Evaluation of Probability Distribution Distance Metrics in Traffic Flow Outlier Detection

Erik Andersen
Mathematics and Computer Science
University of Southern Denmark
Odense, Denmark
erand17@student.sdu.dk

Marco Chiarandini
Mathematics and Computer Science
University of Southern Denmark
Odense, Denmark
marco@imada.sdu.dk

Marwan Hassani
Mathematics and Computer Science
Eindhoven University of Technology
Eindhoven, The Netherlands
m.hassani@tue.nl

Stefan Jänicke
Mathematics and Computer Science
University of Southern Denmark
Odense, Denmark
stjaenicke@imada.sdu.dk

Panagiotis Tampakis
Mathematics and Computer Science
University of Southern Denmark
Odense, Denmark
ptampakis@imada.sdu.dk

Arthur Zimek
Mathematics and Computer Science
University of Southern Denmark
Odense, Denmark
zimek@imada.sdu.dk

ABSTRACT

Detecting outliers in traffic flow measurements on intersections can be helpful to infer correlation patterns in traffic networks. Recent approaches have proven the effectiveness of LOF-based outlier detection when applied over traffic flow probability distributions. However, these approaches used distance metrics based on the Bhattacharyya coefficient when calculating probability distribution similarity. Consequently, the limited expressiveness of the Bhattacharyya coefficient restricted the accuracy of the methods. The crucial deficiency of the Bhattacharyya distance metric is its inability to compare distributions with non-overlapping sample spaces over the domain of natural numbers. Traffic flow intensity varies greatly, which results in numerous non-overlapping sample spaces, rendering metrics based on the Bhattacharyya coefficient inappropriate. In this work, we address this issue by exploring alternative distance metrics and showing their applicability in a massive real-life traffic flow data set from 26 vital intersections in The Hague. The results on these data collected from 272 sensors for more than two years show various advantages of the Earth Mover's distance both in effectiveness and efficiency.

1 INTRODUCTION

Efficiently directing traffic within major cities can alleviate inconveniences caused by traffic congestion which, among others, impact the local economy and aggravate air pollution. Understanding dependencies in the traffic network can help traffic management decisions. By examining not only traffic flow, i.e., the amount of traffic within a given interval, but the distribution of the flow over a given number of intervals, we can determine novel correlations between occurrences of abnormal flow distributions (outliers) within intersections. The benefit of using flow distributions rather than single flow measurements is a more robust insight into the traffic flow. Combined with an outlier-based correlation between traffic intersections, we can discover dependencies of significant changes in the traffic flow of individual traffic intersections. Uncovering such correlations of abnormal traffic flow between traffic intersections provides interpretable values that can be helpful when planning traffic redirection to alleviate the propagation of traffic congestion.

Determining outliers of flow distributions was done by Djenouri et al. [11] using an adaption of the local outlier factor (LOF) algorithm. The algorithm used the Bhattacharyya distance metric to measure distances between distributions. The deficiency of the Bhattacharyya distance for comparing discrete probability distributions, such as traffic flow distributions, becomes evident when comparing two flow distributions that have no overlap between the sets of discrete measurements defining their sample spaces. In such a case, the Bhattacharyya distance is undefined. Undefined distances are not an issue for the Hellinger distance, a metric, which in the case of discrete distributions is, based on the Bhattacharyya coefficient, as the Bhattacharyya distance also is. Kullback-Leibler divergence is a popular similarity measure that also handles the issue well. Though, as the Hellinger distance, it has an upper bound, which we will show impacts its accuracy in the context of using it as a metric for local outlier detection. A perhaps slightly naive way of measuring distances between distributions could be to consider distributions as vectors and determine their squared Euclidean distance. This method has no apparent upper bound. However, since the flow distributions define a probability space, the upper bound is 2. Furthermore, it is unable to determine the lateral displacement of two distributions with non-overlapping sample spaces. That is, it cannot tell how far apart the sample spaces of the distributions are from each other. The Earth Mover's Distance (EMD) can address this final issue by determining distances between distributions by, figuratively speaking, considering one of them as piles of earth and the other one as holes in the ground. The distance is then the energy needed to distribute the piles of earth into the holes.

Recent approaches [10, 11, 13] have proven the effectiveness of LOF-based outlier detection when applied over traffic flow probability distributions. In particular, we extend in this paper the method by Djenouri et al. [11] and compare the suitability of the five previously mentioned metrics for determining flow distribution outliers using data from sensors in traffic intersections of The Hague. We then solidify the benefits of this approach by using the most effective metric, EMD, to determine the correlation of outlier occurrences within these intersections. These correlations are compared with the naively obtained correlations based on traffic flow intensities to highlight how remarkable correlations can be obtained by using flow distribution outliers.

In the remainder of this paper Section 2 provides a brief overview of related work. Definitions of flow probability distributions, the selected set of metrics, and the local outlier factor

algorithm are in Section 3, together with the definition of effectiveness. In Section 4 we apply the method to traffic intensity data obtained from various intersections in The Hague. We analyze several aspects of the metrics. We then compare our best method against a recent proposal and conclude with an example of use of our method for generating new insights from the real-life data set at hand. Section 5 summarizes the main results and outlines the work still to be done in the realm of flow probability distributions for use in density-based outlier detection.

2 RELATED WORK

Traffic flow analysis to study congestion is a critical area of research in the development of an efficient movement infrastructure [1, 23]. Aside from the classic forecasting methods from statistics [19], modeling traffic flow has been done using fuzzy logic and neural networks [2, 22]. Some models are based on sensors placed in the road, while other studies have used trackers on vehicles to model the traffic flow. In this study, the traffic data was collected by sensors placed in the lanes nearby intersections. In [8] different types of sensors are compared.

One definition of an outlier given by V. Barnett and T. Lewis is that it is "an observation (or a set of observations) which appears to be inconsistent with the remainder of that set of data" [3]. There have been several studies on outlier detection in a generalized setting [28]. Some studies within application scenarios [12, 21], compare local outlier detection techniques on spatial data, as we do in this study. Also, many algorithms have been developed to detect outliers in traffic flow [6, 15, 17, 24, 26]. However, all these methods focus on single value outliers, none handles flow distribution outliers.

