Outliers Detection and Comparison of Origin-Destination Flows with Data Depth

Myeong-Hun Jeong[1], Junjun Yin[2], and Shaowen Wang[3]

1 **Department of Civil Engineering, Chosun University, Gwangju, Republic of Korea**
mhjeong@chosun.ac.kr
2 **Social Science Research Institute; Institute for CyberScience, Penn State University, PA, USA**
jyin@psu.edu
3 **Departmet of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, IL, USA**
shaowen@illinois.edu

Abstract

The advances in location-aware technologies have generated a huge volume of trajectory data. In particular, Origin-Destination (OD) trajectories provide rich information to understand urban flow and transport demand. This study presents a new methodology to detect OD flows outliers and conduct hypothesis testing between two OD flows in terms of ~~spread~~the variations of spatial extent (i.e., spread). The proposed method is based on data depth, which measures the centrality and outlyingness of a point with respect to a given data set in $\mathbb{R}^d$. Based on the center-outward ordering property, it is possible to analyze the underlying OD flows characteristics such as location, outlyingness, and spread. The proposed method is compared with Mahalanobis distance approach to detect OD anomalies, and F-test to identify the difference in scale. Empirical evaluation has demonstrated that the proposed method can identify OD flows outliers in an interactive way. Further, it can provide new perspectives by considering the overall structure of data when comparing two different OD flows in scale.

1998 ACM Subject Classification Dummy classification – please refer to `http://www.acm.org/about/class/ccs98-html`

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

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

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 various domains. In particular, the 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 and its applications in characterizing urban human mobility ~~[?]~~[?], accessibility of the activity space [?], and sustainability of the urban systems [?]. Importantly, trajectory mining leads to solutions for addressing important research problems in different fields, such as urban planning [?], transportation [?], environment [?], and public security and safety [?].

This paper presents a new algorithm which not only estimates origination-destination (OD)'s flows anomaly but also conducts hypothesis testing between two sets of different

OD flows. The algorithm is applied on the New York City taxi data set, where each record contains the origin and destination of each trip (without intermediate locations of the actual routes). We believe that our methods for analyzing taxi trip data have a promising application in understanding crowd patterns, and planning urban and transportation investments by the administrative authorities.

In recent years, researchers have investigated a variety of approaches to trajectory data mining. In general, the contemporary trajectory mining methods can be classified into four categories: clustering, classification, frequent/group pattern mining, and outlier detection [**?**, **?**]. These techniques can be used independently or combinatorially for trajectory mining applications. For this study, we focus on outlier detection of OD flows. Outlier detection aims to find trajectories that do not follow the typical flow of trajectory data sets for characterizing the connectivity between regions [**?**]. [**?**, **?**] use Euclidean distance to find outlier patterns from trajectories. [**?**, **?**] raise the questions of Euclidean distance approach due to the loss of local features and unavailability when external factors affect the trajectories (e.g., topography, land cover or weather condition). [**?**, **?**] addresses this issue by using robust distance measurements (i.e., Mahalanobis distance [**?**] and relative distance [**?**]). Further, structural features [**?**] and a data-induced random tree [**?**] are exploited to detect anomalous trajectories instead of using distance or density. In a similar fashion, this study considers the center regions of OD flows with data depth in order to robustly detect the ~~trajectory~~ OD flow outliers.

In addition, utilizing visual analytics is a common approach to analyze OD flow data. Visual representation of massive movement data enables comprehensive exploration of data and results in understanding complex flow trends. Aggregation and generalization of movement data are frequently utilized [**?**, **?**, **?**]. 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, by how much. In this light, this paper uses bivariate hypothesis testing methods based on data depth to understand how different two OD flow data sets are 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). For instance, it is not guaranteed that different aggregations via location can present coherent patterns. Kernel-based flow estimation and smoothing are used to overcome different spatial resolution [**?**]. Instead of focusing on finding the best areal unit to partitioning the urban space and aggregating the OD flows, this study chooses traffic analysis zones in New York City as a base unit. However, the method can be adapted to other choices of areal units. In ~~addition~~this study, New York City taxi trip data include the origin and destination within traffic analysis zones, which ignores the intermediate locations of the actual routes. It is not necessary to reconstruct individual movements for flow estimation (see [**?**]).

