Outlier Detection in Urban Traffic Flow Distributions

Youcef Djenouri, Arthur Zimek, Marco Chiarandini
University of Southern Denmark, Odense, Denmark
{djenouri,zimek,marco}@imada.sdu.dk

***Abstract*—Urban traffic data consists of observations like number and speed of cars or other vehicles at certain locations as measured by deployed sensors. These numbers can be interpreted as traffic flow which in turn relates to the capacity of streets and the demand of the traffic system. City planners are interested in studying the impact of various conditions on the traffic flow, leading to unusual patterns, i.e., outliers. Existing approaches to outlier detection in urban traffic data take into account only individual flow values (i.e., an individual observation). This can be interesting for real time detection of sudden changes. Here, we face a different scenario: The city planners want to learn from historical data, how special circumstances (e.g., events or festivals) relate to unusual patterns in the traffic flow, in order to support improved planing of both, events and the layout of the traffic system. Therefore, we propose to consider the sequence of traffic flow values observed within some time interval. Such flow sequences can be modeled as probability distributions of flows. We adapt an established outlier detection method, the local outlier factor (LOF), to handling flow distributions rather than individual observations. We apply the outlier detection online to extend the database with new flow distributions that are considered inliers. For the validation we consider a special case of our framework for comparison with state-of-the-art outlier detection on flows. In addition, a real case study on urban traffic flow data showcases that our method finds meaningful outliers in the traffic flow data.**

I. Introduction

In the analysis of urban traffic we aim to learn from the behavior of independent participants (cyclists, cars, trucks, and public transportation) under different conditions (weather, events, maintenance of streets) to support decisions of city planners and managers on the layout of streets, regulation systems (e.g., traffic lights), and routes for public transport, or temporarily invasive decisions in planning construction sites. An important basis for the description of the complex traffic system is the estimation of the traffic flow, based on counting the number of objects (e.g., pedestrians, bicycles, cars, trucks, buses) that cross a given location during some time interval by means of various types of sensors in streets, in traffic light systems, or as mobile sensors. The flow, i.e., the number of objects passing a specific location within a specified timeframe (or a "window" over the time axis) varies over the day and between different days of the week. For the city planners it is important to understand the impact of events, particular weather conditions, or planning decisions on the traffic flow in the city. In this study, we therefore consider outlier detection on traffic flow data and the relation of outliers to special circumstances.

The detection of anomalies (outliers) in the traffic flow by the application of adapted outlier detection techniques is one of the main applications in the analysis of urban traffic data. An outlier can be defined as "an observation (or a set of observations) which appears to be inconsistent with the remainder of that set of data" [2]. Outlier detection has been studied intensely in the two last decades in an abstract setting [5], [13], [17], [23] as well as in application scenarios such as spatial data [9], [19]. Also many algorithms have been developed to identify outliers in traffic flow [4], [12], [14], [20], [22]. However, these algorithms detect only single flow outliers and ignore the correlation between the flow values.

In this paper, we propose a different approach, as the interest in collaboration with the city of Odense, Denmark, is foremost not on real-time detection of outliers (unusual flow values in short time frames) but on the impact of certain events on the traffic flow over a longer time frame (e.g., a few hours, a day). We therefore resort to capturing the flow distribution over a longer time frame. The distribution of flows is defined by the set of flows (e.g., cars per minute per location) captured during a specific time period (e.g., rush hour on Mondays), which immediately relates to a probability distribution of flows (or flow probability distribution, FPD).

In this paper, we propose a framework for outlier detection in flow distributions. To the best of our knowledge, we are the first to deal with flow distributions. The main contributions of the paper are summarized as follows: (1) We show that sets of flows can be interpreted as probability distributions of flows. (2) We propose a framework that updates the historical data for dealing with flow probability distribution outliers from the distribution of flows. (3) We propose a strategy of constructing the database of historical flow probability distributions, by taking into account the temporal information of the flow distributions. (4) We propose an adaptation FPD-LOF of the local outlier factor (LOF) algorithm [5] by adapting the Bhattacharyya similarity measure [3] for detecting flow probability distribution outliers. (5) We apply FPD-LOF to a special case to allow for comparison with existing approaches. Experimental analysis shows that FPD-LOF outperforms the state-of-the-art flow outlier detection algorithms. In addition, a case study on real urban traffic flow data demonstrates the practical usefulness of the proposed framework. The results reveal that FPD-LOF using Bhattacharyya metric identifies meaningful outliers relating to unusual weather conditions or special events in the city.