Djenouri et al. [11] introduced the notion of flow probability density (FPD) and used an adaptation of the local outlier factor (LOF) algorithm [7] to calculate the FPD-LOF. In these calculations, they used the Bhattacharyya distance [5] as a similarity metric between distributions, noting that other metrics could be used as well. The FPD-LOF approach to determine distribution outliers has been used in [13] to enhance the effectiveness of a long short-term memory model. In a successive study, Djenouri et al. [9] showed that better results to determine outliers in flow distributions can be achieved by using an adaptation of the kNN algorithm in place of the LOF algorithm and the Kullback-Leibler (KL) divergence in place of the Bhattacharyya distance. A study by Bazan et al. [4] in the field of image classification, indicated that, when comparing as dissimilarity measure: KL divergence, Bhattacharyya distance, χ^2 statistic and Earth Mover's Distance's (EMD), the most effective turns out to be EMD. Other studies in the field of image classification, e.g. [4, 20, 27], have established similar results for EMD.

3 DEFINITIONS

3.1 Flow Probability Distributions

We consider a set of intersections $\mathcal{I}$ in an urban road network. At each intersection $I \in \mathcal{I}$ there are multiple sensors (e.g., one on each lane) collecting data in form of number of vehicles passed in a time interval of duration δ. For a sensor s of an intersection $I \in \mathcal{I}$, we denote the limited sequence of data collected in the set of ordered timestamps $T = \{t_1, \ldots, t_n\}$ each at a distance δ from the previous, as $\langle m_{s,t_1}, \ldots, m_{s,t_n} \rangle$. Each value m_{s,t_i}, $t_i \in T$, indicates the number of vehicles passed in the interval of duration δ starting at timestamp t_i and ending at t_{i+1}, i.e., the interval $[t_i, t_{i+1})$. For each such sequence of data we can calculate its

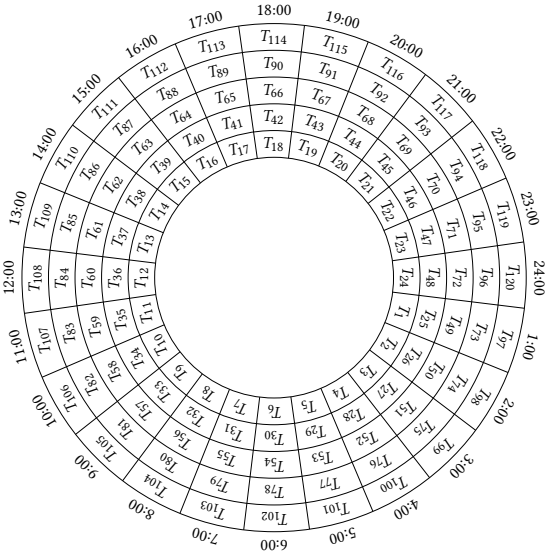

Figure 1: $\mathcal{T}$ partitioned into 120 subsets. Each sector represents a window with a time span of one hour and each inner circular band represents a different working day. The figure is for illustrative purposes only since in our experiments we consider $\mathcal{T}$ partitioned into 24×7 windows.

empirical probability distribution, that we call in this context *flow probability distribution*, FPD and given by the relative frequency of occurrences of some specific measurement $x \in \mathbb{N}_0$ in the sequence of n measurements collected in T. Formally, for a sensor $s \in I \in \mathcal{I}$ and its measurements at timestamps $T = \{t_1, \ldots, t_n\}$ the FPD for $x \in \mathbb{N}_0$ is given by:

$$FPD(x) = \frac{1}{n} \sum_{i=1}^{n} I\left(m_{s,t_i} = x\right) \tag{1}$$

where $I(E)$ is the indicator of event E. We can interpret the FPD as a zero-indexed vector of length $p = 1 + \max\left\{m_{s,t_i} \mid t_i \in \mathcal{T}\right\}$, i.e., we bound x in Eq. (1) by the highest value observed. In such an *FPD-vector*, X, an index, i, corresponds to the observed value, while the value at the index, X_i, represents the relative frequency of i. An FPD spans a time frame of $n \cdot \delta$. In most of our experiments, we will set δ to 5 minutes and $n = 12$ thus an FPD spans an hour. However, our interest will be in comparing FPDs associated to different sensors from $\mathcal{I}$ and different sets T. We will consider a long-time horizon made of a large number of consecutive timestamps $\mathcal{T}$ all at distance δ from the preceding one and we will subdivide $\mathcal{T}$ in subsets $T_1, T_2, \ldots$ of equal size n. Fig. 2 shows an example of traffic intensities and their corresponding FPDs. When comparing two FPD-vectors of different lengths, we use zero padding to extend the length of the shorter vector.

3.2 Distance functions

For two zero indexed vectors of size p, representing FPDs, X and Y, we can define the following alternative distance functions or metrics:

- *Squared Euclidean distance* is:

$$\mathrm{Dist}_E(X, Y) = \sum_{i=0}^{p-1} (X_i - Y_i)^2 \tag{2}$$

- *Kullback-Leibler divergence* adapted to FPDs such that lower values are given to more similar distributions and vice versa, as seen in [9]:

$$\mathrm{Dist}_{KL}(X,Y) = \sum_{i=0}^{p-1} X_i \ln\left(\frac{2X_i}{X_i + Y_i}\right) \tag{3}$$

- *Bhattacharyya distance*:

$$\mathrm{Dist}_B(X,Y) = -\ln(BC(X,Y)) \tag{4}$$

with:

$$BC(X,Y) = \sum_{i=0}^{p-1} \sqrt{X_i Y_i} \tag{5}$$

as the Bhattacharyya coefficient. If BC is 0, then the distance is undefined. In this paper, we handle this by defining the distance as infinite, however, as we shall see, this introduces problems when calculating FPD-LOF scores.

- *Hellinger distance*:

$$\mathrm{Dist}_H(X,Y) = \sqrt{1 - BC(X,Y)} \tag{6}$$

- *Earth Mover's Distance*:

$$\mathrm{Dist}_{W_1}(X,Y) = \sum_{i=0}^{p-1} |\,\mathrm{EMD}_i\,|, \quad \text{where:} \tag{7}$$

$$\mathrm{EMD}_i = \begin{cases} 0 & i = 0 \\ X_{i-1} + \mathrm{EMD}_{i-1} - Y_{i-1} & i \geq 1 \end{cases} \tag{8}$$

which corresponds to the *First Wasserstein distance*[18] for one-dimensional and discrete probability distributions.