In summary, this paper presents a new algorithm which conducts outlier detection as well as a hypothesis testing from OD flows data. Our approach investigates the central regions of OD flows, based on data depth, to detect OD flows anomaly and conduct a hypothesis testing between two different OD flow data sets. The remainder of this paper is organized as follows: Section **??** overviews how to detect OD flows outliers and conduct a hypothesis testing between two different OD flows with the concept of data depth. Experimental design and the evaluation of the methods are presented in Section **??**. These results are discussed in Section **??**. Section 5 concludes with a summary and future perspectives.

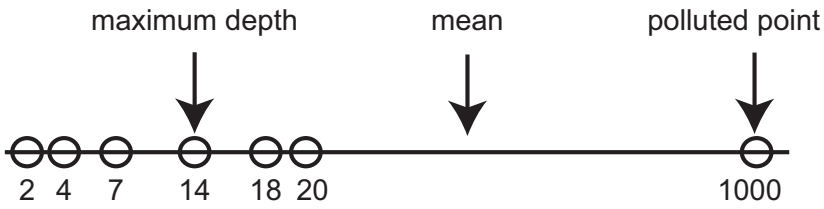


Figure 1 Robustness of halfspace depth for the univariate case

2 Methods

2.1 Data Depth

Data depth measures the centrality of a point with regard to a given data sets in $\mathbb{R}^d$. The notion of data depth (i.e., Tukey's halfspace depth) is originally developed by [**?**], which generalizes the univariate concept of ranking to multivariate data. It presents how deeply a point is located within a given data set by ordering their degree of centrality.

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

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

For the univariate case, all values less than or equal (greater than or equal) to x form a closed halfspace. In Figure **??**, the probability of values less than or equal to 4 is 2/7 and the probability greater than or equal to 4 is 6/7. The halfspace depth of 4 is 2/7, which is the minimum probability carried by any closed halfspace containing 4. Further, 14 has the largest halfspace depth, which is the usual sample median. However, the polluted point inflates the standard error of the sample mean, thereby giving a distorted view of the data.

In a similar fashion, the halfspace depth of x for the bivariate case is defined as the minimal number of data points in any closed halfspace, which is determined by a hyperplane through x [**?**]. For example, the line through x is rotated by 180° in Figure **??**. The halfspace depth of x is determined by the smallest portion of data that are separated by such a hyperplane (e.g., the halfspace depth of x is 3/13, determined by the dotted line).

The property of halfspace depth is a center-outward ordering of points in $\mathbb{R}^d$ and affine invariant [**?**]. These features serve as a useful tool in nonparametric inference, which lead to various applications such as data classification and clustering analysis [**?**, **?**]. There are a couple of approaches to calculate data depth: halfspace depth [**?**], projection depth [**?**], and simplicial depth [**?**]. 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 halfspace depth function proposed by [**?**] of which computation complexity is $\mathcal{O}(n \log n)$.

2.2 OD Trajectory 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, coincide with large distance from the center, can be regarded as outlyingness. [**?**, **?**] utilizes this concept to generate a bag plot, which is analogous to the one-dimensional box plot based on data depth.

