Anomaly Detection in Diurnal CPS Monitoring Data Using a Local Density Approach

Pratik Narang and Biplab Sikdar
Department of Electrical and Computer Engineering, National University of Singapore, Singapore
Email: {pratik, bsikdar}@nus.edu.sg

***Abstract*—Devices that monitor and measure various system parameters or physical phenomena form an integral part of cyber-physical systems. Such devices usually operate continuously and gather important data that is often critical for the operation of the underlying system. Thus, it becomes important to understand and detect abnormal or malicious device behavior, false injection of data by an adversary, or other security threats that may lead to incorrect measurement data. This paper addresses the problem of detection of anomalies in diurnal traffic volume data in an intelligent transportation system. The proposed approach leverages the statistical properties of the data to perform anomaly detection by calculating the 'local density' of the data points. Anomalous behavior in the traffic volumes reported by road segments is calculated based on sparse local density of the data points. Our approach for detecting anomalies does not require any information about the outside factors which might have influenced the data. The proposed approach has been evaluated on attacks simulated on transportation data collected by the New York State Department of Transportation. The proposed approach also extends to other cyber-physical systems where the monitored data exhibits diurnal patterns.**

I. Introduction

Cyber-physical systems (CPS) are complex systems that combine a cyber system and a physical system. The cyber system adds information collection and processing capabilities that can significantly affect the overall operation of the physical system. The monitoring and real-time control facilitated by the cyber system is fundamental in ensuring the reliable operation of the critical tasks performed by the physical systems. The smart devices part of CPS, enabled with sensors, are increasing in number and becoming more common. Be it intelligent transportation systems, smart grids, or water distribution systems, the ability to track on monitor the system parameters is increasingly becoming a part of such systems. The continuous, and often fine-grained measurements provided such monitoring devices result in large volumes of data. This data provides valuable insights into the operating conditions of the systems and the factors that affect them. Finding anomalies and outliers in these data can help in better understanding of the system's behavior as well as detection of abnormal/malicious device behavior, false injection of data by an adversary, or other such security threats and intrusions.

This paper focuses on the problem of detecting data manipulation in CPS monitoring data. In particular, we focus on anomaly detection in CPS data with diurnal patterns and as a specific scenario of interest, we consider vehicular traffic volume data collected from roadside sensors. Such traffic volume data serves as one of the inputs in Intelligent Transportation Systems (ITS), and may be used to control traffic lights, dynamically regulate speed limits, and determine congestion sensitive road toll charges. Road traffic volume data (and other CPS data such as electricity and water usage) show certain patterns during their normal operation. In particular, the traffic is diurnal, with higher activity during the day and lower activity at night. Further, the traffic volume is generally low during weekends. Office hours on working days tend to exhibit a peak in the traffic volume on certain roads. However, the traffic monitoring devices are expected to see low traffic volume on certain days such as Thanksgiving and Christmas in the USA, which are national holidays, irrespective of the day of the week on which the holiday falls. Similarly, many other factors may contribute to low (or high) traffic volume being recorded on a certain day. The traffic on a particular day and a particular road might be influenced by factors such as road accidents, harsh weather conditions, or a variety of other reasons. Other CPS monitoring devices, such as smart electricity meters, may also see similar variations in the data recorded due to influence of outside factors. For example, high electricity or water usage might be recorded at homes at the time of popular sport championships matches since a lot of people are at home watching those matches.

The examples above illustrate the point that many 'outside' factors may influence the data being gathered at such monitoring devices. Given the broad spectrum of applications of such devices, and the fact that the data may be gathered from a large number of devices situated in different locations or countries, it is not always possible to keep track of these outside factors while analyzing the data for purposes such as trend detection, anomaly detection, etc. This work proposes a novel approach for anomaly detection in such scenarios. Our work deals with detection of anomalies in road traffic volume data by leveraging the statistical aspects of the data itself, and in the absence of any information of 'outside' factors. The datasets used in this work correspond to road traffic volume data gathered by the New York State Department of Transportation (NYSDOT), USA, for the state of New York [1] and its different counties. However, our approach may be applied to other CPS 'monitoring data' that exhibit diurnal patterns.