The Euclidean distance Dist_E measures the distance between the two points in the space $[0,1]^p$ represented by the FPD-vectors. The Kullback-Leibler divergence is described as the information gained about a distribution Y by observing a distribution X. The adaptation of the divergence in Dist_{KL} is similar but in opposite direction. That is $\ln(2)$—the maximal value for any FPD-vectors X and Y in Eq. (3)—signifies no information-gain from the observation, and zero mean complete information about the distribution is achieved. The metrics using the Bhattacharyya coefficient, Dist_B and Dist_H, can be thought of as angle-based due to the similarity between BC and the dot product, and the relation between the dot product of two vectors and their angle. The First Wasserstein distance Dist_{W_1} takes into account the horizontal offset, i.e., the indices of the vector, between the mass represented by the corresponding values.

We say that two FPDs have no overlap if the two corresponding FPD-vectors, X and Y, both of length p have no overlapping values, i.e., $\forall i \in \{0 \ldots p-1\}$ it holds that $X_i > 0 \Rightarrow Y_i = 0$ and $X_i > 0 \Rightarrow Y_i = 0$.

Some distances achieve their theoretical maximum on non-overlapping FPD vectors. Let $X = \langle 0,0,0,0,1\rangle$ and $Y = \langle 0,0,0,1,0\rangle$ be two particular non-overlapping FPD-vectors. $\mathrm{Dist}_E(X,Y)$ yields 2, which is its maximum. $\mathrm{Dist}_{KL}(X,Y)$ also yields its maximum, which is $\ln(2)$. The Bhattacharyya coefficient $BC(X,Y)$ becomes 0 and $\mathrm{Dist}_B(X,Y) = -\ln(0)$, which is undefined. We correct these cases setting $\mathrm{Dist}_B(X,Y) = \infty$, which is its maximum. As we will see later, these cases are detrimental to our goals. The Hellinger distance does not suffer from this issue since $\mathrm{Dist}_H(X,Y)$ gives 1; its maximum value, when $BC(X,Y) = 0$. However, Dist_H and Dist_{KL} are not able to discriminate FPDs vectors with no overlaps even if the FPDs differ significantly in terms of mean and variance. For example, let $Z = \langle .5, .5, 0, 0, 0\rangle$ be another FPDs vector non-overlapping with X and Y. Then, $\mathrm{Dist}_{KL}(X,Z) = \mathrm{Dist}_{KL}(Y,Z) = \ln 2$ and $\mathrm{Dist}_H(X,Z) = \mathrm{Dist}_H(Y,Z) = 1$. The Euclidean distance is able to discriminate these cases. Tab. 1 reports the symmetric matrix of distances among X, Y, Z. Distances by $Dist_E$ indicate that X and Y are further apart than X and Y but this does not seem a good assessment since Y is horizontally closer to X than it is to Z. In other terms, under Dist_E FPDs with low variance will appear to be more distant from FPDs with high variance. As also shown in Tab. 1, $Dist_{W_1}$ brings another approach to these cases, and is able to differentiate by taking into account the lateral distance of the distributions.

Table 1: Distances calculated using Dist_E (left) and $Dist_{W_1}$ (right).

	X	Y	Z
X	0.0	2.0	1.5
Y	2.0	0.0	1.5
Z	1.5	1.5	0.0

	X	Y	Z
X	0.0	1.0	4.5
Y	1.0	0.0	3.0
Z	4.5	3.0	0.0

3.3 Local Outlier Factor

The local outlier factor (LOF) [7] assigns to each data point X from a metric space a score indicating the degree of outlierness for that point. It is based on the local density of the k nearest points, $N_k(X)$. With the metrics defined above the space made of probability distributions becomes a metric space and we can define the LOF for probability distributions in the same way as LOF is defined in [7]. Using one of the distance functions $Dist$, let $Dist^k(X)$ be the distance from a point X, now a probability distribution, to its kth nearest point, i.e., $Dist^k(X) = \max\{Dist(X,O) \mid O \in N_k(X)\}$. The reachability distance for a point X with respect to point Y is then:

$$\text{reach-dist}_k(X,Y) = \max\{\mathrm{Dist}^k(Y), Dist(X,Y)\} \tag{9}$$

The local reachability density of X is:

$$\mathrm{lrd}_k(X) = 1/\left(\frac{\sum_{Y \in N_k(X)} \text{reach-dist}_k(X,Y)}{k}\right) \tag{10}$$

which gives the local outlier factor for X:

$$\mathrm{LOF}_k(X) = \frac{\sum_{Y \in N_k(X)} \frac{\mathrm{lrd}_k(Y)}{\mathrm{lrd}_k(X)}}{k} \tag{11}$$

In the original paper [7], it is noted that the LOF-score varies with the value of k. To ensure statistical stability, it is recommend using at least 10 as the minimum value for k. This provides a lower bound on the number of elements needed to form a cluster. Likewise, setting an upper bound on k of 35 indicates the maximum number of elements near a cluster that could potentially be outliers. Thus, they suggest comparing the results for different values of k when calculating the scores and taking the maximum of them. Consequently, we defined the local outlier factor for a point, X, to be:

$$\mathrm{LOF}(X) = \max\{FPD - LOF_k(X) \mid k \in \{10, 15, \ldots, 35\}\}. \tag{12}$$

3.4 LOF for flow probability distributions

We denote by FPD-LOF the LOF when applied to a collection of flow probability distributions introduced in Sec. 3.1. The collections of FPDs that we will consider are derived as follows.

We first partition the timestamps of $\mathcal{T}$ into disjoint subsets, $T_1, T_2, \ldots$ each consisting of n consecutive timestamps. Then we group these subsets into collections, which we call *windows* (see

Fig. 1). A window W contains any T_i from $\mathcal{T}$ where it holds that the timestamps of T_i are within the window span given by a particular time of the day and day of the week. For example, a window could contain all timestamps between 08:00 and 09:00 of all Mondays present in our data. See Fig. 2. For a sensor s and for some window W we calculate the FPD-vectors for each $T_i \in W$. Then, for this collection of FPD vectors derived from a window we calculate the FPD-LOF. The pseudocode of this procedure is in Listing 1.