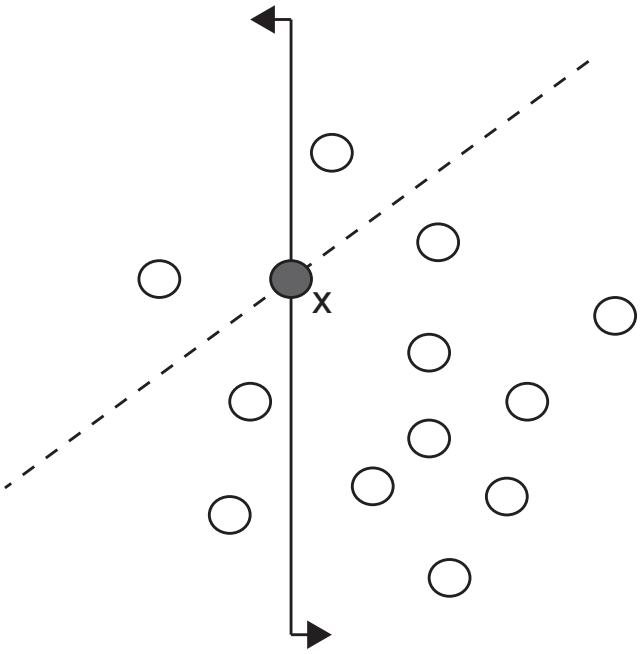


Figure 2 Halfspace depth for the bivariate case

This paper uses the bag plot in order to identify the outliers of OD flows. Before explaining the method of outliers detection, we first introduce a couple of definitions.

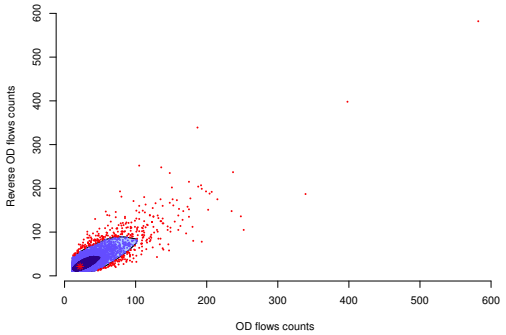
▶ **Definition 1.** Point. A point p is a tuple (o, d, t), where o and d are the origin and destination ID of traffic analysis zones and t is the time when the destination ID is recorded.

▶ **Definition 2.** Trajectory. A trajectory TR is a list of points $(p_1, p_2, p_3, ..., p_n)$, where $p_i = (o_i, d_i, t_i)$ and $t_1 < t_2 < t_3 < ... < t_n$.

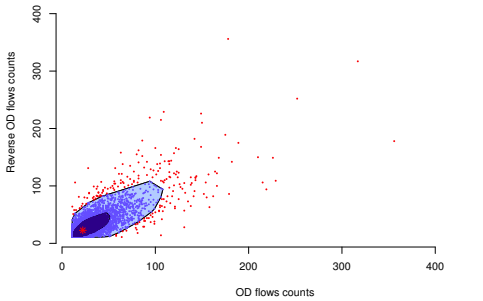
▶ **Definition 3.** OD flow. A OD flow OD is a sum of subset of trajectory TR. $OD = (OD_1, OD_2, OD_3, ..., OD_k)$, where $OD_i = (o_i, d_i, c_i, ts_i, te_i)$, where c_i is the count of the same origin ID (o_i) and destination ID (d_i) between the start time (ts_i) and the end time (te_i), $ts_i < te_i$.

Based on these basic definitions, we plot the OD flows of New York City Taxi data (May 21 and July 1 2014) using a bag plot in Figure **??**. In Figure **??**, the deepest depth of OD flows (i.e., depth median) is represented as a star symbol. This point is surrounded by a bag, which contains the half of OD flows (dark blue area). Magnifying the bag by a factor of 3 relative to depth median constructs a fence (light blue area). The fence is comparable with whiskers in an one-dimensional boxplot. The OD flows outside the fence are outliers (red color circles). The x-axis indicates the counts of forward OD flows and the y-axis indicates the counts of reverse OD flows in Figure **??**. We can similarly present the bag plot of July 1 2014 in Figure **??**.

