Accident Analysis and Prevention 146 (2020) 105703

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Automatic incident detection on freeways based on Bluetooth traffic monitoring

Pedro Mercader, Jack Haddad*

Technion–Israel Institute of Technology, Faculty of Civil and Environmental Engineering, Technion Sustainable Mobility and Robust Transportation (T-SMART) Laboratory, Technion City, Rabin Building, Room 726, Israel

ARTICLE INFO

Keywords:
Road safety
Automatic incident detection
Bluetooth sensors

ABSTRACT

A novel automatic incident detection (AID) method for freeways, based on the use of data provided by Bluetooth sensors and an unsupervised anomaly detection approach, is presented. The two main advantages of the proposed AID system are: (i) the use of Bluetooth sensors offers several practical advantages over inductive loop detectors (ILD), which is one of the preferred sensing technology for traffic flow; and (ii) the unsupervised anomaly detection approach builds a model without the need of incident information. A common problem when designing an AID system is that incident information, i.e., ground-truth data, with enough accuracy is seldom available. Isolation forest is the unsupervised anomaly detection approach adopted in this work. This method is based on characterizing anomalous traffic conditions by exploiting the fact that anomalies tend to be isolated. The most remarkable feature of this anomaly detection method is its high detection performance while having a very simple tuning procedure and an extremely low computational demand. Finally, the effectiveness of the presented AID method is demonstrated using real traffic data collected by a network of Bluetooth sensors installed in Ayalon Highway, Tel Aviv.

1. Introduction

Intelligent transportation systems (ITS) emerge with the purpose of improving the level of utilization of road networks, without the need for large investments in new traffic infrastructures. Among the goals of an ITS are to relieve traffic congestions and to reduce the risk of traffic incidents. Automatic incident detection (AID) systems are regarded as a key component of ITS. This is due to the fact that traffic incidents are the cause of most of disruptions in freeway traffic flow; therefore, an accurate and fast detection of incidents is of vital importance in order to take actions aimed at restoring a normal traffic flow and notifying emergency personnel. In addition, an early incident detection could potentially decrease the probability of secondary incidents, and hence, improve road safety (Park et al., 2018; Wang et al., 2019). Although it is possible to detect incident by non-automatic techniques like collecting incidents information from phone calls performed by other drivers, a properly designed AID will be more efficient at managing the resolution of the incident. Recently, the use of data collected by social media has also been proposed for performing incident detection (Gu et al., 2016). Similarly to the previous case, the quality of information becomes a major challenge, since sometimes humans may behave in an untrustworthy manner (Restuccia et al., 2017).

AID systems require real-time data about the state of the traffic in the freeway sections under consideration. Traditionally, AID systems rely on data provided by inductive loop detectors (ILD) and to a minor extent video based systems (VBS). This work proposes the use of data collected by Bluetooth sensors. During the last decade, Bluetooth technology has become a mature sensing technology in ITS, see, e.g., Antoniou et al. (2016). The operation principle of this sensing technology is very simple, as it is based on the detection and re-identification of visible Bluetooth devices on-board of vehicles. Bluetooth detectors are deployed along the traffic networks to scan for Bluetooth devices in a short range. Every time a Bluetooth device is detected, the system captures a unique electronic identifier of the device, denoted as Machine Access Control (MAC) address, and the time-stamp data of the detection (Bhaskar and Chung, 2013). A monitoring system, that receives the data from all the Bluetooth detectors, calculates travel times by matching these device identifiers at successive sensor locations. The same monitoring system processes the data in order to eliminate multiple detections on the same vehicle, e.g., several devices in a bus, and possible outliers, e.g., Bluetooth devices belonging to pedestrians. In this way, the monitoring system is able to obtain the travel times of vehicles equipped with a Bluetooth device traveling through the freeway sections where the Bluetooth sensors are deployed. Finally, this

* Corresponding author.
E-mail address: jh@technion.ac.il (J. Haddad).

https://doi.org/10.1016/j.aap.2020.105703
Received 5 April 2020; Received in revised form 23 July 2020; Accepted 24 July 2020
Available online 21 August 2020

event-based data is aggregated into time intervals to create time series corresponding to the average travel time, or velocity, of the considered freeway sections.