Listing 1: Pseudocode for the calculation of FPD-LOF scores of the FPD vectors associated to a window W and sensor s. The function `calculateLOFs` returns the LOFs using Eq. (12).

```
FPD-LOF(s, W)
  FPDs = []
  for T_i in W
    FPDs.push(FPD-vector(s, T_i))
  p = FPDs.length()
  Let M be a p×p matrix of zeroes
  for i in {0..p}
    for j in {i+1..p}
      M[i,j] = M[j,i] = dist(FPDs[i], FPDs[j])
  LOFs = calculateLOFs(M)
  return LOFs
```

3.5 Measuring effectiveness

We are interested in assessing how effective the distance metrics are in the calculations of FPD-LOFs. Since we have no ground truth about outliers, we will assume that measurements from sensors located in different lanes of the same intersection will have outliers at approximately the same time. If, for instance, one traffic lane is congested, several drivers may consequently switch lanes. It might not always be the case that outliers co-occur in all sensors within the same intersection. However, when this is the case, the FPD-LOFs of all the sensors should indicate this.

With these assumptions we define the *effectiveness* of a distance metric at an intersection $I \in \mathcal{I}$ by comparing the temporal overlap of the outliers between sensors in an intersection within the same window. More precisely, let $X_{s,i}$ be the FPD vector associated to a set of timestamps T_i from W. For example, if W is the window from 8 to 9 on Mondays $X_{s,1}$ could be the FPD vector associated to the first Monday of the year, i.e., `2020-01-06`, $X_{s,2}$ the one associated to the second, i.e., `2020-01-13`, $X_{s,3}$ the one associated to the third, i.e., `2020-01-20`, and so forth. Let also $X_s^{(1)}, X_s^{(2)}, ..., X_s^{(n)}$ be the FPD vectors sorted in decreasing order of their FPD-LOF scores within W and let $\alpha(X_s^{(j)})$ be the date to which $X_s^{(j)}$ is associated. In the previous example, if $X_{s,2}$ had the largest FPD-LOF score in W, $X_{s,3}$ the second largest, and $X_{s,1}$ the third one, then $\alpha(X_s^{(1)})$ =`2020-01-13`, $\alpha(X_s^{(2)})$ =`2020-01-20`, $\alpha(X_s^{(3)})$ =`2020-01-06`. Finally, let $O_{s,h} = \cup_{j=1}^{h}\{\alpha(X_s^{(j)})\}$. We can now define the effectiveness of a given metric on an intersection I with sensors $s_1, s_2, \ldots s_\ell$ for a window W with n FPDs as:[1]

$$e(I) = \frac{1}{n}\sum_{h=1}^{n}\frac{\left|O_{s_1,h} \cap O_{s_2,h} \cap \cdots \cap O_{s_\ell,h}\right|}{h}. \tag{13}$$

[1]The effectiveness e should be parameterized also by the metric and the window. However, for the sake of lighter notation we do not make this explicit. The reference will be clear from the text.

In the running example, if we have another sensor r with $\alpha(X_r^{(1)})$ = `2020-01-20`, $\alpha(X_r^{(2)})$ = `2020-01-13`, $\alpha(X_r^{(3)})$ = `2020-01-06`, then $e(I) = 1/3(0+1+1) = 2/3$. As exposed earlier, Dist_B might be unable to determine a valid FPD-LOF for some sensor at some T_i. In this case, to ensure that the sensors used for calculating e are the same for all metrics, we remove the sensor where Dist_B fails also in the calculations of e for the other metrics. Finally, we determine the *overall effectiveness* of the metric:

$$\bar{e} = \frac{1}{|\mathcal{I}|}\sum_{I\in\mathcal{I}} e(I). \tag{14}$$

4 ANALYSIS

In this section, we set out to assess the FPD outlier detection with the different metrics and to exemplify the insights that the method can generate in a real-life context.

We implemented everything in Python,[2] using Pandas [25] and Numpy [14] for most parts. For the Dist_{W_1} calculations we used the implementation available in the submodule `stats` of SciPy [16].

4.1 Data set

Evaluation of the metrics was performed on data from 26 intersections of The Hague, shown in Fig. 4, collected during the time-frame from January 2018 to April 2020. Each intersection contains from 7 to 21 relevant sensors, thus we considered in total 272 sensors.

In addition, we had data on tweets about traffic accidents and events in and around The Hague. For the accidents the data available were a timestamp of the occurrence and a brief textual description of the location and severity. The severity of the accident was indicated by a priority of 1 to 5. A priority of 1 indicates personal injury, a priority of 2-3 indicates injury to property, a priority of 4-5 indicates events such as a demonstration, a water leak, etc. For the events we had information about the type of event, e.g., 'large barbeque', 'outdoor concert', etc., the start and end date of the event, and its approximate latitude and longitude. Finally, we had data on holidays. For each date from January 2018 to August 2019 we had indication on whether it was a 'school holiday' or 'other holiday'.

We focused on the window from 08:00 to 09:00 on Mondays. Therefore, each of the 272 sensors yielded up to 113 FPDs corresponding to the 113 Mondays occurring from January 2018 to April 2020. Each FPD was derived from 12 measurements, one every 5 minutes. In fact, to avoid issues with missing measurements, we considered only the FPDs where at least one of the measurements was a value strictly larger than zero. We chose this particular time and day because of the usually high traffic intensity, thus only few zeros are present in the data and only few FPDs had to be discarded. Each sensor had finally after removal still at least 100 FPDs.

4.2 Maximal distances

For each distance metric introduced earlier, we examined 1,993,942 distances. For only one metric, Dist_{W_1}, we do not know a theoretical upper bound on the distance between two FPDs. Hence, we cannot say if it ever reached its upper bound. For the others we observed the following. The calculated Dist_E were never maximal, i.e., equal to two. Dist_B, Dist_H and Dist_{KL} were maximal 13066 of the times, i.e., the distance was infinite, 1 or $\ln(2)$, respectively.