In the remainder, we survey existing outlier techniques for traffic data (Section II), we present the overall framework and the adaptation FPD-LOF for outlier detection in flow probability distributions (Section III), we perform an experimental analysis of the framework and method on synthetic data as well as a case study on real data (Section IV), and conclude the paper with a perspective on potential future work (Section V).

II. Related Work

Several surveys on outlier detection algorithms for traffic flow data have been published [10], [11], [15]. We refer to our recent overview [8] and sketch here only the methods that we use in the experiments as competitors.

Ngan et al. [16] used a DPMM (Dirichlet Process Mixture Model) for deriving outliers in urban traffic flow data. First, the set of all flow values $F = \{f_1, f_2, \ldots, f_{|F|}\}$ is projected into an n-dimensional space, where the i^{th} object is defined by the flow values $\{f_i, \ldots, f_{i+n-1}\}$. The obtained dimensions are then reduced by PCA (Principal Component Analysis) to a two-dimensional space. Then, the Chinese restaurant process [1] is performed to cluster the flow values with an infinite number of clusters. Each flow value is assigned to a new cluster with a probability proportional to a concentration parameter α, otherwise, it is assigned to the previously created cluster. Afterwards, all flow values belonging to the cluster having a maximum number of elements are considered inliers, the remaining flow values are deemed outliers.

Ye et al. [22] present an anomaly-tolerant traffic matrix estimation approach called SETMADA (Simultaneously Estimate Traffic Matrix and Detect Anomaly). It estimates the traffic matrix and uses it for anomaly detection. Based on the prior low-rank property and temporal characteristic of the traffic flow, the outlier detection is formulated as a prior information-guided matrix completion problem.

Dang et al. [7] proposed a combination between kNN [17] and PCA for outlier flow detection. A dimensionality reduction is performed by PCA. In the derived subspaces the kNN outlier detection [17] is applied.

Tan et al. [21] proposed a density-based bounded application of LOF for large scale traffic flow data in Hong Kong. A three dimensional space is derived by PCA, then the LOF algorithm [5] is applied on this reduced space to find local outliers in the flow data.

III. Outlier Flow Probability Distribution Detection

A. Problem statement

In this paper, we focus on detecting anomalous flow probability distributions. A *traffic flow* is defined as the number of vehicles passing through a location (a point in the road network) during a given time interval. A *flow probability distribution* (FPD) links flow values to their likelihood of occurrence during a given period of time. We estimate traffic flow probability distributions on the basis of their empirical counterparts based on real-life measurements.

Let I be the set of time instants at regular time intervals at which we collect flow measurements at a specific location and let $X = [x_1, ..., x_{|I|}]$ be the list of corresponding flow values. Let λ be the duration of the time interval between two consecutive measurement instants and δ the duration considered by each flow measurement $x \in X$. We will assume $\lambda = \delta$.

From I we can extract a collection $\mathcal{T} = \{T_1, \ldots, T_\tau\}$ of non-intersecting subsets of μ consecutive time instants. For a subset T_j, $j = 1, \ldots, \tau$, we identify the time instant where the subset begins by $\iota(T_j)$, that is, $T_j = \{\iota(T_j), \iota(T_j) + 1, \ldots, \iota(T_j)+\mu\}$. To each T_j, $j = 1, \ldots, \tau$, there is associated a set of flow measurements X_{T_j}. Thus, for example, we can create the collection of sets X_{T_j}, $j = 1 \ldots 7$ each containing the flow measurements between 7:00 and 10:00 on the seven different days of a week.