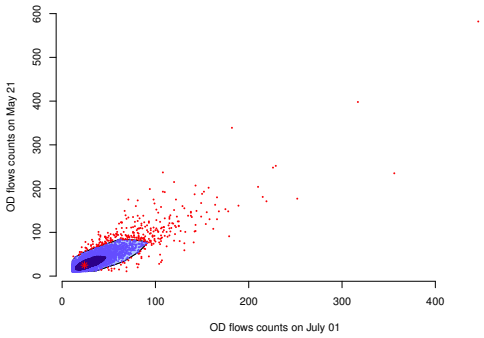
The bag plot is able to present the location (the depth median), spread (the ~~size~~ spatial extent of bag), correlation (the orientation of the bag), and skewness (the shape of the bag and the fence) of data [**?**]. For example, we can see which forward OD flows have higher counts compared to the reverse OD flows; the relatively linear correlation between forward OD flows and reverse OD flows; and the skewness of forward (reverse) OD flows in Figure **??**.



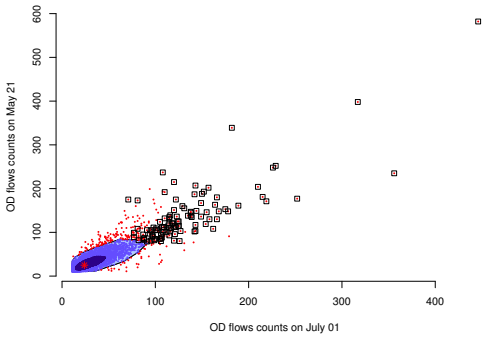
(a) May 21 2014



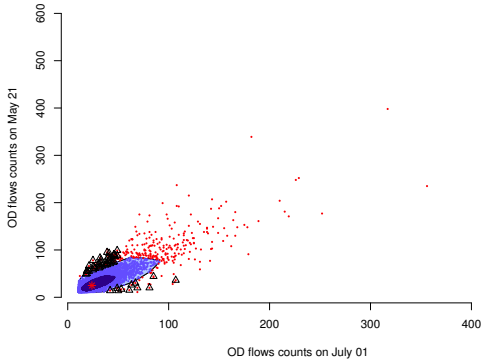
(b) July 1 2014



(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

In addition, it is possible to detect the outliers of OD flows of two different time stamps. In Figure **??**, we visualize the OD flows of two different days. Therefore, we can not only see where is the central region of OD flows, but also differentiate which OD flows are significantly different compared with two different OD flows.

Further, the OD flows of high active areas in a city are more likely to have large volume of trips. We use set operations to detect such outliers. For example, we regard OD flows on July 1 as control data set (*control*); OD flows on May 21 as test data set (*test*); and the combination of two OD flows as combination data set (*combination*) in Figure **??**. Then we can calculate the intersection of three outliers sets ($control \cap test \cap combination$), which are represented as rectangle symbols in Figure **??**.

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. For example, we define the union of points in the bag at time t_1 and t_2. Then we can calculate the intersection of two sets: the outliers of combination set and the previous union set. These outliers are represented as triangle symbols in Figure **??**. For example, these outliers are typical OD flows at time t_1 which are located in the central regions in the bag plot. When we consider two OD flows together, they become unusual OD flows (e.g., some have more trips and some have less trips compared with the control data set). Thus, we can detect and treat outliers in an interactive way based on data depth.

2.3 OD Trajectory Comparisons Based on Depth

Data depth can compare bivariate data related with two independent groups. A t-test can be used to compare means from two independent groups. For example, the t-test reveals whether two OD flows' means are different at two different temporal ranges. However, it is worth examining how groups differ in terms of scale (spread). The comparisons of central regions in data depth compare the marginal distribution, thereby considering the overall structure of the data [**?**].

Let X and Y be the random variables having distributions F and G for two independent groups. The quality index proposed by [**?**] is the probability that the depth of Y is greater 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 a randomly sampled observations according with the distribution F. [**?**] presents the range of Q is $[0, 1]$ and $Q(F, G) = 0.5$ if and only if $F = G$. If $Q < 0.5$ or $Q > 0.5$, a scale increases or decreases from F to G. It is therefore possible to detect differences in scale based on 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 as 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 below.