[2]Source code is available at: https://anonymous.4open.science/r/2243

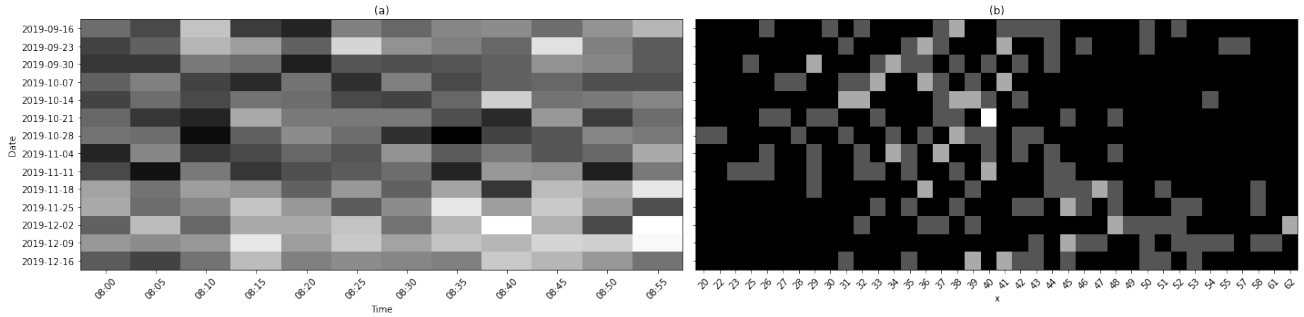

Figure 2: (a) Measurements in a window spanning 08:00 to 09:00 on Mondays in the period 2019-09-16 to 2019-03-16. White indicates the highest traffic intensity in the window and black indicates a traffic intensity of 0, i.e., no cars passed the sensor at the given point in time. (b) The corresponding FPDs of (a) where the four shades ranging from black to white represent $FPD(x)$ of 0, $1/12$, $2/12$, and $3/12$, respectively.

Proportionally, this corresponds to 0.6% of the distance pairs examined for a metric. The maximal distances were distributed among 72 sensors. On these sensors the maximal distance was achieved between 43 and 129 (with median of 56) times in the calculations for all pairwise distances between FPDs associated to a sensor. Note that in the LOF calculations, having infinite distances can result in LOF scores that are infinite, and hence make it impossible to compare the outlierness of points. It can also result in undefined LOF scores (Eq. 12) when a point X has $\text{lrd}_k(X) = \infty$ and another point $O \in N_k(X)$ also has $\text{lrd}_k(O) = \infty$. When calculating FPD-LOF, we looked at neighborhood sizes, k, up to 35. Thus, we could expect a high number of maximal distances occurring in an affected sensor would impact FPD-LOFs. Indeed, for FPD-LOFs calculated using Dist_B approximately 0.5% FPD-LOFs were infinite, which is a clear indication against the use of Dist_B in the context of LOF. The effect on FPD-LOFs of hitting the maximal distance with Dist_{KL} and Dist_H is harder to determine, since the maximum value is finite and hence ultimately absorbed in the calculations with other values, but it still might negatively influence the results with a lack of sensitivity.

4.3 Variance of FPD-LOF scores

When using the LOF algorithm, two approaches for determining whether a point is an outlier are commonly used. One is based on a "pollution" parameter, that is, a fixed number of points to consider as outliers given by the highest LOF scores. The other approach is based on a fixed threshold ϵ for the LOF score, and any point with a score above ϵ is considered outlier. Usually, inliers have a score of around 1, while higher values indicate the degree of outlierness for the corresponding FPDs. However, for determining a threshold we need to consider the variation of the FPD-LOF scores from Eq. (12). As we saw with Dist_B the FPD-LOF scores turn out sometimes infinite consequently making variance infinite. Beside this peculiar case with Dist_B, we observed the highest variance between scores based on Dist_{W_1}. Usually, a high variance is beneficial because it makes it easier to find a suitable value for ϵ, since it is not as sensitive as it is for LOF scores with low variance. Thus—as we will see later—finding a good value for the threshold on this metric is not a problem.

4.4 Effectiveness

In Fig. 3, we show the effectiveness $e(I)$ for each metric on the 26 intersections. The overall effectiveness $\bar{e}$ is reported in Tab. 2.

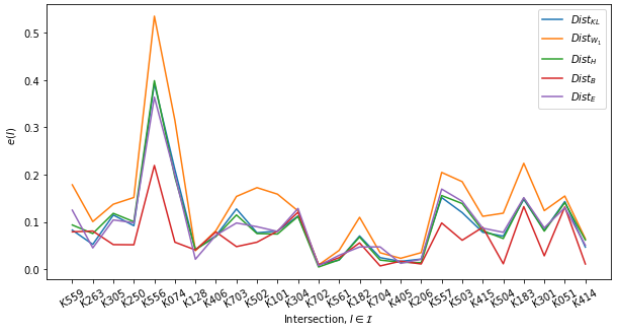

Figure 3: Effectiveness $e(I)$ of the metrics for all intersections $I \in \mathcal{I}$.

Higher values of $e(I)$ and $\bar{e}$ indicate better performance. Hence, Dist_{W_1} is more effective in nearly all intersections and overall.

Trying to explain the difference in effectiveness among the intersections evident in Fig. 3, we studied the impact of the length of the FPD vectors and of the number of sensors in the calculations of $e(I)$. A correlation analysis throughout the different intersections indicated that: i) there is a slight positive impact on the effectiveness of the metrics when the FPDs have more values; ii) there is a considerable negative impact when the number of sensors in an intersection increase. Clearly, with more sensors it becomes more difficult to have large set intersections in Eq. (13).

4.5 Computational cost

In Tab. 2, we report also the computation times to calculate the overall effectiveness $\bar{e}$ from Eq. (13) on a system with 8 × Intel Core i7-4770 CPU @ 3.40GHz processor and 16 GB of RAM. We see that the computation times vary considerably with the metrics and that Dist_{W_1} leads to the fastest computations, up to half of the time of Dist_{KL}.