The flow measurements in each set X_{T_j}, $j = 1, \ldots, \tau$, can be represented as discrete random variables $Y_j \in \mathbb{N}_0$ to capture the uncertainty related to those measurements. Consequently, each Y_j can be described by its probability mass function $f_{Y_j}\colon \mathbb{N}_0 \to [0, 1]$ defined as $f_{Y_j}(y) = \Pr(Y_j = y)$, $y \in \mathbb{N}_0$. To estimate $f_{Y_j} = f_j$, we use the empirical probability distribution $\hat{f}_j$ of the flow given by the relative frequency of the measurements contained in X_{T_j}, that is:

$$\hat{f}_j(y) = \frac{|\{x = y \mid x \in X_{T_j}\}|}{|X_{T_j}|}, \quad y \in \mathbb{N}_0, j = 1, \ldots, \tau.$$

We name such an estimated probability distribution of flow values a flow probability distribution (FPD). Outliers in a set of FPDs could be defined by some outlier scoring function and some threshold as follows:

Definition 1 (FPD Outliers): Given a family of empirical FPDs $F = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_\tau\}$ derived from a collection of flow measurements $\{X_{T_1}, \ldots, X_{T_\tau}\}$, a scoring function $s\colon F \to \mathbb{R}$ that assigns outlier scores to some FPD, and some threshold θ; FPD outliers are the members of the set $O \subseteq F$, such that:

$$O = \{\hat{f}_j \in F \mid s(\hat{f}_j) \geq \theta\}.$$

B. Outlier detection in a growing database of traffic flow data

In the application scenario with our partners in the municipality, observations of the traffic flow are collected continuously over time.

There are specific questions of interest that can be answered based on the aggregation of flows over pre-specified time-intervals (e.g., rush hour during weekdays, afternoon and evening during weekends and holidays) to study the impact of interferences with the normal traffic behavior (e.g., closing of routes for construction during the rush hour, impact of special events such as sport events or festivals during weekends and holidays). Detected outliers can be verified and excluded from the database of normal traffic behavior.

Our infrastructure for the detection of outliers consists of:

1) Construction and update of a database of historical FPDs:

a) extraction of information from raw traffic flow data received from sensors,
b) synthesizing data for pre-specified time intervals via empirical FPD.

2) (Online) detection of outlier FPDs: The historical FPDs are used as a reference set for each new FPD in order to detect outliers.
   - If the new flow distribution is not an outlier, it is added to the historical data.
   - If it is suspected to be an outlier, it is not added to the historical reference database.

The first component follows the definitions in Section III-A. To carry out the second task we adapt LOF [5], as we discuss in the following.

C. *Adaptation of LOF for FPD outlier detection*

We base the outlier procedure applied in our framework on the classic method LOF (Local Outlier Factor) [5]. Other methods could be adapted in a similar way. We chose LOF as it is well-known and it has been shown to be still state-of-the-art [6] and suitable for generalizations to different data types and scenarios [19].

1) Similarity measure for FPDs: For the representation of an empirical FPD $\hat{f} : X_T \to [0, 1]$ we use a vector $\mathcal{A}$ of length $d = \max\{X_T\}$ to represent $\hat{f}$. $\mathcal{A}$ contains the estimated probability density values of all possible flow values up to d:

$$\mathcal{A}[m] = \hat{f}(m) \quad \forall m \in [0, d\}]. \tag{1}$$

To compare two FPDs with vector representation of different size, we project the lower size vector to the greater size one as follows.

Definition 2 (Vector projection): Let $\mathcal{A}_i$ and $\mathcal{A}_j$ be vector representations of two FPDs $\hat{f}_i$ and $\hat{f}_j$. Without loss of generality, let $d_i \geq d_j$. $\mathcal{A}_j$ is projected to a d_i-size representation by setting all the missing values of $\mathcal{A}_j$ to 0, i.e.,

$$\mathcal{A}_j(m) = \begin{cases} \mathcal{A}_j(m) & \text{if } m \in [0, d_j] \\ 0 & \text{if } m \in [d_j + 1, d_i] \end{cases}$$

Many distance measures for computing the similarity between two probability distributions exist in the literature such as the Euclidean distance, the Jaccard similarity, the Kullback-Leibler-divergence, and the Bhattacharyya distance [3]. Here we choose the Bhattacharyya distance.