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

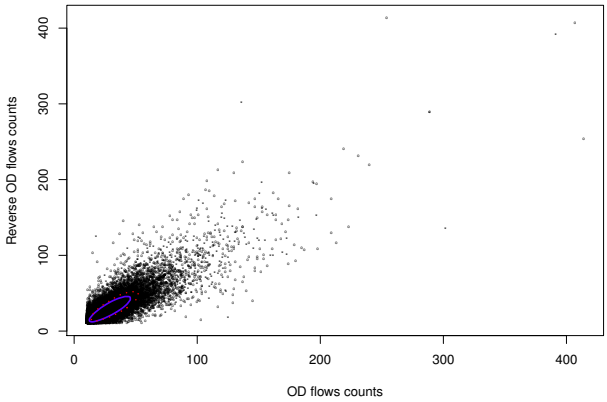


Figure 4 Central regions of two OD flows: ∘ indicates the OD flows for Saturday, March 29 2014 and * indicates the OD flows for a list of Saturdays; blue line presents the central region of the OD flows for the list of Saturdays and red dotted line presents the central region of the OD flows on March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

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 [**?**, **?**].

For ease of understanding, Figure **??** presents the central regions of two OD flows. One data set is OD flows for Saturday, March 29 2014 and the other is the data set of Saturdays which include March 1, 8, 15, 22, and April 5. March 29 has the highest number of taxi trips of the year (i.e., 552,064 taxi trips). The data set for the list of Saturdays has 2,621,703 taxi trips. The bootstrap method reveals that the confidence interval is 0.02467486 and 0.0595938. This confidence interval does not include zero, which results in rejecting the H_0. This means that the amount of scale is significantly changed between two OD flows. Further, it is possible to see the OD flows from a list of Saturdays is nested within the OD flows corresponding to March 29. This is additional perspective based on data depth comparisons.

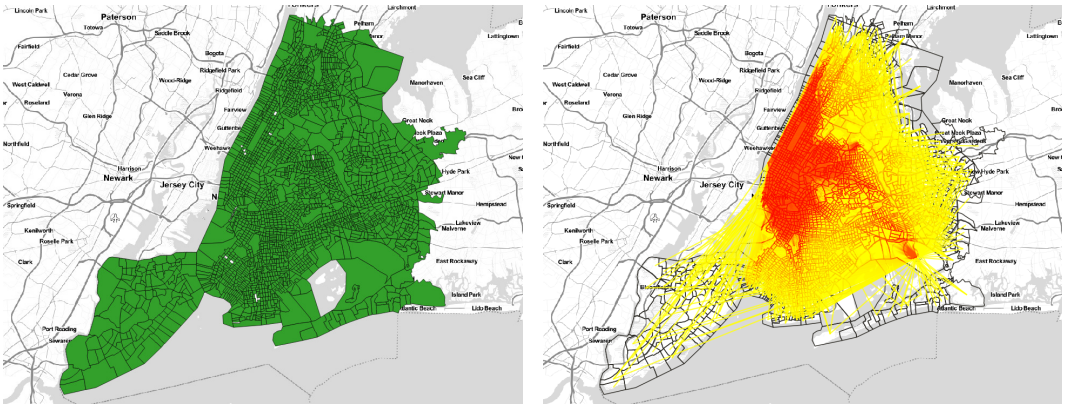
In addition, it is apparent that the bootstrap method is a time consuming process. We generate 2,000 bootstrap samples. In order to scale up the bootstrap computation, we ~~spread~~ distribute work across multiple nodes and cores by implementing an embarrassingly parallel R code. Thus, we can effectively improve execution ~~times~~efficiency.

3 Experiments

3.1 Data

This study uses New York City Taxi data in 2014 to evaluate the effectiveness of the proposed approach. Figure **??** presents taxi analysis zones in New York City which indicate the origin and the destination IDs of OD flows. Figure **??** shows OD flows on July 1. Red lines indicate the dominated OD flows.