4.6 Co-occurrence with events

The number of times outliers co-occur with holidays, $\rho_{\mathcal{H}}$, and with tweets on traffic accidents, $\rho_{\mathcal{A}}$, is reported in Tab. 2. More precisely, let us say that an intersection $I \in \mathcal{I}$ has an outlier at a given date, if any of its sensors has an outlier with respect to a fixed threshold, ϵ, for FPD-LOF score of at least 1.3. Let $D(I)$ denote the set of dates where at least one sensor of intersection I has an outlier, and let $\mathcal{X}$ denote either the set of dates

Table 2: Computation times (user CPU time) for calculating $\bar{P}$, i.e., the performance averaged over all 26 intersections. In total, 272 sensors each with at least 100 FPDs have been processed. $\rho_{\mathcal{H}}$ and $\rho_{\mathcal{A}}$ are number of outliers co-occurred with holidays and tweet events, respectively.

Metric	$\bar{e}$	*Time*	$\rho_{\mathcal{H}}$	$\rho_{\mathcal{A}}$
Dist_B	0.0636	63 min 02 s	318	195
Dist_{KL}	0.0945	88 min 26 s	244	109
Dist_H	0.0956	58 min 46 s	99	14
Dist_E	0.0957	59 min 41 s	488	206
Dist_{W_1}	0.1380	46 min 35 s	594	282

Figure 4: Locations of a subset of traffic intersections in The Hague.

that are holidays, $\mathcal{H}$, or that are accidents announced with a tweet within 08:00 and 09:00 on Mondays, $\mathcal{A}$. We then define the number of outliers co-occurring with one of such events as $\rho_{\mathcal{X}} = \sum_{I \in \mathcal{I}} |D(I) \cap \mathcal{X}|$. Note that for $\mathcal{A}$ we do not take into account the location with respect to the intersections, assuming that it may affect anywhere in the network.

It is difficult to interpret these values. The value $\rho_{\mathcal{H}}$ tells how many outliers occur in a holiday and we conjecture that on a Monday, which is a holiday, the traffic flow between 8 and 9 will be low. So $\rho_{\mathcal{H}}$ should indicate the ability of the approach with the given metric to detect as outlier FPDs of exceptional low flow intensity. On the contrary, the value of $\rho_{\mathcal{A}}$ should indicate the ability of the approach to detect outlier FPDs with exceptional high flow intensity, as it is expected that in the case of accidents traffic jams arise. Thus, we can interpret the results as indicating that Dist_{W_1} can find most of these outliers and that it is the best at discovering outliers of both types. However, we desist from making an assessment on the values of $\rho_{\mathcal{H}}$ and $\rho_{\mathcal{A}}$ in absolute terms since without a ground truth it is not possible to give a reliable estimate about how many outliers we should have found. If 10 out of 113 Mondays in our data are holidays we should have approximately $10 \cdot 272$ outliers detected but not all those might be outliers. On the other side, classifying everything as outliers would cover all $10 \cdot 272$ but would not be very useful.

4.7 Comparison with kNN-FPD

In [9], Djenouri et al. used an adaptation of the kNN algorithm to find FPD-outliers. They show that kNN-FPD with Dist_{KL} outperforms in terms of percentage of detected outliers FPD-LOF with Dist_B. Here, we compare their kNN-FPD against FPD-LOF with Dist_{W_1}, having determined that Dist_{W_1} is both more effective and more efficient than all other metrics in an FPD-LOF approach.

Since we are interested in outliers caused by unusually high traffic intensity instead of those caused by noisy readings or sensor dropouts, we constructed a synthetic data set where inlier-FPDs had traffic flow values sampled from a uniform distribution $\mathcal{U}\{500, 600\}$ and outlier-FPDs had values sampled from $\mathcal{U}\{550, 650\}$. We let 10% of the 100 FPDs be outliers and the rest, inliers. Each FPD had a sample size of 10.

The kNN-FPD with Dist_{KL} classifies an FPD as an outlier if the distance to its kth nearest neighbor is above a specified threshold ϵ. For a fair comparison, we modify our LOF approach for FPD such that it also depends on k, that is, stopping the calculations at Eq. (11) rather than Eq. (12). Thus, in the FPD-LOF with Dist_{W_1}, k represents the neighborhood size in the LOF_k calculations and the method classifies as outlier an FPD if its LOF score from Eq. (11) is above a threshold ϵ. The AUC-ROC curves on the synthetic data for the kNN-FPD and FPD-LOF methods are shown in Fig. 5.

First, we observe that FPD-LOF yields a better precision: the highest ROC-AUC score achieved by kNN-FPD is 0.972 while FPD-LOF achieves ROC-AUC scores of 1.0 with the right parameters. Even for the worst performing ϵ value, FPD-LOF reaches a maximal ROC-AUC score of 0.967. Second, FPD-LOF gives a stable classification of outliers for a wide range of ϵ-values when k is above 17. Thus, the method is rather robust with respect to the parameters and tuning them is relatively easy compared to kNN-FPD, where the two parameters k and ϵ are highly interdependent and good performance is achieved only in narrow space.

Indeed, calibrating kNN-FPD seem an intricate task. Assume we have a set of points where no distance between two points in the set is similar. When we increase k in the kNN-FPD algorithm, then the distance to the kth nearest neighbor will increase correspondingly. Hence it follows that ϵ must also be increased to correctly classify points. Thus the two parameters needs to be balanced with respect to each other. In the synthetic data used for Fig. 5, outliers are high-intensity traffic flows and therefore have some homogeneousness, i.e., their mutual distances are less than their distance to the inliers. When this is the case, k should be set such that it is at least as high as the number of expected outliers. Otherwise, outliers would be classified as inliers. On the other hand, if k is too high, then the distance to the kNN for inliers increases, which would result in classifying inliers as outliers. The choice of ϵ depends on which distance metric is being used. Also, ϵ needs to reflect the minimum distance we would expect between an outlier and an inlier. Thus, we need to have a deep understanding of the distance metric used. Using $Dist_{KL}$ somewhat alleviates this problem since it is bound to the interval $[0; 1]$. However, finding the exact value that separates outliers from inliers is still hard. These shortcomings do not seem to appear in the FPD-LOF approach that seems more robust to the choice of the distance metric and the values of k. Moreover, it can be easily made even more robust with the use of Eq. (12).