The Bhattacharyya distance $B(\mathcal{A}_i, \mathcal{A}_j)$ expressing the similarity between two FPDs represented by $\mathcal{A}_i$ and $\mathcal{A}_j$ is defined as follows:

$$B(\mathcal{A}_i, \mathcal{A}_j) = -\ln \sum_{m=1}^{d_i} \sum_{k=1}^{d_j} \sqrt{(|m - k|) + (\mathcal{A}_i(m)\mathcal{A}_j(k))} \tag{2}$$

2) Local outlier factor for FPDs: Let $F = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_\tau\}$ be a family of FPDs, $\hat{f}$ a new FPD not in F, and $k < \tau$ a parameter for the size of the set $k\text{NN}(\hat{f}) \subseteq F$ consisting of the k FPDs from F that are most similar to $\hat{f}$. We denote by $k\text{NN-}dist(\hat{f}_i)$ the distance between some FPD $\hat{f}_i$ and the k^{th} most similar FPD.

In LOF, the kNN-$dist$ is the most fundamental ingredient for density estimates. Outliers are objects with a relatively low local density as compared to their k nearest neighbors. These density estimates typically relate to Euclidean space. However, the general LOF pattern has been extended to many other, non-Euclidean applications as well [19]. In our adaptation we have the equivalent in Bhattacharyya space. The components to derive the local outlier factor (LOF) are [5] the local reachability density (lrd) and the local outlier factor (LOF) based on the lrds:

The local reachability density (lrd) is defined as follows:

$$\text{lrd}(\hat{f}) \quad := \quad 1 \Bigg/ \frac{\sum_{\hat{f}_i \in k\text{NN}(\hat{f})} \text{reach}_k(\hat{f}, \hat{f}_i)}{|k\text{NN}(\hat{f})|} \tag{3}$$

where reach_k is the so called reachability distance, given by:

$$\text{reach}_k(p, o) \quad := \quad \max\{k\text{NN-}dist(o), dist(p, o)\}. \tag{4}$$

The function $dist$ designates the basic distance measure used in the data space. In standard applications, often the Euclidean distance is used. Here we use the Bhattacharyya distance (Eq. 2) for measuring the distance between two FPDs.

Local reachability densities (lrds) are local density models for each FPD. The local outlier factor (LOF) [5] is the average ratio between the lrds of the FPDs in $k\text{NN}(\hat{f})$ and $\text{lrd}(\hat{f})$:

$$LOF(\hat{f}) \quad := \quad \frac{1}{|k\text{NN}(\hat{f})|} \sum_{\hat{f}_i \in k\text{NN}(\hat{f})} \frac{\text{lrd}(\hat{f}_i)}{\text{lrd}(\hat{f})} \tag{5}$$

The intuition is that those FPDs are deemed unusual that have on average a larger Bhattacharyya distance to the k most similar other FPDs than those k most similar FPDs in turn have to their k most similar FPDs. Thus, $LOF(\hat{f}) > 1$ signals outlierness of $\hat{f}$. We use therefore 1 as a conservative cut-off threshold, i.e., if $LOF(\hat{f}_{new}) \leq 1$, then $\hat{f}_{new}$ will be considered an inlier and can be added to the database of historical records.

IV. Experimental Evaluation

A number of experiments have been carried out to demonstrate the performance of the proposed framework using both synthetic data and real urban traffic flow data.[1] The evaluation is performed using F-measure and the area under the curve of the receiver operating characteristic (ROC AUC), common measures for the evaluation of outlier detection methods [6].

A. *Time Series Datasets*

In a first experiment, we compare FPD-LOF to the recent use of LOF for detecting outliers in time series (BLOF [21]). We use three real-world time series datasets.[2] The datasets are described in [18] and collected from various domains.

1) **EnronInc** This dataset comprises four years (1999–2002) of Enron email communications. The Enron email network contains a total of 80.884

[1] Source code and data are available at: http://dss.sdu.dk/projects/its.html.
[2] http://odds.cs.stonybrook.edu/

points. The ground truth identifies the major events in the company's history, such as revenue losses and restatements of earnings.

