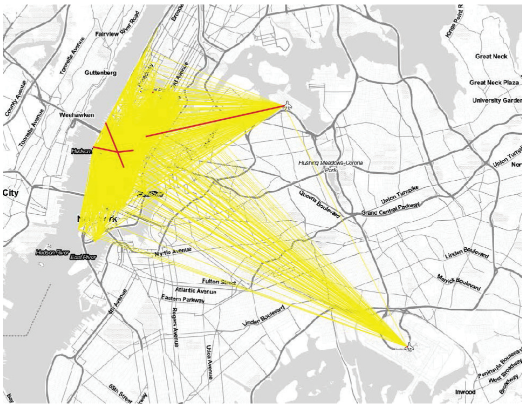


Figure 9 Outlier OD flows on weekdays and weekends based on Mahalanobis distance.

some meaningful decrease in OD flows during the weekends in business district, as depicted by the blue lines in Figure 8b.

We also detected outlier OD flows using Mahalanobis distance. The results are presented in Figure 9. Far fewer outlier OD flows were detected using Mahalanobis distance than by the proposed approach. The Mahalanobis method only considers the forward OD flows of the two datasets. It identified OD flow outliers with high volume of trips because Mahalanobis distance considers the correlations between two OD flows. Thus, Mahalanobis distance is more likely to identify outliers when two OD flows have large trip volumes. In fact, the OD flows outliers from Mahalanobis distance are a subset of the outliers identified by the proposed method, as depicted in Figure 7b. Furthermore, the Mahalanobis distance approach could not detect the outliers detected by the proposed approach in Figure 8 because the Mahalanobis distance approach cannot compare two flows to evaluate significant increases or decreases.

3.3.2 Comparisons in scale

We further investigated how two OD flows differ. Our approach is sensitive to the difference in scale. Hypothesis testing of differences between two central regions in Figure 10 inadvertently

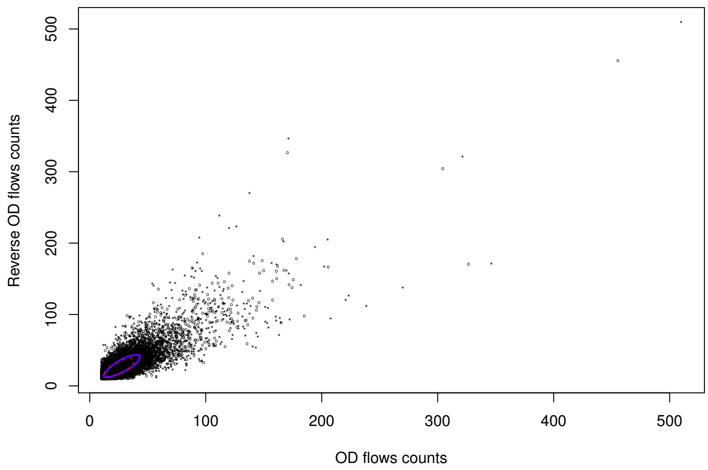


Figure 10 OD flows comparisons based on data depth: ○ indicates the OD flows on weekdays and * indicates the OD flows on weekends; blue line presents the central region of the OD flows for the weekdays and red dotted line presents the central region of the OD flows on weekends.

revealed that the confidence interval was -0.0277 and 0.0157, which includes zero. Thus, failing to reject the null hypothesis. The two central regions were similar in terms of the spread.

Interestingly, the standard statistic F-test was significant, $F(9530, 7637) = 1.1786$, $p \leq 0.05$. The variances of two groups were significantly different. This result directly opposed that of the proposed approach.

4 Discussion

The results demonstrate that the proposed method effectively identifies outlier OD flows based on data depth. It is also possible to detect outlier OD flows by querying with conditional clauses, such as which outlier OD flows always have high trip volumes during time t_1 and time t_2.

As an alternative method, the state-of-the-art Mahalanobis distance approach detected similar outlier OD flows. However, the number of outliers detected was different. This occurred because our OD flows data had heavy tail distributions, that is, many of the OD flows with a long distance from the depth median depicted in Figure 8a. Mahalanobis distance is known to be inadequate when the underlying data have heavy tail distributions [27]. Thus, the presence of outliers may mask the detection of other outliers in Mahalanobis distance approach. Furthermore, it can only detect OD flow outliers with high numbers of trips during time t_1 and time t_2. It is difficult to detect OD flows outliers that have different properties, such as substantial differences in the number of trips when comparing with time t_1 and time t_2.

In terms of the difference in spread, our approach used a bootstrap method to compare the central regions of data depth. This approach investigated the difference in scale as well as the structure of data. It can provide information how deeply points from group 1, OD flows at t_1, tend to be located within group 2, OD flows at t_2. General statistics such as

F-test only provide their difference in variation and do not further specify how groups differ.

Interestingly, the F-test results revealed a statistically significant difference in terms of variation of OD flows on weekdays and weekends. Our approach showed no statistically significant differences. The difference may be caused by the sensitivity of F-test to non-normality [6], which increases the Type-I error rate. Conversely, data depth makes no assumptions about the distributions of the underlying dataset.

5 Conclusions and Future Work

This paper provides a new methodology for identifying outlier OD flows and the difference in scale between two different OD flows at t_1 and t_2. The proposed method is based on the concept of data depth. Data depth is robust statistics, which is suited to non-Gaussian distribution of the underlying datasets. Compared with standard statistics, this approach enhances understanding of the differences and the magnitude of the differences between tow OD flows.

This study made no attempt to consider geographic contexts such as locational circumstances or surrounding environment in understanding OD flows. Further investigation will focus on integrating the analysis of OD flows with geographic context. Such an effort will lead to knowledge discovery and understanding the dynamics of urban flow.