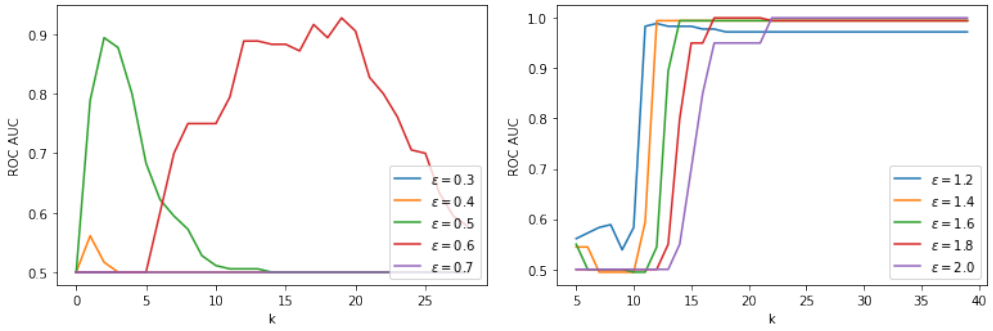

Figure 5: AUC-ROC performance of kNN-FPD (left) and FPD-LOF (right) for various values of k and ϵ.

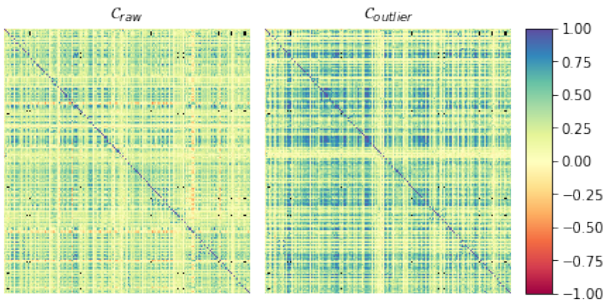

Figure 6: C_{raw} shows the correlation coefficients based on traffic flow intensities and $C_{outlier}$ shows those based on FPD-LOF. 279 sensors were used so sensor names are omitted to avoid cluttering the axes. However, sensors are grouped along the axes based on the intersection they belong to. The tiny black dots signify that two sensors do not have any valid measurements in common, so no correlation could be determined.

4.8 FPD-LOF based correlation analysis

We use FPD-LOF with Dist_{W_1} to determine structural dependencies in the road network. In other terms, we wish to gain insights of the kind: if we detect a traffic jam in intersection A, then there will also soon be a jam arising in intersection B. We describe how to carry out this analysis with the FPD-LOF for outliers approach, and we show that the result can be quite different from what would be otherwise achieved by a more traditional analysis based on raw traffic intensity data.

We applied the FPD-LOF analysis of Eq. (12) to 24×7 windows, that is, every hour of every weekday. We required that all measurements of an FPD must be valid and excluded sensors where one or more windows had less than 100 valid FPDs. Thus, the resulting data is a table with rows indexed by timestamps and columns indexed by sensor names. The cells in the table are the FPD-LOF values. Finally, we calculated correlations between the sensors. For comparison, we calculated the correlation of raw data between the sensors. Since we only used a subset of the data for the FPD-LOF correlation, we extracted the raw data corresponding to the timestamps and sensors previously used.

We show the obtained correlation matrices for the FPD-LOF approach and for the raw data approach denoted by $C_{outlier}$ and C_{raw}, respectively, in Fig. 6. Generally, $C_{outlier}$ gives stronger correlations than C_{raw}.

Focusing on the rankings of the correlations relative to their approach, we found, in general, that C_{raw} yielded higher correlations for sensors in proximity of each others, whereas $C_{outlier}$ was able to find strong correlations also among sensors not in immediate proximity of each other. This observation might not be immediately apparent from Fig. 6, but it was revealed by examining more in detail the locations of the sensors. For example, a pair of sensors from different intersections having a high ranking under both approaches, is made of the sensors both named '081' in intersections 'K556' and 'K559'. The two sensors turn out to be both located in the street Lozerlaan (see Fig. 4) in the trait connecting the same lane between the two intersections, just 2 minutes of travel apart. Thus, the strong correlation for this pair is not surprising. However, $C_{outlier}$ shows that sensor '081' of intersection 'K556' is also strongly correlated with sensor '111' of intersection 'K702'. These intersections are 15-20 minute apart and hence the finding can be more instructive.

Correlations based on FPD-LOFs and those based on traffic intensity measures have evident semantic differences and the calculated correlation coefficients underline this difference. Using the FPD-LOF approach, outliers may propagate from one sensor to another within an hour since we use hourly measurements when creating the FPDs. In such a case, the correlation between the sensors will be higher. A correlation of similar causation will not occur when using raw data to determine sensor correlation. First and foremost, there is no notion of outliers, and secondly, the correlation leaves no room for temporal displacement since correlation is determined laterally. Conclusively, the FPD-LOF correlation provides a unique insight into traffic flow patterns.

5 CONCLUSION

5.1 Summary

We compared distance functions Dist_B, Dist_H, Dist_E, Dist_{KL} and Dist_{W_1} in the context of FPD outlier detection. We defined a measure for the effectiveness of these metrics and found that Dist_{W_1} is the most effective and also the most efficient to use in large data sets. Our measure of effectiveness is biased by the number of sensors in each intersection and by the average length of the FPDs vectors but the influence seems to be the same on all metrics thus not affecting the conclusion.

Assuming holidays would cause abnormal decrease of traffic flows in The Hague on Mondays between 8 and 9, and that traffic accidents would cause an abnormal increase of traffic flows, we observed that LOF with Dist_{W_1} was able to find these different types of FPD outliers more often than with other metrics evaluated.

The interpretability of the distances given by Dist_{W_1} is a further factor of appeal, since it provides a way of explaining the changes in the distribution of traffic flow. On the contrary we highlighted the shortcomings of metrics with upper bounds on the distance values. Specifically, we estimated that the upper bound on these metrics influenced 0.5% of the measurements in our real-life data. We also argued that Dist_E biases distances by the individual variances of the FPDs rather than their variance relative to each other, thus making also this distance less appealing.