2) **RealityMining** This dataset contains the communication flow data at MIT university recorded continuously via preinstalled software on their mobile devices over 50 weeks. The sequences of weekly temporal flows are built for three types of relations, voice calls, short messages, and bluetooth scans. The ground truth captures semester breaks, exam weeks, and holidays.
3) **TwitterWorldCup** This collection contains data related to the World Cup 2014, June 12 to July 13. The tweets are filtered by popular or official World Cup hashtags, such as #worldcup, #fifa, #brazil, etc. The ground truth contains the goals, penalties, and injuries in all the matches that involve at least one of the following renowned teams (Brazil, Germany, Argentina, Netherlands, Spain, and France).

While BLOF works on the time series as such, FPD-LOF works on the distributional representation of the time series. Figures 1(a), 1(b), and 1(c) present the ROC AUC value on the three time series datasets (EnronInc, RealityMining, and TwitterWorldCup) of FPD-LOF and BLOF. By varying the size of the neighborhood from 10 to 100, FPD-LOF outperforms BLOF in all settings. In addition, the difference between both algorithms increases for RealityMining and TwitterWorldCup datasets. This can be explained by the fact that the FPD-LOF algorithm uses sequence flow values by taking into account the correlation between the flows as opposed to BLOF where individual flow values are used to determine outliers. This result confirms that using a distributional representation can be superior to the classic time series approach.

B. Real Data

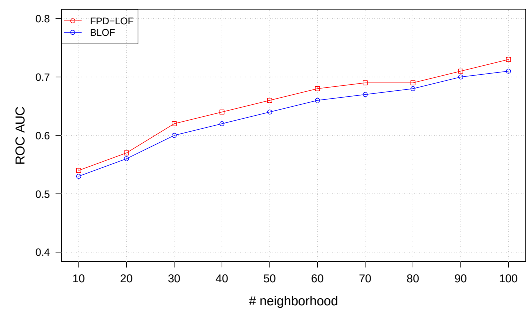
From our collaboration with the city of Odense we have data from several test locations throughout the city area. Each data entry contains information related to the vehicle detected at specific locations such as: gap, length, date, time, speed, and class (i.e., type of vehicle). For ten locations, sensor infrastructure has been installed in a pilot experiment. The ten locations have different characteristics (traffic density, counters for cars or for bikes) as described in Table I. The traffic data were obtained between January 1^{st}, 2017 and September 30^{th}, 2017.

C. Quantitative Analysis

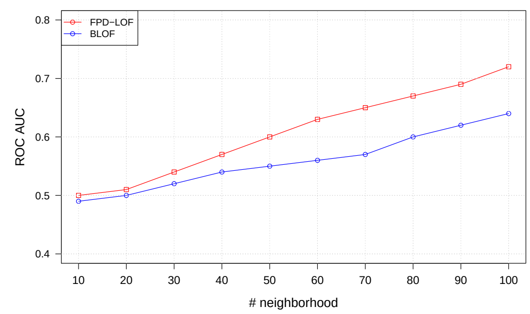
A common problem in the evaluation of outlier detection techniques using new data is that outliers are not labeled. To facilitate a quantitative evaluation on the real data, we inject in $F = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_\tau\}$ with $d^* = \max\{d_j \mid j = 1..\tau\}$ synthetic outliers f_i in different variants:

1) Null FPD: In the null FPD, the flow distribution is equal to 0 for any positive flow. In other words, the street is always empty during the observation (see Figure 2(a)). Formally, a null FPD is defined as:

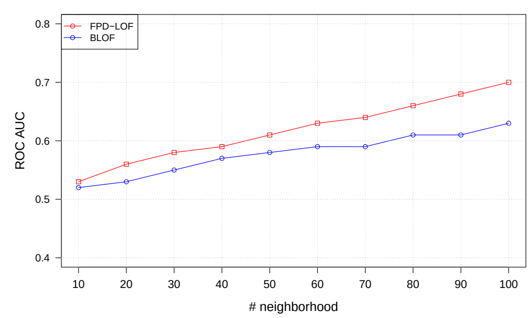
$$f_i(m) = 0, \quad m = 1 \ldots d^* \tag{6}$$