To research contributions of this paper are as follows:

- We develop an anomaly detection methodology for diurnal traffic volume data, which extends to other diurnal 'monitoring data' gathered by CPS devices.
- We leverage the statistical properties of the data to perform anomaly detection based on the calculation of the 'local density' of the data points. Anomalous behavior in road segments (and not individual moving objects) is calculated based on sparse local density which utilizes the Local Outlier Factor algorithm [2].
- Our approach allows for the detection of anomalies in the data without the need for any information about

the outside factors which might have influenced the data.

The rest of this paper is organized as follow. A brief literature survey of related works is presented in Section II. In Section III, we provide an overview description of the data analysis problem of traffic volume and our anomaly detection approach. Section IV gives the detailed steps involved in the anomaly detection of traffic volume data, which includes extracting relevant metrics from the data and performing local density based anomaly detection using those metrics. Section V presents an evaluation of the proposed approach, while Section VI presents the conclusions and future scope of work.

II. Related Work

Due to the complex nature of cyber-physical systems, developing security solutions for them requires a multi-faceted approach. For the purpose of this work, we restrict our discussion to security threats and attacks that lead to anomalous patterns in data monitoring devices. False injection attacks in power grid systems were demonstrated in [3], wherein the authors demonstrate attack vectors which exploit the configuration of a power system to launch attacks to successfully introduce arbitrary errors into certain state variables, while bypassing existing techniques for bad measurement detection. Their experimental results demonstrate the feasibility of such attacks, and reiterate the necessity of better security approaches in CPS environments.

From the more general perspective of anomaly detection, authors in [4] demonstrate the utility of Pareto or Log-Normal distribution to discover abnormal traffic patterns under the presence of diurnal/seasonal trends in data. The work of authors, however, does not lay much emphasis on night patterns in detecting anomalous patterns.

With respect to anomalies in road traffic volume, authors in [5] study the root cause of road traffic anomalies by modeling road traffic as a time-dependent flow on a network formed by partitioning a city into regions bounded by major roads. Their work models the flow in the network as an origin-destination matrix and adopts a novel approach to identify link anomalies through an optimization problem. However, the two-step approach proposed by the authors is expected to be slow and thus not suitable for real-world applications involving quick computations. Authors in [6] propose a methodology for detecting temporal outliers in vehicle traffic data by considering historical similarity trends between data points. Outliers are calculated from drastic changes in the trends. Although this approach of similarity between data points is more robust, the authors are forced to rely on an expensive graph search in order to compute the temporal outliers. This limits the practical utility of the approach. Work in [7] performs detection and classification of anomalies by combining the spatio-temporal changes in the variability of microscopic traffic variables, such as relative speed, inter-vehicle time gap and lane changing. Their work relies on extracting a number of microscopic variables. It should be noted that, in the data collection process, it is not uncommon for many of these variables to be corrupt or unavailable altogether. This may limit the deployment scenarios of the approach proposed by the authors.

The anomaly detection approach proposed by us is not dependent on a large number of variables or computation of multiple statistical metrics (as in [7]). Neither do we require to solve expensive optimization problems (as in [5]) or perform prohibitive graph searches (as in [6]). Our approach computes simple metrics based on traffic volume data. These metrics are easy to compute, and do not pose any performance bottleneck. We leverage the statistical properties of the data itself, and demonstrate that the local density based metrics are sufficient to provide good accuracy of anomaly detection.

III. System Description

The road traffic volume data obtained from NYSDOT corresponds to ten years of data for different road segments in the state of New York and its counties. The data is available in the form of the tuple: RC_station (road ID), timestamp, direction, lane, count. A sample of traffic volume data, available in the form of 5-tuple, is shown in Table I. Each tuple indicates the traffic volume count on that road segment for a particular direction and lane at a given time. The tuples in the dataset have been captured periodically, which is once in 15 minutes for certain counties such as Rensselaer, and once in an hour for New York state's highway road segments.