Bluetooth sensors offer many advantages over ILD like: low installation and maintenance costs, low failure rate, and sectional data instead of local data. On the other hand, Bluetooth sensors are only able to collect data from vehicles that contain a visible Bluetooth device, and hence, they are not able to provide reliable measurements of density and occupancy. Regarding VBS, these systems are able to measure density and velocity for all vehicles traveling in a freeway section, assuming idealized conditions. However, the main drawbacks of VBS are the cost and the complexity of the required infrastructure, for example, these systems demand much more computational power to process the data than the Bluetooth detectors. In addition, re-identification of vehicles using VBS is not exempt of complications, e.g., some researches show relatively high figures for the ratio of unmatched vehicles, see Zhan et al. (2015), and this technology is also more sensitive to adverse weather conditions than Bluethooth sensors (Luvizon et al., 2016). To sum up, the main drawback of the Bluetooth technology is the inability to provide density and occupancy data. This issue poses a great challenge for AID since information about density or occupancy is needed by most of the existent AID methods (Hossain et al., 2019). Although the number of detected vehicles could be used as a proxy for density, this would rely on the assumption of an approximately constant or variable but known penetration rate. Experimental evidences show that the penetration rate has a stochastic time-varying nature along the hours of the day, see, e.g., the studies presented in Sharifi et al. (2011) and Mercader et al. (2020). The lack of density and occupancy information motivates the development of novel AID methods tailored for Bluetooth sensors.

Some related works on this research line include AID methods relying on automatic vehicle identification (AVI) technologies, specifically, methods based on electronic toll collection (ETC) systems as a measurement technology that appeared in the late nineties. Some examples of these methods that utilize data provided by ETC systems are given in Hellinga and Knapp (2000) and Khoury et al. (2003). These methods relied on classical statistic approaches. Nowadays, there exist many advantages with respect to these previous AID methods: (i) Bluetooth is a mature technology whose feasibility in transportation has been proved in many applications and the penetration rate of this technology is expected to grow in oncoming years, and (ii) recent advances in both statistics and computation allow to use the existing data in a more effective way than the classical statistical methods existing some decades ago. While classical statistical methods provide interpretable models, their performance is rather limited. In recent years, a wide range of machine learning techniques that overcome the performance of classical statistical methods have been presented. These methods include, among others, support vector machines, artificial neural networks, and ensemble methods. Many of these have been successfully applied to transportation applications, see, e.g., Zhang et al. (2011), Lv et al. (2014), Theofilatos et al. (2019), Li et al. (2020). It is worth to mention that while the use of Bluetooth technology is becoming a mature technology for traffic data measurement, its use for AID is rather scarce. Some examples where Bluetooth sensors have been proposed for AID are given in the works Yu et al. (2014), Margreiter (2016), but these methods also rely on classical statistic approaches and do not fully exploit the available data. Alternatively, some approaches combine the use of Bluetooth and adaptive signal control data, see, Yuan et al. (2018). Finally, related works using unsupervised approaches to detect non-recurrent traffic congestions (incidents) on urban road networks have been also presented, see, e.g., Li et al. (2007), Anbaroglu et al. (2014). The idea behind these methods is based on clustering congested links. Therefore, it is required that the incident affects several links. This assumption is difficult to meet in freeways, since the length of freeway sections (defined by the location of sensors) may be of the order of a few kilometers. Therefore, these methods are not suitable for detecting incidents in freeways. Another alternative approach based on spatial and temporal dependencies of the traffic flow is presented in Liu et al. (2019).

This work investigates an automatic incident detection method that relies on unsupervised anomaly detection approach, i.e., it uses exclusively data provided by Bluetooth sensors without assuming knowledge about incident information. In addition, the algorithm is able to detect incidents affecting only to one freeway section. Unsupervised anomaly detection methods generally require to build a predictive model representing the normal traffic instances and identify incident as mismatches with respect to this model, as e.g., Chandola et al. (2009), Mercader and Haddad (2018), Chakraborty et al. (2019). The main drawback of building a prediction for the normal traffic is the complexity of the design because tuning the parameters of the model in order to optimize the detection performance is a challenging task. Instead of this approach, this paper uses an *alternative* unsupervised anomaly detection technique known as isolation forest (Liu et al., 2008, 2012) that isolates the anomalous traffic conditions and uses that for generating an anomaly score. This method exploits the fact that anomalies are *few and different*, and hence, they can be easily isolated in the input space. Moreover, the fast execution speed and simplicity in its design make isolation forest very suitable for detecting anomalies in streaming data.