(a) EnronInc



(b) RealityMining

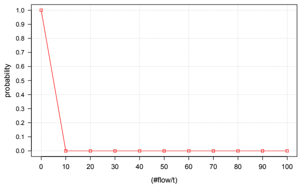


(c) TwitterWorldCup

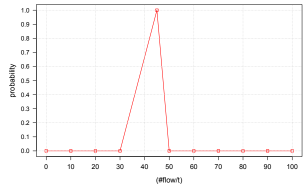
Fig. 1. ROC AUC of FPD-LOF and BLOF for time series data sets.

TABLE I
DATA DESCRIPTION

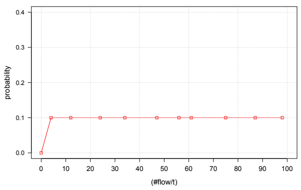
Address	ID	Type	#(Cars or Bikes)
Falen	L_1	Cars	16.932
Anderupvej	L_2	Cars	25.310
Åløkke Alle	L_3	Cars	238.775
Thomas B Thriges Gade A Syd	L_4	Bikes	46.978
Niels Bohrs Alle	L_5	Bikes	445.883
Rødegårdsvej Østgående	L_6	Bikes	575.089
Rugårdsvej	L_7	Cars	2.318.852
Nyborgvej	L_8	Cars	2.352.930
Grønlandsgade	L_9	Cars	2.955.464
Odins Bro	L_{10}	Cars	3.921.746



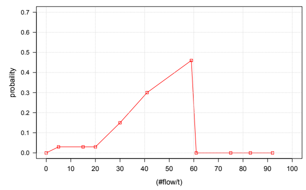
(a) Null FPD



(b) Stable FPD



(c) Regular FPD



(d) Unexpected FPD

Fig. 2. Simulated unusual FPDs (schematic examples).

2) Stable FPD: In the stable FPD, the flow distribution is equal to 1 for flow equal to x, 0 otherwise. In other words, the flow is always the same, x (see Figure 2(b)). Formally:

$$f_i(m) = \begin{cases} 1 & \text{if } m = x \\ 0 & \text{otherwise} \end{cases}, \quad m = 1 \dots d_i \tag{7}$$

3) Regular FPD: The flow here is equally distributed, i.e., all flow values are equally likely to occur (Figure 2(c)):

$$f_i(m) = \frac{1}{d_i}, \quad m = 1 \dots d^* \tag{8}$$

4) Unexpected FPD: These FPDs mock the behavior observed when an unusual event occurs with a strong impact on the traffic flow (e.g., festivals or accidents that cause some road closings). We have three stages (Figure 2(d)): a stable flow from 1 to x, a cumulated flow from x to y, and a null flow from y to d_i:

$$f_i(m) = \begin{cases} \epsilon & \text{if } 1 \le m \le x \\ \Psi(m) & \text{if } x < m \le y \\ 0 & \text{if } y < m \le d^* \end{cases}, \; m = 1 \dots d^* \tag{9}$$

Here $\Psi(m)$ is some function $[x...y] \to [\epsilon...(1-x\epsilon)]$ with the following properties:

$$\forall (m_1, m_2), m_1 \le m_2 \iff \Psi(m_1) \le \Psi(m_2) \tag{10}$$

$$\sum m\Psi(m) = (1 - x\epsilon) \tag{11}$$

5) Noise FPD: Noise FPDs are generated by adding Gaussian noise of variance σ_i with a certain probability $p \sim \mathcal{U}(0,1)$ and a threshold γ_i:

$$f_i = \begin{cases} f_i + n \sim \mathcal{N}(0, \sigma_i^2) & \text{if } p \ge \gamma_i \\ f_i & \text{otherwise} \end{cases} \tag{12}$$

We compare our FDP-LOF against three competitors: the work of Dang et al. [7], SETMADA [22], and the work of Ngan et al. [16] (see Section II).

