Measure for Traffic Anomaly Detection on the Urban Roads Using Speed Transition Matrices

L. Tišljarić*, Ž. Majstorović*, T. Erdelić*, T. Carić*
* Faculty of Transport and Traffic Sciences, University of Zagreb, Vukelićeva 4, 10000 Zagreb, Croatia
ltisljari@fpz.unizg.hr

Abstract – Road traffic anomaly detection is an essential research topic within the Intelligent Transport System (ITS) context. Urban road anomaly detection systems are a crucial part of the ITS regarding the trip planning, road security, and congestion estimation applications. In this paper, the method for traffic anomaly detection using Speed Transition Matrices (STM) is presented. The paper's main goal is to present the novel method for measuring the distance between two STM, as standard distance measures are inapplicable for the anomaly detection and road traffic analysis interpretation. The method is based on the Euclidean distance measure between STM's Center of Mass (COM), and the average STM that represents normal traffic conditions. The Global Navigation Satellite System (GNSS) data on the road-network of City of Zagreb were used as a case study, as it is, the capital and the largest city in Croatia, suitable for the application of the proposed methodology. The anomaly detection method resulted in 73 anomalous points which are presented on the digital map. The proposed method is compared to the other distance metrics used in the literature, and advantages over each of the metrics are highlighted.

Keywords – Anomaly detection; Speed transition matrix; Quantifying anomaly; Intelligent transport systems; Traffic state estimation; Urban road network

I. INTRODUCTION

As a process of detecting potentially dangerous events, anomaly detection is a crucial part of the ITS [1]. The importance of anomaly detection lies in the potential useful actionable information for traffic management systems. Anomalous traffic in an urban road network could indicate a severe traffic accident, traffic congestion, or a violation of the regulations.

The focus of this research is on detecting the recurrent anomalies on the urban road networks. In [2], authors report that recurrent anomalous traffic is covering around 85% of all congestion on the urban road networks. Although, the algorithms for anomaly detection in other research areas have reached high efficiency and performance, the development of traffic anomaly detection methods in urban areas is still in the early stages. The authors in [3] present the review on the traffic anomaly detection and report that this is an early-stage research area, and further research is required in every direction.

In this paper, a method for anomaly detection based on the STM is proposed. The methodology consists of the three main steps: (i) data preprocessing and STM generation, (ii) COM estimation, and (iii) measure of distance between STMs. For the STM generation, a large real-life GNSS data are used. Anomaly detection is based on the measure of distance between the COMs of the observed STMs and the median STM, which represents normal traffic conditions.

Contributions of this paper are as follows: (i) novel methodology for the anomaly detection based on the COM measurement and (ii) results of the anomaly detection are analyzed on the urban road network in the City of Zagreb, Croatia.

The rest of the paper is organized as follows. Section 2 presents the review on the literature regarding the anomaly detection method, and data used to model the traffic parameters. Section 3 presents the methodology for the anomaly detection based on the measure of distance between STMs’ COMs. Section 4 presents the case study results on the urban road network in the City of Zagreb. Alongside the anomaly detection results, a comparison between standard and most used distance metrics is conducted. In section 5, the future work, and the conclusion is given.

II. RELATED WORK

In [4], the authors developed a framework for detecting multiple public events which were differentiated from regular crowd moving patterns. The method is based on the tensor factorization where crowd movement data, collected from the bike trips in a bike-sharing system, is augmented with semantic social activity data. In [5], the authors relied on multiple data sources to demonstrate the advantages of the proposed method to detect anomalies with unknown external influences, and to detect an anomaly before it reaches its peak. The tensor decomposition methods [6] are used by authors in [7] and [8] for traffic data modelling, which contained spatial and temporal data, and the Local Outlier Factor algorithm, to successfully detect the anomalies, and to determine the time interval of the anomaly for the certain area.

The authors in [9] based their research on a data collected from the traffic loop sensors and a dictionary-based compression theory for analyzing the traffic data to identify the features of the spatial and temporal patterns. The proposed method resulted with clear geographically distributed features of spatial and temporal traffic flow pattern, represented as a heatmap, where anomalies can be easily identified. Taxi vehicles are also a valuable source of the GPS data considering the nature of their activities. In [10] authors used Taxi GPS data as traffic flow data to build traffic flow matrix and proposed anomaly detection method based on the wavelet and principal component analysis for
detecting traffic anomalies in urban subregions. The proposed method proved to be effective in discovering anomalies in the two adjacent regions as well as identifying traffic flows that cause the traffic anomalies. The authors in [11] also used GPS data collected from taxi vehicles and proposed deviation-based outlier detection method, which focuses on the road segments instead of paths. The proposed method resulted in detection of a large number of events (anomalies), where each event was associated with the road segment and time interval when the event happened.