The remainder of this work is structured as follows. The proposed unsupervised anomaly detection system is described in Section 2. The application of isolation forest to AID is presented in Section 3. Section 4 presents the application of the proposed method to real traffic data collected in Ayalon Highway, which is a major freeway in Tel Aviv, Israel. Finally, conclusions and future work are drawn in Section 5.

2. Unsupervised anomaly detection for AID

The AID problem in freeways is tackled in this paper by using an unsupervised anomaly detection approach, see Aggarwal (2015) for more details on unsupervised approaches. This approach offers many advantages over the supervised classification approach that is typically used for the AID problem in the literature, see, e.g., Chen and Wang (2009), Wang et al. (2013), Zhu et al. (2018a), Parsa et al. (2020). A supervised classification approach requires incident information, i.e., labels indicating whether the traffic state corresponds to a normal or an incident condition, and then, the labeled set of data is used to train a model to classify new traffic conditions as normal or anomalous. There are two major issues associated with the use of a supervised classification approach for an AID problem: (i) there is an extremely high imbalance in the data between normal traffic and incident conditions, which may pose challenges to data pre-processing and classification algorithms (Parsa et al., 2019), and (ii) obtaining accurate labels classifying the traffic state as normal or anomalous is usually a challenging and costly process that requires human intervention. In the problem considered in this work, information about traffic incidents are generally recorded by human operators in a traffic control center. Due to this manual method to collect data, it is very costly to improve the quality and accuracy of incident information. These issues motivate the consideration of an unsupervised model-based anomaly detection approach as an alternative to a supervised classification approach.

Most of the unsupervised model-based anomaly detection approaches rely on assumptions that instances associated to normal conditions are much more frequent than instances associated to anomaly conditions, and that it is possible to learn a model during the training phase that is robust to these few anomalies (Chandola et al., 2009). Therefore, this model could be used to determine whether an instance corresponds to a normal traffic condition or an incident traffic condition. The main drawback of this kind of approaches is the complexity of the design due to the hyper-parameters tuning. It should be noted that this process may be very time consuming if the number of hyper-parameters to optimize is large. An alternative unsupervised anomaly

detection approach is adopted in this paper, where a method called isolation forest is used to detect incident in the incoming stream of traffic data measured by the Bluetooth sensors. This approach is able to detect anomalous conditions, without having to build explicitly a model that predicts the behavior of the system under normal conditions. A set of variables measured by the Bluetooth sensors for each freeway section is used as input variables for the isolation forest method. The list of input variables used in this work includes: time of the day, velocity, and different relative changes of velocity that account for temporal and spatial variations. Isolation forest takes advantage from the fact that anomalies are *few and different* as opposed to normal instances that are *many and similar* (Liu et al., 2008, 2012). In the context of AID, it is expected that normal traffic instances will be much more numerous than anomalous traffic instances and that the input variables corresponding to normal traffic instances will be similar among them. On the other hand, it is also expected that the input variables corresponding to anomalous traffic instances will differ considerably from input variables corresponding to normal traffic instances. The underlying principle that isolation forest assumes is that anomalies are easier to isolate than normal instances, therefore, it is possible to generate an anomaly score based on the isolation level and use a threshold over this score to detect anomalous traffic conditions. Both the model and the threshold are designed using historic data. This method is broadly illustrated in Fig. 1, where the model generates the anomalous score for a given instance of data and the threshold decides whether this value corresponds to a normal or anomalous traffic condition. The isolation forest algorithm is further described in the section ahead.