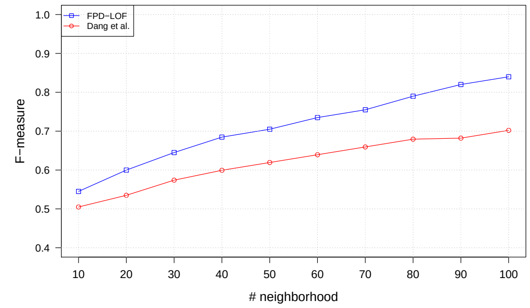


Fig. 3. F-measure of FPD-LOF and baseline1 on 10000 observations and over increasing neighborhood

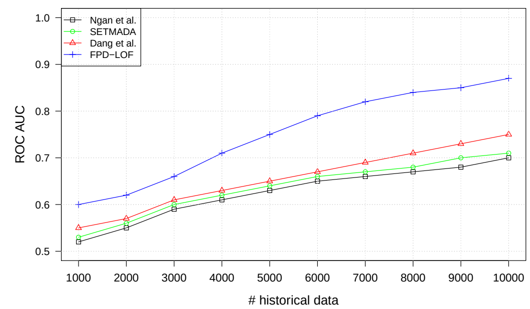


Fig. 4. ROC AUC of FPD-LOF and state-of-the-art outlier flow detection algorithms

In a first comparison, we evaluate the impact of the neighborhood size head-to-head with the work of Dang et al. [7] that uses the kNN approach with Euclidean distance for detecting outliers.

Figure 3 presents the F-measure of our FPD-LOF and Dang et al. [7] procedure with different size of the neighborhood on 10 000 observations sampled from location L_1 and 10% injected outliers randomly selected from the five types. For both algorithms, the quality increases with a larger neighborhood on this dataset. In all cases, FPD-LOF outperforms Dang et al. [7].

When we fix the neighborhood size to 100, we can compare all three competitors on the basis of the given data. Figure 4 presents the ROC AUC value achieved by the methods over varying database size (amount of historical data used). More data is helpful for all methods, but FPD-LOF benefits more. Anyway, in all cases FPD-LOF is superior to the competing approaches that do not take a distributional representation into account.

D. *Qualitative Analysis (Case Study)*

Tables II shows the top three outliers for each location along with an interpretation by connecting the dates to weather information or the event calendar of the city. Some outliers can be related to the weather information, others can be related to the city events.

We can remark from this table, that some flow distribution outliers can be justified by the weather information (very

TABLE II
THE TOP FPD-LOF OUTLIERS AT THE TEN LOCATIONS

Location	Date	Interpretation
L_1	04-04-2017	children sport event
	26-03-2017	very windy
	28-02-2017	very windy
L_2	04-04-2017	children sport event
	01-01-2017	new year holiday
	15-03-2017	very rainy
L_3	10-09-2017	very rainy
	09-09-2017	very windy
	03-09-2017	very windy
L_4	09-02-2017	farmer's market
	08-02-2017	farmer's market
	23-06-2017	very windy
L_5	10-03-2017	very windy
	08-02-2017	farmer's market
	09-02-2017	farmer's market
L_6	08-03-2017	women's day
	23-02-2017	national sport event
	14-02-2017	saint valentine's day
L_7	03-05-2017	very windy
	26-06-2017	very windy
	14-02-2017	saint valentine's day
L_8	23-02-2017	national sport event
	01-01-2017	new year holiday
	08-03-2017	women's day
L_9	23-01-2017	very cold
	08-03-2017	women's day
	19-05-2017	very windy
L_{10}	23-02-2017	national sport event
	01-01-2017	new year holiday
	12-07-2017	very windy

windy, very rainy, and very cold) of that day. For instance, the FPD of the day 28-02-2017 is an outlier because the weather was so windy that a substantial increase was observed in people that took cars and buses instead of bikes. Other flow outliers can be related to major events. For instance, the first day of the year (01-01-2017) has strong impact in three locations, also the celebration of the women's day (08-03-2017) shows up as an outlier in three locations. We can justify the first case by the fact that people tend to stay at home and take some reset after a celebration at the last night of 2016. However, we can justify the second case by the fact that women celebrate their day in public places (restaurant, cinemas, theaters, and so on).