As a case study, this study used OD flows during weekdays and weekends in June~~.~~, 2014. The weekdays data set includes taxi trajectories on June 3, 10, 17, and 24. There are 1,721,655 taxi trips. The weekends data set includes taxi trajectories on June 8, 15, 22, and



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

Figure 5 Experimental data: New York City Taxi data

29. Its taxi trips are 1,593,480.

3.2 Procedure

The performance of the proposed method was compared with alternative methods. Trajectories anomalies detection, based on Mahalanobis distance [**?**], was used to compare the performance of outliers detection. The Mahalanobis distance considers the correlations of the data (e.g., two OD flows), which distinguishes it from Euclidean distance. According to [**?**], the anomaly detection threshold can be defined as below.

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

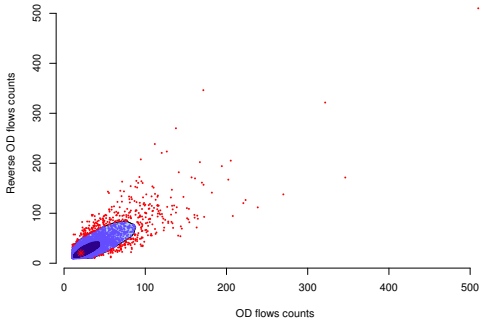
where OD_{t_1} is the current OD flow, and $\mu_{[t_0,t_1)}$ is the median of all OD flows during $[t_0, t_1)$. In addition, we visualized the results in order to compare them and make the difference easier to understand. In terms of the difference of scale, we used standard statistics such as F-test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

With regard to data cleaning process, this study used Hadoop with Pig. We have developed a Hadoop program to handle the large volume of data (173 million taxi trip records), remove trips with invalid OD coordinates, and assign each OD locations into the corresponding traffic analysis zone. R was used to implement the ~~code in~~ OD flow outlier detection. The computing environment uses Amazon Web Service and Bridges supercomputer ~~resources~~at the Pittsburgh Supercomputing Center. Further, we used OD flows that have more than 10 trips. It is reasonable to remove the majority of low number of OD flows, which may distort 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).

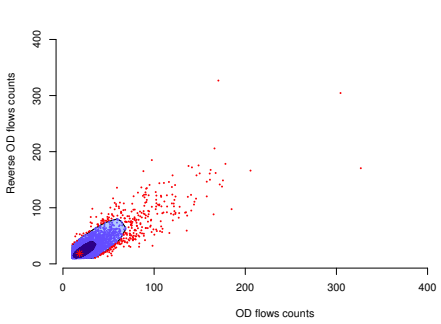
3.3 Case study: weekdays vs weekends

3.3.1 Outlier Detection

The bag plot presented OD flow outliers on weekdays and weekends separately in Figure **??**. The outliers are detected by considering forward OD flows and reverse OD flows together.