3. Isolation forest applied to AID

The state of the traffic of a freeway section during a given time interval is characterized by a set of p input variables denoted by X_1, X_2, …, X_p. The purpose of the anomaly detection method is to train a model which generates an anomaly score that makes possible to classify as normal or anomalous a given instance of data. The model is trained using an unlabeled set of data known as training set with n data instances $\{x_1, x_2, \ldots, x_n\}$, where each data instance contains information about the p input variables that describe the state of the freeway section $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, \ldots, n$. Each data instance, or observation, x_i is constituted by the numerical values of the p input variables X_1, X_2, …, X_p. The isolation forest algorithm is based on building a collection of isolation trees, usually denoted as forest of ensemble. An isolation tree performs a random binary partition of the input space X_1, X_2, …, X_p until each instance in the training set is isolated to an external node, see an example shown in Fig. 2. Each isolation tree is generated by randomly choosing a variable and its splitting value, this is referred as a test. This process ends with a tree whose all nodes are either terminal-nodes or internal-nodes with one test and two child nodes. The forest of trees is formed by a number of trees where each tree has been trained using a different training set obtained by sub-sampling without replacing the original training set. The only two design parameters involved in the model are: (i) number of trees in the forest N, and (ii) sub-sampling size S. The values $N = 100$ and $S = 256$ are proposed in Liu et al. (2008) as a rule of thumb that has provided satisfactory results in a wide range of data sets.

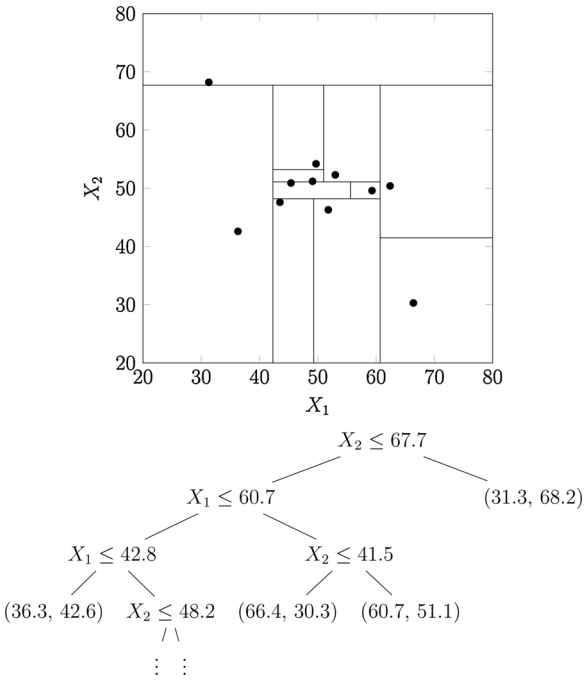

Fig. 2. Partition generated by a random binary tree (top) and truncated binary tree (bottom). At each internal node, the left branch emanates when the condition is true.

The criterion employed by the isolation forest algorithm to measure the degree of anomaly of a data instance x_i is a function of its path length $h(x_i)$. The path length of a data instance x_i, $h(x_i)$, is defined as the number of edges that a data instance x_i traverses from the root node until the terminal node. For example, considering the tree shown in Fig. 2 (bottom), the path length of the data instance (31.3, 68.2) is h ((31.3, 68.2)) = 1. The isolation forest algorithm considers the average of path lengths among all the isolation trees in the forest. Recall that the model consists of a forest with N trees. The rationale behind the choice of the average of path lengths as a measure of anomaly is that anomalies can be isolated easily and, hence, the path length for anomalies is expected to be considerable shorter than the one corresponding to the normal instances. This is illustrated in Fig. 2 (top) that shows a random binary partition of the input space (X_1, X_2) that isolates all the considered data instances in this simple example. Fig. 2 (bottom) shows the corresponding truncated tree for the aforementioned partition. It is appreciated that a random binary tree generates shorter path for anomalous instances, like for example (31.3, 68.2) and (66.4, 30.3), than for normal instances. In order to derive a normalized degree of anomaly, the average path length given the size of sub-sampling S is considered, that is given by the expression