TABLE I: Sample of traffic volume data

RC_station	Timestamp	Direction	Lane	Count
140040	09/22/2011 12:00	1	1	105
140040	09/22/2011 12:00	5	1	80
140040	09/22/2011 12:15	1	1	87
140040	09/22/2011 12:15	5	1	86
140040	09/23/2011 01:00	1	1	4
140040	09/23/2011 01:00	5	1	8
140040	09/23/2011 01:15	1	1	8
140040	09/23/2011 01:15	5	1	6

Figure 1 shows the measurement data for one road segment in Rensselaer country (RC station 140040) for a week's worth of data. The two directions of the road are marked 1 and 5. The roads have a single lane, and thus the lane number does not bear any significance. The traffic volume count is plotted on the Y axis against the timestamp on the X axis. The plot clearly highlights the diurnal pattern in the traffic volume data. The variations in traffic volume for day and night times, as well as weekday and weekend (around the middle of the plot) are clearly evident (for direction 1 as well as 5).

Our work addresses the problem of detecting anomalies in the traffic volume data in such a scenario. We do not assume the presence of any information which might influence the traffic volume, be it road accidents, weather conditions, or national holidays. Our approach considers the 'locality' of each data point to compute certain simple metrics which remove the diurnal trend of the data and compute each data-point's position with respect to its locality or neighborhood. Based on these metrics, we calculate the local density of a data point, and identify anomalous points based on sparse density from the neighbors. We also evaluate our approach by injection of anomalous data points in the dataset.

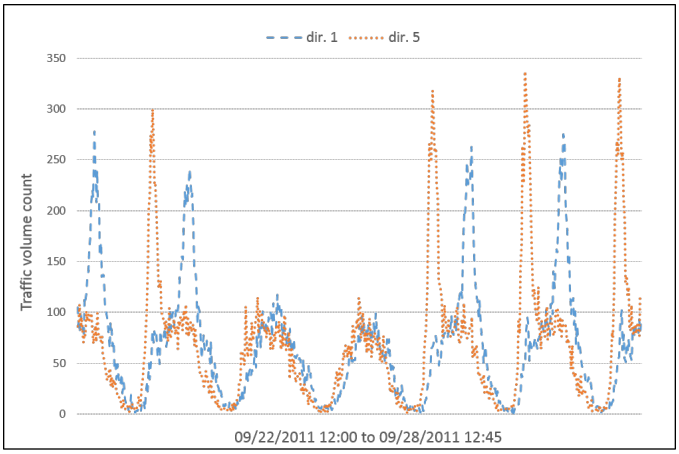


Fig. 1: Diurnal trend in road traffic volume of RC station 140040.

IV. Anomaly Detection for Traffic Volume Data

Intuitively, a simple method for anomaly detection is to compare the traffic volume data against thresholds obtained using historical data. However, a simple threshold on the 'count' of the traffic volume cannot serve to identify anomalies in the traffic data. This is easily understood from the fact that the definition of 'normal' traffic varies at different times of the day and on different days (weekday/weekend). The sample data shown in Table I clearly illustrates this. The traffic count around noon is in the range between 80 and 105, while post-midnight it is between 4 to 8. Since the traffic count in itself is not a suitable indicator of normal or anomalous traffic, we use other metrics computed from the traffic volume data. These metrics provide us with each data-point's position with respect to its neighboring points, and also remove the diurnal trend from the data.

A. Extracting Metrics

Let x_i denote the i-th data point (i.e. the i-th traffic volume measurement). We then perform the following operations on each traffic count data points and calculate the metrics given below. Let the variables p_j denote the difference of the traffic count of x_i from the traffic count of j-th preceding data point ($j = i-1, i-2, \cdots$), divided by the difference in their timestamps. Similarly, let n_k denote the difference of the traffic count of k-th subsequent data point from the traffic count of x_i ($k = i+1, i+2, \cdots$), divided by the difference in their timestamps. These metrics are calculated separately for both directions of each road in the dataset. Lanes are ignored for the sake of simplicity. Thus,

$$\begin{aligned} p_{i-1} &= \frac{x(count)_i - x(count)_{i-1}}{x(timestamp)_i - x(timestamp)_{i-1}} \\ p_{i-2} &= \frac{x(count)_i - x(count)_{i-2}}{x(timestamp)_i - x(timestamp)_{i-2}} \\ &\vdots \end{aligned} \tag{1}$$

$$\begin{aligned} n_{i+1} &= \frac{x(count)_{i+1} - x(count)_i}{x(timestamp)_{i+1} - x(timestamp)_i} \\ n_{i+2} &= \frac{x(count)_{i+2} - x(count)_i}{x(timestamp)_{i+2} - x(timestamp)_i} \\ &\vdots \end{aligned}$$

Note that as the difference in the timestamps grows larger, the values of the variables p_j and n_k becomes insignificant. Thus, the variables with higher values of j and k do not need to be considered for practical purposes.