Simple features such as speed or distance can be used as the anomaly indicator as well. Based on the physical distance between the origin and the destination, the authors in [12] developed a method which long taxi route marks as an anomalous event. The authors in [13] used the vehicle speed recorded by the mobile device sensors to detect speed bumps and road damages, which turn out to represent anomalies in urban roads. In [14], the authors detected detect traffic jams based on the estimated free-flow speed and actual speed for each road segment.

III. METHODOLOGY

Given the set of precomputed STMs, this paper aims to obtain a method for traffic anomaly detection on the urban road network. The proposed methodology is presented in Figure 1. This section briefly describes the three main steps of the method that includes data preprocessing, STMs generation, and anomaly detection based on the COM estimation.

A. Speed Transition Matrix

In most of the literature, the traffic data are represented by using a time series vector [15]. In some recent papers, the traffic data are represented as images by using a two-dimensional matrix that represents the spatial and temporal characteristics of the traffic flow [16], [17].

In this paper, the STM is used for the representation of traffic data. The STM is a matrix that represents a probability of changing the speed value when a vehicle travels between two consecutive road segments. Similar data modeling was used in [18] for predicting the traffic density by using the GNSS data. The change of speed between two consecutive road network segments is called a transition. The transition can be defined as a change in vehicle trajectory when traveling between two consecutive road segments \( r_i \) and \( r_j \) in the time interval \( t \). Then, the average speed on the input road segment \( r_i \) is labeled as source speed \( v_{s} \), and the average speed on the output segment \( r_j \) is labeled as destination speed \( v_{d} \).

Two examples of the transition are visually presented in Figure 2a). The first transition can be presented with the vehicle traveling between road segments \( a \) and \( d \) (red), and the second transition with the vehicle traveling between the road segments \( c \) and \( a \) (blue). Source speeds \( v_{s} \) are average speeds on the edges \( a \) and \( c \), and destination speeds \( v_{d} \) are average speeds on the road segments \( d \) and \( a \) for the corresponding transitions. The STM can be constructed by counting the \( v_{s} \) and \( v_{d} \) for the time interval under observation. Then, the speed counts are transformed into the speed transition probability distribution to get the probabilities for every transition. Figure 2b) and c) show the examples of the STMs that represent anomalous and normal traffic conditions. On the x-axis are the destination speed, and on the y-axis is the source speed, with the discretization period of 5 km/h. Based on the speed count, the probability matrix \( X(t) \) is computed. The dimensions of the matrix depend on the chosen resolutions (sensitivity).

Figure 1. Methodology for the anomaly detection using STMs

Figure 2. Examples for two types of transition of the STMs containing the anomaly and the normal traffic flow
of the speed change and the maximal speed that can be captured:

\[
X(t) = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1n} \\
p_{21} & \ddots & & \vdots \\
\vdots & & \ddots & \vdots \\
p_{m1} & \cdots & & p_{mn}
\end{pmatrix}
\]

(1)

where \( p_{ij} \) presents the probability of the observed transition that vehicle had source speed \( v_s \) and destination speed \( v_d \) in the time interval \( t \). For this paper, 5 km/h is the chosen discretization period and 100 km/h for the maximal possible speed, which resulted in matrix dimensions \( 20 \times 20 \). The specific maximal speed value is chosen because case study is conducted on the road segments with a speed limit between 50 and 80 km/h.

B. Quantifying anomaly

Determining the COM is a key factor for detecting anomalous road segments. It can be observed that the position of the grouped data in the STMs indicates the traffic conditions on the observed transitions. If grouped data are positioned further from the center of the matrix, an anomalous event can be detected. For example, if the data are grouped in the far upper left corner, the STM shows that the observed speed transitions are extremely low, which can indicate an anomaly (Figure 2b). Similar conclusions can be given if the data are grouped in the other corners of the STM.