We also compared a LOF approach to a kNN approach for determining FPD outliers. The results showed that FPD-LOF using $Dist_{W_1}$ outperforms kNN-FPD in terms of precisely classifying outliers generated by higher intensity flows. Furthermore, FPD-LOF is robust to parameter tuning providing best results for $k > 17$ and $\epsilon \in [1.2; 1.8)$.

Finally, we showed the usefulness of FPD-LOFs for determining dependencies between sensors in the road network. In particular, the FPD-based outlier approach turns out to provide more instructive insights than a more naive approach of correlating raw traffic intensity data.

5.2 Further Research

Having established the utility of Earth Mover's distance in the context of determining flow distribution outliers and correlating them with tweets on traffic incidents,

It would be interesting examining the location of the outliers and their correlated accidents. Moreover it would be interesting enlarging the study to other windows, i.e., beside the one occurring between 08:00 and 09:00 on Mondays.

It is very possible that the kNN approach for outlier detection of [9] can be improved by replacing the Kullback-Leibler divergence with the Earth Mover's distance. The apparent drawback of the kNN approach to outlier detection persists, though. However, our results indicate the approach will produce more stable results.

Outliers found using FPD-LOF with $Dist_{W_1}$ could possibly enhance the performance of recurrent neural networks for predicting traffic flow outliers even further than what has been shown in [13]. In the FPD-LOF-based correlation analysis, we have restricted ourselves to correlate outlier FPDs appearing within the same window. It would be interesting to add also adjacent windows since this might discover dependencies temporally and spatially farther apart. Having discovered spatial and temporal dependencies through the FPD-LOF-based correlation analysis, we could test these insights in predictive tasks. However, it might first be instructive to look deeper into the correlation analysis proposed. For example, we need to assess whether the high correlations reported are meaningful or caused by many inliers whose LOF values are close to 1.0. In this regard, an alternative approach to calculating correlation would be interesting. Specifically, transforming the LOF scores into binary values indicating whether an FPD is an outlier or not.

Finally, we should investigate the effect of the amount of data on the outlier analysis. Too little data might make the detection too sensitive while too much data might make outliers less anomalous and hence not detectable.

REFERENCES

[1] Nawaf O. Alsrehin, Ahmad F. Klaib, and Aws Magableh. 2019. Intelligent Transportation and Control Systems Using Data Mining and Machine Learning Techniques: A Comprehensive Study. *IEEE Access* 7 (2019), 49830–49857.

[2] Kęstutis Balsys, Algimantas Valinevičius, and Danielius Eidukas. 2010. Traffic Flow Detection and Forecasting. *Electronics and electrical engineering* 5 (2010), 91–94.

[3] Vic Barnett and Toby Lewis. 1994. *Outliers in Statistical Data* (3rd ed.). John Wiley & Sons.

[4] Eric Bazan, Petr Dokládal, and Eva Dokladalova. 2019. Quantitative Analysis of Similarity Measures of Distributions. In *British Machine Vision Conference 2019, BMVC 2019*. Cardiff, United Kingdom.

[5] Anil K. Bhattacharyya. 1943. On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions. *Bulletin of the Calcutta Mathematical Society* 35 (1943), 99–109.

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

[7] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD '00)*. Association for Computing Machinery, New York, NY, USA, 93–104.

[8] Andrzej Czyzewski, Andrzej Sroczynski, Tomasz Smialkowski, Piotr Hoffmann, Sebastian Cygert, Grzegorz Szwoch, Jozef Kotus, Dawid Weber, Maciej Szczodrak, Damian Koszewski, Kazimierz Jamroz, and Wojciech Kustra. 2019. Comparative study on the effectiveness of various types of road traffic intensity detectors. In *Comparative study on the effectiveness of various types of road traffic intensity detectors*. IEEE, 1–7.

[9] Youcef Djenouri, Asma Belhadi, Jerry Chun-Wei Lin, and Alberto Cano. 2019. Adapted K-Nearest Neighbors for Detecting Anomalies on Spatio–Temporal Traffic Flow. *IEEE Access* 7 (2019), 10015–10027.

[10] Youcef Djenouri and Arthur Zimek. 2018. Outlier Detection in Urban Traffic Data. In *WIMS*. ACM, 3:1–3:12.

[11] Youcef Djenouri, Arthur Zimek, and Marco Chiarandini. 2018. Outlier Detection in Urban Traffic Flow Distributions. In *2018 IEEE International Conference on Data Mining (ICDM)*. 935–940.

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

[13] Wesley Fitters, Alfredo Cuzzocrea, and Marwan Hassani. 2021. Enhancing LSTM Prediction of Vehicle Traffic Flow Data via Outlier Correlations. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. 210–217.

[14] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362.

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

[16] Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python.

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

[18] Elizaveta Levina and Peter Bickel. 2001. The Earth Mover's distance is the Mallows distance: some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision (ICCV)*. 251–256.

[19] Alexander Novikov, Anastasiya Shevtsova, and V. Vlasov. 2012. Definition of perspective scheme of organization of traffic using methods of forecasting and modeling. *IOP Conference Series: Materials Science and Engineering* 327 (2012).

[20] Jan Puzicha, Joachim M. Buhmann, Yossi Rubner, and Carlo Tomasi. 1999. Empirical evaluation of dissimilarity measures for color and texture. In *Computer Vision and Image Understanding - CVIU*, Vol. 2. 1165 – 1172 vol.2.

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

[22] Bharti Sharma, Vinod Kumar Katiyar, and Arvind Kumar Gupta. 2014. Fuzzy Logic Model for the Prediction of Traffic Volume in Week Days. *International Journal of Computer Applications* 107, 17 (2014).

[23] Grzegorz Sierpiński. 2011. Travel Behaviour and Alternative Modes of Transportation. *Mikulski J. (eds) Modern Transport Telematics. TST 2011. Communications in Computer and Information Science* 239 (2011).

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

[25] The pandas development team. 2020. *pandas-dev/pandas: Pandas*.

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

[27] Leonidas J. Guibas Yossi Rubner, Carlo Tomasi. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40 (2000), 99–121.

[28] Arthur Zimek and Peter Filzmoser. 2018. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 8, 6 (2018).