Obtaining statistical metrics by taking the difference of the traffic data count from its previous and next neighbors serves to remove the diurnal trend from the data. Removal of diurnal patterns was necessary, as highlighted earlier. Our approach of dividing the difference of traffic count by the difference of timestamp allows us to put greater emphasis on the immediate neighbors, while lower emphasis is paid to the neighbors which lie farther away. This was a natural choice because intuitively, data from nearby time instants are expected to be correlated and the correlation decreases as the time difference is increased. Thus for the purposes of anomaly detection, it suffices to only consider recent values.

A sample of these metrics is presented in Table II. Figure 2 shows these metrics for the traffic volume data shown in Figure 1, limited to direction 5. The p_j and n_k variables are shown for $i-1$ and $i-2$, and $i+1$ and $i+2$ only. The two plots are quite similar to each other, which is natural since both the variables are computed by the approach of taking difference from the same single variable of traffic count.

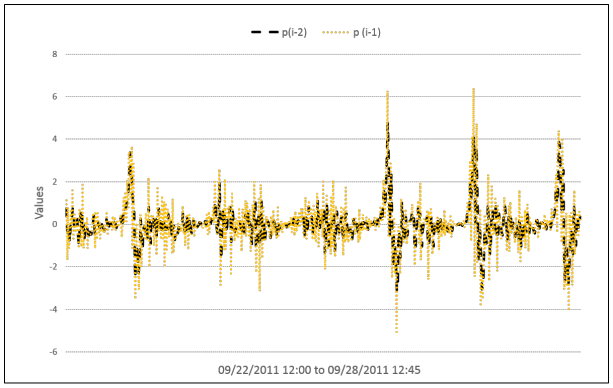
TABLE II: Sample of metrics obtained for RC station 140040, direction 5

Timestamp	p_{i-1}	p_{i-2}	n_{i+1}	n_{i+2}
09/22/2011 13:00	-1.67	0.267	0.8	-0.167
09/22/2011 13:15	0.8	-0.433	-1.133	-0.833
09/22/2011 13:30	-1.33	-1.67	-0.53	0.03
09/22/2011 13:45	-0.53	-0.83	-0.6	0.23

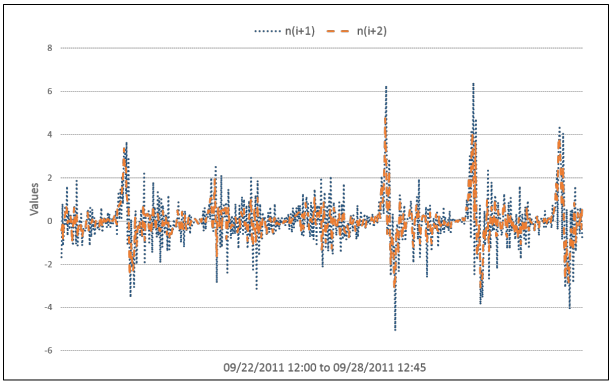
B. Local Density Based Anomaly Detection

As the first step towards anomaly detection, the variables p_j and n_k ($\forall j, k$) are obtained by taking the difference of x_i from its previous and next neighbors. The difference is divided by appropriate weights (i.e. the difference in the timestamp values) in order to quantify the contribution of the metrics based on their nearness to x_i. The variables p_j and n_k have been thus chosen to represent the 'locality' of x_i, along with appropriate weights assigned to the variables based on how 'near' they are to x_i. This nearness of x_i to the objects in a restricted neighborhood is used to identify anomalous data points.

Anomaly detection is performed based on the local outlier factor (LOF) algorithm [2], which is a local density based outlier detection approach that relies on nearest neighbors search. The approach provides a score to each data point by computing the ratio of the average densities of the point's neighbors to the density of the point itself. The local density



(a) p_j variables



(b) n_k variables

Fig. 2: Locality based statistical metrics obtained from traffic volume data.

is estimated by the typical distance at which a point can be "reached" from its neighbors. The estimated density of a point p is the number of p's neighbors divided by the sum of distances to the point p's neighbors. Thus, a point is assigned a high LOF score if its density is low relative to its neighbors' densities.