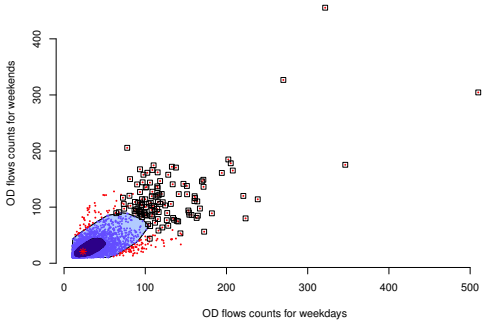


(a) Bag plot on weekdays

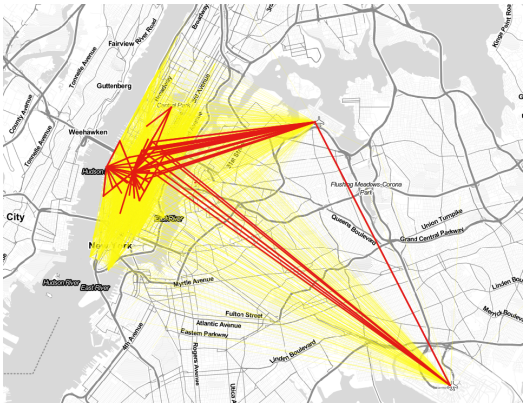


(b) Bag plot on weekends

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



(a) High volume of OD flows



(b) OD flow map for high volume of OD flows

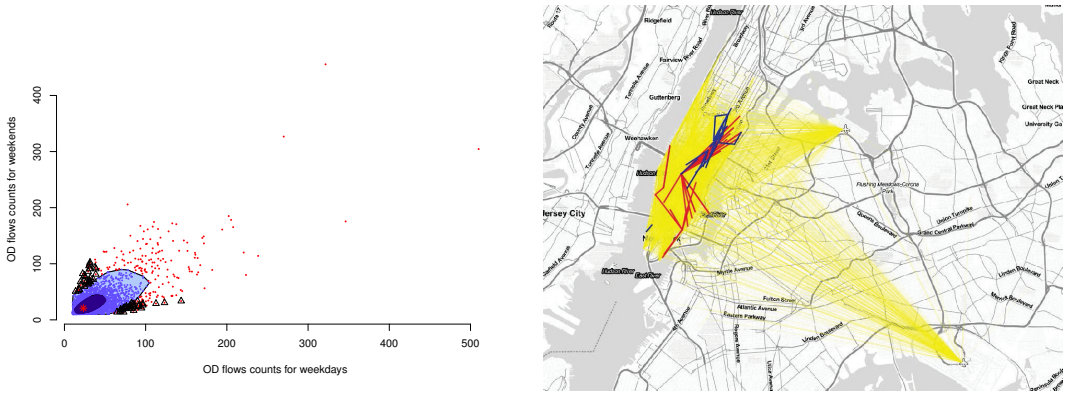
Figure 7 Outliers with high volume of trips on weekdays and weekends: Rectangles in Figure **??** coincide with red lines in Figure **??**.

In order to find the difference between two data sets, we considered two forward OD flows together with the bag plot. Then we found the outliers of OD flows in Figure **??**. In particular, the outliers with rectangle symbols indicate that these OD flows have high volume of taxi trips during weekdays and weekends. These outliers are superimposed on a map with red lines in Figure **??**. The yellow lines present OD flows except the high volume of OD flows on weekends.

As this case very clearly demonstrated, the majority of OD flows occurred in broadly two places: within Manhattan, and between the center of Manhattan and two airports (i.e., J.F.K International airport and LaGuardia airport).

In addition, we investigated which OD flows are abnormal on weekends. However, they are typical OD flows on weekdays. These OD flows can have significantly more taxi trips or less taxi trips compared with the OD flows on weekdays. Figure **??** presents these OD flows outliers with triangle symbols. These OD flows are represented on a map with red lines in Figure **??**.

Further, we detected OD flows outliers with Mahalanobis distance. The results are



(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 **??** coincide with red lines in Figure **??**.

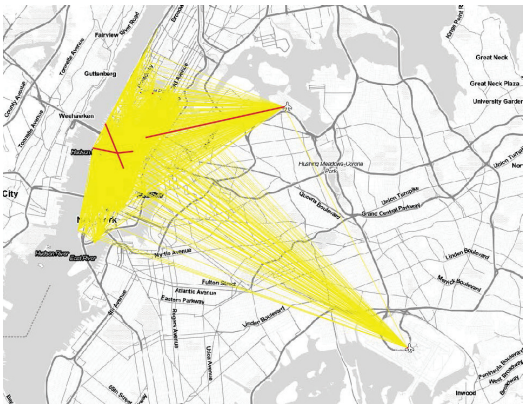


Figure 9 OD flows outliers on weekdays and weekends based on Mahalanobis distance.

presented in Figure **??**. The number of OD flows outliers are very small compared with our approach. In particular, this method only considers forward OD flows with two data sets. It identified OD flows outliers with high volume of trips because Mahalanobis distance takes into account the correlations between two OD flows. It is more likely to be outliers when two OD flows have high volume of trips. In fact, the OD flows outliers from Mahalanobis distance are the subset of them from our methods in Figure **??**.