For COM estimation, a method based on the determination of the expected values is selected [19]. First, the marginal distributions for the \( x \) and \( y \) coordinates of the STM are calculated. Then, the coordinates of the COM are calculated as follows:

\[
c_x = \sum_{i=1}^{20} p_x(x_i) \cdot i
\]

(2)

\[
c_y = \sum_{j=1}^{20} p_y(y_j) \cdot j
\]

(3)

where \( p_x(x_i) \) represents the value of the marginal distribution of the \( x \) coordinate, \( i \) represents the index of the \( x \) coordinate, \( p_y(y_j) \) represents the value of the marginal distribution of the \( y \) coordinate, \( j \) represents the index of the \( y \) coordinate. Applying the equations (2) and (3) on the STMs gives the result as a coordinate of COM denoted as \( c_x \) and \( c_y \). The examples of the COM estimation are shown in Figure 3 represented as red and green dots.

Figure 3a) shows an example of STM with determined COM, and figure 3b) shows the results of the COM estimation on the median STM. The median STM is used as the representation of the normal traffic on the observed transitions. The median matrix is characterized by the speeds grouped diagonally from the center of the matrix to the lower right corner, which indicates high-speed values close to the speed limit.

With known coordinates of COM, Euclidean distance can be calculated between the COM coordinates of the STM under observation and the COM coordinates of the median STM:

\[
d(\mathbf{c}_{\text{STM}(i)}, \mathbf{c}_M) = \sqrt{(c_{x2} - c_{x1})^2 + (c_{y2} - c_{y1})^2}
\]

(4)

where \( c_{x1} \) and \( c_{y1} \) represent the COM coordinates of the STM under observation denoted as \( \mathbf{c}_{\text{STM}(i)} \) and \( c_{x2} \) and \( c_{y2} \) represents the COM coordinates of the median STM denoted as \( \mathbf{c}_M \). To get more interpretable measure of the anomaly level, Euclidean distance is expressed as the relative distance to the longest distance in a matrix (diagonal):

\[
d_{rel} = \frac{d(\mathbf{c}_{\text{STM}(i)}, \mathbf{c}_M)}{200\sqrt{2}}
\]

(5)

where \( d_{rel} \) is the relative Euclidean distance. The Euclidean distance is divided by \( 200\sqrt{2} \) because STM has dimensions \( 20 \times 20 \), and its diagonal can be calculated by Pythagorean theorem. This method gives the normalized results in interval \([0, 1]\), and the value quantifies the traffic anomaly. If the result is closer to 1, the speed data are grouped on the edges of the STM, which indicates a significant anomaly. This can indicate significantly the lower or higher speeds on the source and destination road segments represented by the STM in the traffic context. The result closer to zero indicates that data are grouped closer to the COM of the median STM, which represents the normal traffic conditions in the source and destination road segments. This indicates that the traffic flow has the speed closer to the speed limit.

Given the relative Euclidean distance values \( d_{rel} \), we consider the standard statistical classification of the distance as an outlier [20]:

\[
d_{rel} > Q_3 + 1.5 \cdot IQR
\]

(6)
where Q3 represents the value of the $d_{rel}$ that is on the 75% of the sorted data, and the IQR represents the interquartile range of the data.

IV. RESULTS

A. Data

The GNSS data used for the case study are acquired from the vehicles equipped with the tracking devices. The data are provided by the company Mireo Inc. as a part of the SORDITO project [15, 21]. Geographical longitude and latitude, speed, timestamp, and heading are included in every record. Most of the data are sampled with the sampling rate of 100 m for vehicles in driving mode and every 5 min for turned off vehicles. GNSS records are map-matched to the road segments in a digital map of Croatia. The data were recorded in a five-year period between August 2009 and October 2014 by approximately 4200 vehicles, which were versatile, mostly consisting of taxi cars and delivery vehicles (vans and caddies). The total number of records was around 6.55 billion. The data were analyzed, and anomalies were detected using the proposed method for a large road segments in the City of Zagreb. The seasonality of the traffic flow is considered to lower the deviation. Summer months, July and August, are excluded from the case study. They significantly influence the results due to the different, and lower traffic flows caused by vacations. The data from the different, and lower traffic flows caused by vacations [22]. Further filtering is done by segmenting the data on working and weekend days. Traffic conditions during the working days are different from the weekend traffic conditions mostly due to the daily commuters. The data used for the case study include only working days with summer months excluded to extract only the most relevant traffic conditions.