A local density based approach applied to 'locality' based metrics can effectively segregate normal data points from anomalous data points based on the sparse density of the anomaly with respect to its surroundings. This approach is useful for us because traffic volumes lying far apart (say 11 AM and 7 PM) are not expected to impact each other and are not highly correlated. The metrics chosen by us have been suitably obtained to quantify the locality of an object. The distance of an object from its neighbors in the restricted neighborhood can suitably quantify the local density of an object. Outliers are detected by segregating the data points assigned a high LOF score.

The LOF methodology in [2] assigns a degree of outlierness to each object in the dataset, thereby quantifying how outlying that object is. All normal objects will have high local density. The outlier factor for most such points is near to 1. The outlier factor moves above 1 for objects which have sparse local density. However, these cannot be immediately called as outliers, and an observer will have to use his/her understanding of the dataset to judge them, and tune the algorithm's parameters if required. Objects with outlier factor significantly above 1 are clearly identified as anomalies. For the purpose of analyzing the results obtained in this work, a local outlier factor of above 3 was flagged a clear outlier, while a value between 1.5 to 3 was flagged as a suspicious outlier. Below the value of 1.5, we consider the value as an inlier.

Our approach is presented in an algorithmic form in Algorithm 1.

Algorithm 1: Local density based outlier detection

Data: $TrafficVolumeData$
Result: $outlier$
begin
  List inlier;
  List [suspicious, confirm] outlier;
  for *Tuple* x_i *in* $TrafficVolumeData$ **do**
    Compute p_j and n_k $(\forall j, k)$ using x_i;
    $OutlierScore = LOF(x_i, p_j, n_k)$;
    if $OutlierScore \leq 1.5$ **then**
      $x_i \leftarrow inlier$;
    else if $1.5 < OutlierScore \leq 3$ **then**
      $x_i \leftarrow outlier[suspicious]$;
    else
      $x_i \leftarrow outlier[confirm]$;
  return $outlier$

C. Data formulation

The dataset published by NYSDOT is very large, consisting of highway road segments in the state of New York as well as various road segments in its different counties. An evaluation on the entire dataset will be outside the scope of this paper. For the purpose of this work, we present our analysis for single road segments, and report the findings on a single road segment of Rensselaer county: RC station 140040.

The raw traffic volume count (as in Table I) was parsed for both directions (labeled 1 and 5), and relevant metrics were extracted (as in Table II). First, in order to establish the baseline data, LOF was directly used on these metrics. LOF algorithm uses a pre-defined distance metric (euclidean distance in this case) to determine the 'distance' of a data point from its neighbors in a restricted neighborhood. The restricted neighborhood is defined by a minimum and a maximum limit on the number of neighbors to consider while calculating the distances and the outlier factor. By varying and experimenting with different values for these limits, we obtained suitable limits so that the LOF values of the dataset can serve as a baseline for the detection of future anomalies. The limits were so adjusted that the LOF values for most of the points is below or around 1. This was done with the view that the data of traffic volume counts obtained from NYSDOT should be considered free from anomalous patterns and can be used as a baseline. For some data points, however, the LOF value were far above 1 (around 3 or 4). These points are clear outliers in the baseline dataset itself.

V. Evaluation

An evaluation of the anomaly detection approach was performed by injecting false traffic volume data in the base dataset. False values were injected at occasional time-intervals: 1 PM, 7 PM, 1 AM, 7 AM, and so on. The traffic volume count of these false values was chosen to be $v\%$ higher than the original value at that timestamp. The value of x was varied as 10, 20, 30, 40 and 50. Thus, multiple test datasets were created (based on different values of v), and relevant metrics (variables p_j and n_k) were extracted from them. Local density based anomaly detection was performed on the extracted metrics (for each dataset) using LOF. Ten false data entries were added for each value of x at different timestamps in the dataset. The timestamps were chosen so as to distribute the data entries at different times of the day.