3.3.2 Comparisons in scale

We further investigated how two OD flows differ. Our approach is sensitive to the difference in scale. The hypothesis testing whether two central regions in Figure **??** are different by chance revealed that the confidence interval was -0.02769909 and 0.01573812, which includes zero. Thus, it failed to reject the null hypothesis. They were the same in terms of the spread.

Interestingly, the standard statistic such as F-test was significant, $F(9530, 7637) = 1.1786$, $p \leq 0.05$. The variances of two groups were significantly different. This is the opposite result compared with our approach.

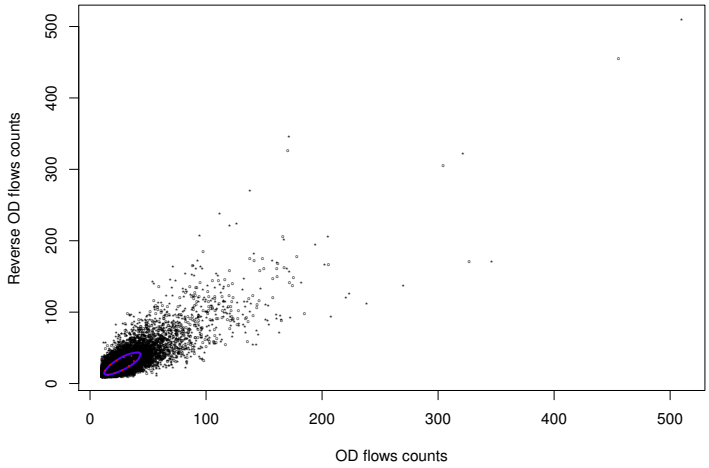


Figure 10 OD flows comparisons based on data depth: ∘ indicates the OD flows on weekdays and * indicates the OD flows on weekends; blue line presents the central region of the OD flows for the weekdays and red dotted line presents the central region of the OD flows on weekends.

4 Discussion

Our results demonstrate that the proposed method can identify OD flows outliers interactively, based on data depth. It is possible to detect OD flows outliers by querying with conditional clauses (e.g., which OD flows outliers have always high volume of trips during time t_1 and time t_2?).

The state-of-the-art Mahalanobis distance approach as an alternative method detects similar OD flows outliers. However, the number of outliers is quite different. The reason behind it is that our OD flows data has heavy tail distributions (e.g., there are a lot of OD flows with a long distance from the depth median in Figure **??**). Mahalanobis distance is well known to be unsatisfactory when the underlying data has heavy tail distributions [**?**]. Thus, the very presence of outliers may mask the outliers detection in Mahalanobis distance approach. Further, it can only detect OD flows outliers that have high number of trips during time t_1 and time t_2. It is difficult to detect OD flows outliers that have different properties (e.g., which OD flows outliers have significantly high number of trips compared with time t_1 and time t_2?).

In terms of the difference in spread, our approach uses a bootstrap method to compare the central regions of data depth. This approach can investigate not only the difference in scale, but also the structure of data. It can provide information how deeply points from group 1 (e.g., OD flows at t_1) tend to be located within group 2 (e.g., OD flows at t_2). General statistics such as F-test only provide their difference in variation. However, it does not give how groups differ.

Interestingly, F-test presents there are statistically significant difference in terms of variation for the OD flows on weekdays and weekends. However, our approach shows there are no statistically significant differences. The difference may be caused by the sensitivity of F-test to non-normality [**?**], which increases the Type-I error rate. Conversely, data depth has no assumptions about the distributions of the underlying data set.

5 Conclusions and Future Work

This paper provides a new methodology for identifying OD flows outliers and the difference in scale between two different OD flows at t_1 and t_2. The method is based on the concept of data depth. Data depth is robust statistics, which is suited to non-Gaussian distribution of the underlying data sets. Compared with standard statistics, it adds different perspective how two OD flows differ and by how much.

However, this study makes no attempt to consider geographic context such as locational circumstances or surrounding environment for understanding OD flows. Ultimately, further investigation should focus on integrating the analysis of OD flows with geographic context. Such an effort will lead to knowledge discovery for understanding the dynamics of urban flow.

Acknowledgements.