B. Anomaly detection

By applying the proposed method, the 73 anomalous matrices out of total 25,188 STMs were detected. Matrices that were declared anomalous, with the relative Euclidean distance closer to 1 (e.g. $d_{rel} = 0.41$) will have data grouped on the edges of STM visible in Figure 4a). On the contrary, matrices with relative Euclidean distance closer to 0 (e.g. $d_{rel} = 0.03$) will have the data grouped around the center of matrix as shown in Figure 4b).

The visualization of detected anomalies on the map of Zagreb is shown in Figure 5. It is noticeable that the anomalies (the red lines) are detected in different parts of the city. Results can be segmented in two types of anomalies: the first type are anomalies detected on the roads such as city’s access roads, positioned on the edges of the city, and the second type are the anomalies detected on the city streets closer to the city center. The anomalies detected on the city access roads are the result of people’s daily migrations from surrounding cities, and those anomalies occur in peak hours during the morning and the evening rush hour. The second type of anomalies that were detected in the city center is the result of the migration of people within the city. Those anomalies occur not only during the morning and evening rush hour, but also in the other parts of the day. Such an outcome is the result of a high number of people working in the city, a lot of tourists, and availability of wide a variety of entertainment and nightlife.

C. Comparison

In this section, the proposed method for measuring the distance between STMs is compared to other distance metrics. Table 1 presents the results for the anomaly detection using different distance metrics. The proposed relative Euclidean distance between COMs is compared to the Manhattan, Cosine, and Jaccard distances, which are one of the most used distance metrics for the comparison of the matrices. For the comparison, Q1, Q3, IQR, $Q_3 + 1.5 \cdot IQR$ values, and anomaly count are used.

1) Metrics

Manhattan distance is also known as Taxicab geometry. It is called Manhattan because it is the distance that a car or a taxi would drive in a city such as Manhattan, which has straight streets with intersections at right angles and buildings in square blocks formation. If two points A and B are at the intersections on the same street, the distance between them is measured as a straight line (Euclidean distance), and the measure for the distance is expressed by counting the number of blocks. In another case, if two points A and B are not on the same street, then the distance between two points is the number of blocks taxicab must travel to get from A to B using the shortest possible route. In contrary to the Euclidean distance, where the shortest distance is unique, when using the Manhattan
distance, it is possible to get the shortest distance with multiple paths through the traffic network. Manhattan distance can be calculated as [23]:

\[
d_M(STM, M) = \sum_{i=1}^{n} |STM(i) - M(i)|
\]  

where \(d_M\) is Manhattan distance, \(STM(i)\) represents \(i\)-th value of the vectorized STM, \(M(i)\) represents \(i\)-th value of the vectorized values of median STM, and \(n\) is the number of samples.

**Cosine distance** is a distance metric for comparing two datasets where both datasets are represented as a vector with its own direction and magnitude. If two datasets are represented as vectors, the similarity between those two datasets corresponds to an angle between those two vectors, which is quantified as the cosine of the angle between vectors. The Cosine similarity between the two datasets is defined as [24]:

\[
d_C(STM, M) = \frac{STM \cdot M}{\|STM\| \times \|M\|}
\]  

where Cosine distance is denoted as \(d_C\), and \(M\) represents flattened median STM matrix. Result values of Cosine distance are in \([-1, 1]\) range, where value closer to 1 means higher similarity between the compared datasets (vectors with same orientation) and the result with value closer to -1 indicates that compared matrices are opposite (vectors are the same, but with different orientation). Result with value 0 means that the compared matrices are unrelated (orthogonal vectors).

**Jaccard distance** is the distance metric that measures the distance as similarity of two matrices. It is calculated as a ratio between the number of elements of intersected data between two matrices, and a number of union elements of those two matrices.

\[
d_J(STM, M) = \frac{|STM \cap M|}{|STM \cup M|}
\]  

where Jaccard distance is denoted as \(d_J\) with the results in range \([0,1]\). The result closer to 1 indicates that the compared matrices are more similar, and the result closer to 0 indicates that compared matrices are more different.