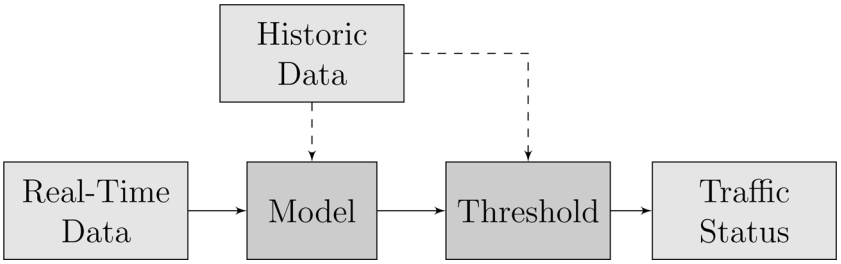

Fig. 1. Unsupervised (model-based) anomaly detection framework for incident detection. Dashed lines represent off-line processes and solid lines represent on-line processes.

$$c(S) = 2H(S-1) - 2\frac{S-1}{S},$$

where $H(S)$ stands for the harmonic number of S, that is given by

$$H(S) = \sum_{i=1}^{S} i^{-1}.$$

Then, the normalized anomaly score proposed in Liu et al. (2008) is built as follows:

$$s(x_i, S) = 2^{-E(h(x_i))/c(S)},$$

where $E(h(x_i))$ is the average of path lengths between all the trees included in the forest for the data instance x_i. The anomaly score takes values between 0 and 1, where values close to 1 denote that the sample is an anomaly. Note that as $E(h(x_i))$ tends to 0 the value of $s(x_i, S)$ tends to 1. Samples with values smaller than 0.5 are regarded as normal instances. The threshold to decide whether a sample corresponds to an anomaly can be designed by specifying the ratio of outliers R in the training set, i.e., the threshold for $s(x_i, S)$ is chosen in a way that the ratio of outliers identified by the method is equal to R. The ratio of outliers R is a third design parameter of the isolation forest model. The value of this parameter depends strongly on the data set under consideration and its tuning involves a trade-off between true and false positive rates, as it will be shown in the receiver operating characteristic (ROC) curve in the next section. The reader is referred to the survey paper Bradley (1997) for further details about the use of ROC curves for measuring the performance of classifying algorithms.

4. Application to real data

This section is devoted to the application of the proposed AID method to real traffic data collected by Bluetooth sensors placed along Ayalon Highway, Tel Aviv, see Fig. 3. The data used in this section was collected during 90 days, and corresponds to an 11 km segment that is divided into 7 sections. The number of individual trips detected in the 7 sections during 90 days amounts to more than 7 million. The incident detection analysis is performed on the 5 central sections, due to the need of data corresponding to the downstream and upstream sections. This is a common limitation of models that use information about neighbor sections. Firstly, raw data measured by the sensors have to be processed to generate a set of variables that will characterize the traffic state. Two sets of data are generated, train (first 83 days) and test (last 7 days) sets, for each of the 5 sections. Secondly, N different train sets are generated by sub-sampling S instances from the initial train set and these N sets are used to build the models for each section. Finally, the trained models are applied to instances in the test set to evaluate the anomaly detection performance. The performance of the designed AID is measured using ground-truth data from a traffic incident report provided by the traffic control center. The traffic incident report for the days corresponding to the test set was manually checked in order to verify its accuracy.

4.1. Data pre-processing

The AID method proposed in this work is fed by data provided by Bluetooth detectors. Table 1 shows some examples of data recorder by the monitoring system that receives the measurements of the Bluetooth sensors. Each row corresponds to a trip in a freeway section and includes a token that is associated with the MAC address, the origin (O) and destination (D), the instants when the vehicle was detected in O and in D, and finally, the trip time in seconds. The average velocity during a trip can be obtained using the distance between O and D. This event data corresponding to every individual trip (shown in Table 1) is aggregated into time intervals of 5 min of duration for each section obtaining the following time series:

- Velocity [km/h] in section i during the time interval t, $v_i(t)$.
- Time of the day measured in 5 minutes time intervals, t.