These findings show that a traffic flow probability distribution can be unusual in different ways, be it higher or lower flow values than usual being dominant. Using a distributional representation of the flow values supports detecting such different manifestations of outlierness apparently in a better way than the traditional approaches.

V. CONCLUSION

We studied the representation of urban traffic flow data as flow probability distributions (FPDs) and proposed an adaptation of LOF for outlier detection in a database of FPDs, using the Bhattacharyya distance. Several experiments show the benefits of this novel treatment of traffic flow data compared to standard approaches. In an additional case study, looking into possible explanations of top outliers found on ten locations, we related these outliers to unusual weather and to city events.

REFERENCES

[1] Amr Ahmed and Eric Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *Proc. SDM*, pages 219–230, 2008.

[2] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley&Sons, 3rd edition, 1994.

[3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[4] Monowar H. Bhuyan, D. K. Bhattacharyya, and Jugal K. Kalita. A multi-step outlier-based anomaly detection approach to network-wide traffic. *Inf. Sci.*, 348:243–271, 2016.

[5] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *Proc. SIGMOD*, pages 93–104, 2000.

[6] Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Min. Knowl. Discov.*, 30(4):891–927, 2016.

[7] Taurus T. Dang, Henry Y. T. Ngan, and Wei Liu. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In *Proc. DSP*, pages 507–510, 2015.

[8] Youcef Djenouri and Arthur Zimek. Outlier detection in urban traffic data. In *Proc. WIMS*, pages 3:1–3:12, 2018.

[9] Marie Ernst and Gentiane Haesbroeck. Comparison of local outlier detection techniques in spatial multivariate data. *Data Min. Knowl. Discov.*, 31(2):371–399, 2017.

[10] Zhenni Feng and Yanmin Zhu. A survey on trajectory data mining: Techniques and applications. *IEEE Access*, 4:2056–2067, 2016.

[11] Manish Gupta, Jing Gao, Charu C. Aggarwal, and Jiawei Han. Outlier detection for temporal data: A survey. *IEEE Trans. Knowl. Data Eng.*, 26(9):2250–2267, 2014.

[12] Tingshan Huang, Harish Sethu, and Nagarajan Kandasamy. A new approach to dimensionality reduction for anomaly detection in data traffic. *IEEE Transactions on Network and Service Management*, 13(3):651–665, 2016.

[13] E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proc. VLDB*, pages 392–403, 1998.

[14] Anukool Lakhina, Mark Crovella, and Christophe Diot. Diagnosing network-wide traffic anomalies. In *Proc. SIGCOMM*, pages 219–230, 2004.

[15] Xiaolei Li, Zhenhui Li, Jiawei Han, and Jae-Gil Lee. Temporal outlier detection in vehicle traffic data. In *Proc. ICDE*, pages 1319–1322, 2009.

[16] Henry YT Ngan, Nelson HC Yung, and Anthony GO Yeh. Outlier detection in traffic data based on the dirichlet process mixture model. *IET intelligent transport systems*, 9(7):773–781, 2015.

[17] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proc. SIGMOD*, pages 427–438, 2000.

[18] Shebuti Rayana and Leman Akoglu. Less is more: Building selective anomaly ensembles. *TKDD*, 10(4):42:1–42:33, 2016.

[19] Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Min. Knowl. Discov.*, 28(1):190–237, 2014.

[20] Dihua Sun, Hongzhuan Zhao, Hang Yue, Min Zhao, Senlin Cheng, and Weijian Han. St td outlier detection. *IET Intelligent Transport Systems*, 11(4):203–211, 2017.

[21] Jialing Tang and Henry YT Ngan. Traffic outlier detection by density-based bounded local outlier factors. *Inf. Techn. Industr.*, 4(1):6–18, 2016.

[22] Wencai Ye, Lei Chen, Geng Yang, Hua Dai, and Fu Xiao. Anomaly-tolerant traffic matrix estimation via prior information guided matrix completion. *IEEE Access*, 5:3172–3182, 2017.

[23] A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *WIREs DMKD*, page e1280, 2018.