References

1. Marina Alberti, John M Marzluff, Eric Shulenberger, Gordon Bradley, Clare Ryan, and Craig Zumbrunnen. Integrating humans into ecology: Opportunities and challenges for studying urban ecosystems. *AIBS Bulletin*, 53(12):1169–1179, 2003.
2. Maike Buchin, Somayeh Dodge, and Bettina Speckmann. Similarity of trajectories taking into account geographic context. *Journal of Spatial Information Science*, 2014(9):101–124, 2014.
3. Chao Chen, Daqing Zhang, Zhi-Hua Zhou, Nan Li, Tülin Atmaca, and Shijian Li. B-planner: Night bus route planning using large-scale taxi GPS traces. In *2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 225–233. IEEE, 2013.
4. Srinivas Devarakonda, Parveen Sevusu, Hongzhang Liu, Ruilin Liu, Liviu Iftode, and Badri Nath. Real-time air quality monitoring through mobile sensing in metropolitan areas. In *Proc. 2nd ACM SIGKDD International Workshop on Urban Computing*, page 15. ACM, 2013.
5. Matt Duckham, Marc van Kreveld, Ross Purves, Bettina Speckmann, Yaguang Tao, Kevin Verbeek, and Jo Wood. Modeling checkpoint-based movement with the earth mover's distance. In *International Conference on Geographic Information Science*, pages 225–239. Springer, 2016.
6. Andy Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage, London, UK, 2012.
7. Vitor Cunha Fontes, Lucas Andre de Alencar, Chiara Renso, and Vania Bogorny. Discovering trajectory outliers between regions of interest. In *Proc. XIV GeoInfo*, pages 49–60, 2013.
8. Yizhao Gao, Ting Li, Shaowen Wang, Myeong-Hun Jeong, and Kiumars Soltani. A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science*, 0(0):1–22, 2018.
9. Diansheng Guo and Xi Zhu. Origin-destination flow data smoothing and mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2043–2052, 2014.

10 Myeong-Hun Jeong, Yaping Cai, Clair J Sullivan, and Shaowen Wang. Data depth based clustering analysis. In *Proc. 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 29. ACM, 2016.

11 Mei-Po Kwan. Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis*, 30(3):191–216, 1998.

12 Tatjana Lange, Karl Mosler, and Pavlo Mozharovskyi. Fast nonparametric classification based on data depth. *Statistical Papers*, 55(1):49–69, 2014.

13 Jae-Gil Lee, Jiawei Han, and Xiaolei Li. Trajectory outlier detection: A partition-and-detect framework. In *IEEE 24th International Conference on Data Engineering*, pages 140–149. IEEE, 2008.

14 Liangxu Liu, Shaojie Qiao, Yongping Zhang, and JinSong Hu. An efficient outlying trajectories mining approach based on relative distance. *International Journal of Geographical Information Science*, 26(10):1789–1810, 2012.

15 Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, pages 405–414, 1990.

16 Regina Y Liu and Kesar Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993.

17 Jean Damascène Mazimpaka and Sabine Timpf. Exploring the potential of combining taxi GPS and flickr data for discovering functional regions. In *AGILE 2015*, pages 3–18. Springer, 2015.

18 Jean Damascène Mazimpaka and Sabine Timpf. Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, 2016(13):61–99, 2016.

19 Karl Mosler. *Robustness and Complex Data Structures*, chapter Depth Statistics, pages 17–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

20 Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proc. 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 344–353. ACM, 2013.

21 Peter J Rousseeuw and Ida Ruts. Algorithm AS 307: Bivariate location depth. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):516–526, 1996.

22 Peter J Rousseeuw, Ida Ruts, and John W Tukey. The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

23 Ran Tao and Jean-Claude Thill. Spatial cluster detection in spatial flow data. *Geographical Analysis*, 48(4):355–372, 2016.

24 John W Tukey. Mathematics and the picturing of data. In *Proc. International Congress of Mathematicians*, volume 2, pages 523–531, 1975.

25 Rand R Wilcox. Approximating tukey's depth. *Communications in Statistics-Simulation and Computation*, 32(4):977–985, 2003.

26 Rand R Wilcox. Two-sample, bivariate hypothesis testing methods based on tukey's depth. *Multivariate Behavioral Research*, 38(2):225–246, 2003.

27 Rand R Wilcox. *Introduction to robust estimation and hypothesis testing.* Academic Press, 2012.

28 Hans Peter Wolf and Uni Bielefeld. aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3r, plotsummary, plothulls, and some slider functions, 2014. R package version 1.3.0. URL: `https://CRAN.R-project.org/package=aplpack`.

29 Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. Depicting urban boundaries from a mobility network of spatial interactions: A case study of great britain with geo-located twitter data. *International Journal of Geographical Information Science*, 31(7):1293–1313, 2017.

30 Guan Yuan, Shixiong Xia, Lei Zhang, Yong Zhou, and Cheng Ji. Trajectory outlier detection algorithm based on structural features. *Journal of Computational Information Systems*, 7(11):4137–4144, 2011.

31 Daqing Zhang, Nan Li, Zhi-Hua Zhou, Chao Chen, Lin Sun, and Shijian Li. iBAT: Detecting anomalous taxi trajectories from GPS traces. In *Proc. 13th International Conference on Ubiquitous Computing*, pages 99–108. ACM, 2011.

32 Yu Zheng. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3):29, 2015.

33 Yijun Zuo and Robert Serfling. General notions of statistical depth functions. *The Annals of Statistics*, 28:461–482, 2000.