Table 1 - Comparison of the results for the anomaly detection using different distance measures

<table>
<thead>
<tr>
<th>Measures</th>
<th>(d_{rel})</th>
<th>(d_M)</th>
<th>(d_C)</th>
<th>(d_J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>0.08</td>
<td>6363</td>
<td>0.341</td>
<td>0.9822</td>
</tr>
<tr>
<td>Q3</td>
<td>0.21</td>
<td>6390</td>
<td>0.666</td>
<td>0.9873</td>
</tr>
<tr>
<td>IQR</td>
<td>0.13</td>
<td>27</td>
<td>0.325</td>
<td>0.0051</td>
</tr>
<tr>
<td>Q3+1.5*IQR</td>
<td>0.405</td>
<td>6436.5</td>
<td>1.1535</td>
<td>0.99495</td>
</tr>
<tr>
<td>Anomaly count</td>
<td>73</td>
<td>1339</td>
<td>0</td>
<td>69</td>
</tr>
</tbody>
</table>

Jaccard distance can be calculated as follows [25]:

\[
d_J(STM, M) = \frac{|STM \cap M|}{|STM \cup M|}
\]  

Figure 5. Visualization of detected anomalies in the City of Zagreb

Figure 6. Distributions of the distance measures used for the anomaly detection
D. Results of the anomaly detection

Alongside the mentioned measures, the comparison is conducted by examining the distributions of distance values and it is presented in Figure 6. By examining the distribution of relative Euclidean distances, it can be noticed that most of the results are grouped in the first half the distributions, which is not the case with other distance measuring methods. The other measuring methods have distributions that are dispersed and/or compressed in certain parts.

**Manhattan distance:** Out of 25,188 matrices Manhattan distance detected 1,339 anomalous matrices, which is significantly more than by using relative Euclidean distance method. Considering the distribution of the distance values (Figure 6) it can be concluded that the data are grouped mostly at the beginning of the distribution, which indicates that this measure is not well suited for the anomaly detection based on the STMs. This claim is justified by examining the results in Figure 7. The Manhattan distance for the examples a), and b) are very high, and the anomaly is detected. This is not the correct assumption because on the example a) the speeds are scattered, and on the example b) the destination speeds are high and cannot be classified as the anomaly. On the other hand, relative Euclidean distance for the presented examples is very low, and the anomaly is not detected.

**Cosine distance:** By using this metric, anomalous STMs cannot be detected with proposed approach because the measure $Q3 + 1.5 \cdot IQR$ resulted in $d_c > 1.15$. This value is not suited for anomaly detection because it exceeds the range of $[-1, 1]$. High values for $Q1$, $Q3$ and high $IQR$ value indicate that the result values are dispersed through entire distribution in range $[0,1]$ which is visible in the distribution of Cosine distances shown in Figure 6.

**Jaccard distance:** detected 69 anomalous matrices, which is the result similar to the relative Euclidean distance. Considering the results of statistical processing, $Q1$ and $Q3$ have the values closer to 1 and are very closely spaced with low $IQR$ value. It can be concluded that all the values are grouped in the upper part of the distribution, which is visible in Figure 6. Such grouped result suggests that this method is not appropriate for anomaly detection based on the STMs. The claim is justified by examples of detected anomalous STMs by the Jaccard distance, shown in Figure 8. The Jaccard distances for the examples a) and b) are very high, and the anomaly is detected. Since the speeds are high and scattered on both examples a) and b), they cannot be classified as an anomaly. On the contrary, relative Euclidean distances for the mentioned examples are very low, and anomalies were not detected.

**CONCLUSION**

In this paper, the method for anomaly detection based on the STMs is presented. The method consists of the three mains steps: data preprocessing and STM generation, COM estimation, and distance measurement. As a distance metric, relative Euclidean distance between the COMs of the observed STM and the median STM, which represents the normal traffic conditions, is proposed. The results show that the proposed metric is better suited for the anomaly detection problem based on the STMs because it can use the valuable traffic information, which depends on the position of the grouped data. The method is validated by comparison to other distance metrics from the literature. It showed the best results for the anomaly detection because other distance metrics are not able to use the information about the speeds position in the STM.

Future work will be related to extending the analysis with the temporal components of the anomalies by utilizing the tensor decomposition methods. This type of anomaly detection algorithm will provide more detailed spatiotemporal traffic anomaly information and can be useful for the potentially actionable traffic information, routing applications, or to the road network maintenance authorities.
ACKNOWLEDGMENTS

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATA CROSS). Data used for this research is collected during the SORDITO project, European Regional Development Fund under contract RC.2.2.08-0022. Authors are also very grateful to industrial partner MIREO Inc. and the Faculty of Transport and Traffic Sciences.

REFERENCES