Lower values of v (namely, 10 and 20) naturally do not lead to much change in the original traffic volume count. Hence, our approach could not trigger an 'anomaly' alarm for these false values. For $v = 30$, some false injections were detected, while some escaped. But for higher values of x ($v = 40, 50$), our algorithm was successful in flagging off most of the injected anomalies (in the form of giving a high outlier score).

Irrespective of the value of v, it was noticed that the false positives (an anomaly not recognized as anomaly) were higher for the data points with their timestamp in the late night hours. This can be attributed to the low amount of traffic seen during late night hours as can be seen from the sample given in Table I. Even a 100% increment to the value 4 (timestamp 09/23/2011 01:00) will turn it into 8, which is the value for the next timestamp itself. Because of very slight changes in the metrics p_j and n_k, these false injections could not be flagged as anomalies. However, it is not a serious limitation because the majority of traffic is seen during the day, and particularly during the peak 'office hours'. It is more important for an algorithm to correctly categorize data points of the peak traffic. Also, it has been established that anomaly detection for diurnal data can be effective even when night data is not considered [4].

Values of $v \geq 60$ are not reported in this work since almost 100% accuracy for detection of anomalies was achieved with these values. Accuracy achieved in the detection of anomalies with respect to change in v is plotted in Figure 3.

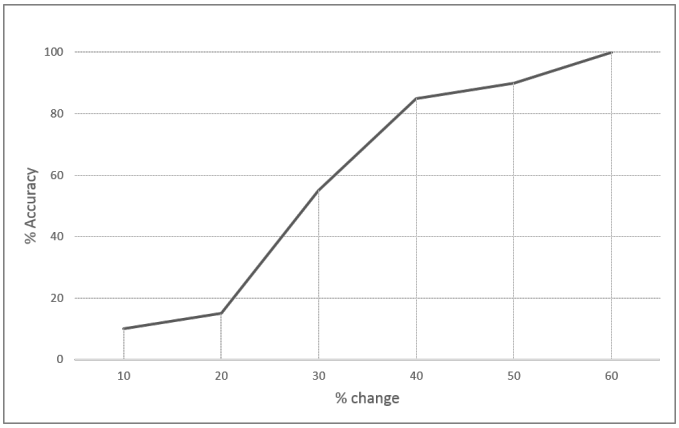


Fig. 3: % accuracy achieved in anomaly detection w.r.t % change in v

VI. Conclusion and Future Work

This paper presented an anomaly detection approach for CPS monitoring data. Traffic volume data from NYSDOT was used to test and evaluate the proposed anomaly detection approach. Our approach calculated locality based metrics from the traffic volume count at a given timestamp. These locality based metrics were used to perform local density based anomaly detection using the LOF algorithm. The approach was evaluated with false injection of traffic counts at occasional intervals. The false injections were chosen to be $x\%$ higher than the original value. Our algorithm was not successful in detecting these injections when the value of x was small. Even for higher value of x, we could not correctly categorize the false injections during the night hours. These are not serious shortcomings of our approach since small changes to the original value (or, even bigger changes when the original value itself is very small) are unlikely to have significant impact on the ITS applications.

Our approach evaluated anomalies in traffic volume data while considering the data points from that particular road segment itself. This work can be extended by considering the traffic volume count from other road segments which are understood to be 'similar' to the road segment in question. If the locality based metrics at a given timestamp deviate significantly from those at other similar road segments, an anomaly can be flagged.

References

[1] "Traffic volume data, Department of Transportation, USA," https://www.dot.ny.gov/divisions/engineering/technical-services/highway-data-services/hdsb, accessed on 20 July 2016.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[4] F. Mata, P. Żuraniewski, M. Mandjes, and M. Mellia, "Anomaly detection in diurnal data," *Computer Networks*, vol. 60, pp. 187–200, 2014.

[5] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 141–150.

[6] X. Li, Z. Li, J. Han, and J.-G. Lee, "Temporal outlier detection in vehicle traffic data," in *2009 IEEE 25th International Conference on Data Engineering*. IEEE, 2009, pp. 1319–1322.

[7] J. A. Barria and S. Thajchayapong, "Detection and classification of traffic anomalies using microscopic traffic variables," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 695–704, 2011.