This aggregation of event data corresponding to individual trips into aggregated time series also serves to filter out possible extreme behaviors of individual drivers and corrupted measurements. The freeway sections are delimited by the positions of two consecutive Bluetooth detectors and they are numbered from 1 through 5, starting from the upstream-most section. Additionally, boundary sections numbered by 0 and 6 are also considered, see Fig. 3. These time series (velocity and time of the day) are used as input variables and to generate additional input variables for the isolation forest algorithm. The set of input variables used in this work to describe the condition of the traffic of a given freeway section, indexed by i, are given as follows:

- Time of the day measured in 5 min time intervals, t.
- Velocity in section i during the time interval t, $v_i(t)$.
- Relative change of velocity with respect to the previous time interval

$$r_i^p(t) = \frac{v_i(t-1) - v_i(t)}{v_i(t)}.$$

- Relative change of velocity with respect to the upstream section

$$r_i^u(t) = \frac{v_{i-1}(t) - v_i(t)}{v_i(t)}.$$

- Relative change of velocity with respect to the downstream section

$$r_i^d(t) = \frac{v_{i+1}(t) - v_i(t)}{v_i(t)}.$$

- Relative change of velocity with respect to the upstream section at the previous time interval

$$r_i^{u,p}(t) = \frac{v_{i-1}(t-1) - v_i(t)}{v_i(t)}.$$

- Relative change of velocity with respect to the downstream section at the previous time interval

$$r_i^{d,p}(t) = \frac{v_{i+1}(t-1) - v_i(t)}{v_i(t)}.$$

These seven variables are used to characterize the state of the traffic in a given freeway section in the proposed AID. The train and test sets are generated by computing these variables for the corresponding period of time. Here, the train set includes data of 83 days and the test set includes data of the 7 following days, and the numbers of data instances of the train and test sets for each section are approximately 24,000 and 2000, respectively, see data shown in Table 2. The difference in the number of data instances is due to the time intervals in which there were no detections. An example of some data instances included in the train and test sets is shown in Table 3.

4.2. Anomaly detector design

The design of the unsupervised (model-based) anomaly detector using the isolation forest algorithm is relatively simple and does not require excessive computational effort. This is crucial for real implementation since a different model is required for each freeway section. It is emphasized here that the isolation forest algorithm has only two design parameters for obtaining the anomaly score, namely, the number of trees in the ensemble N and the sub-sampling size S, for the

Fig. 3. Location of the Bluetooth sensors (black dots) placed along Ayalon Highway, Tel Aviv.

Table 1
Example of trips recorded by the monitoring system. O stands for origin, that is beginning of the section, and D for destination, that is the end of the section.

Token	O	D	Time at O	Time at D	Trip time [s]
83278124	13	14	13:04:59	13:05:46	47
89255584	15	16	13:05:02	13:05:46	44
85658501	13	14	13:05:07	13:05:46	36
89369456	19	20	13:04:07	13:05:47	100
⋮	⋮	⋮	⋮	⋮	⋮
83278124	14	15	13:05:46	13:07:02	76

Table 2
Number of samples in train and test sets.

Section	Samples in train set	Samples in test set
1	22,953	1902
2	23,524	1961
3	23,883	1989
4	23,897	1989
5	23,748	1984

Table 3
Example of data in train and test sets.

t	v_5	r_5^p	r_5^u	r_5^d	$r_5^{u,p}$	$r_5^{d,p}$
1	76.68	0.0079	0.3107	0.5896	0.3207	0.5562
2	76.68	−0.0007	0.3952	0.5855	0.3098	0.5885
3	77.40	−0.0109	0.3604	0.5388	0.3798	0.5680
4	82.08	−0.0546	0.3068	0.5099	0.2861	0.4548
5	78.84	0.0390	0.3172	0.4975	0.3578	0.5689
6	78.48	0.0054	0.3151	0.5032	0.3244	0.5057

model that generates the anomaly score $s(x_i, S)$ presented in the previous section. Additionally, the threshold to decide whether an instance is classified as an incident is defined by the ratio of outliers R that is a third design parameter.

The model that generates the anomaly score is built using the train set, and its parameters, N and S, are designed to optimize a performance criterion evaluated over the test set, i.e., instances of data which are not used for building the model. The performance in anomaly detection tasks is usually evaluated by the area under the receiver operating characteristic (ROC) curve, denoted as area under the curve (AUC). A value of AUC equal to 1 corresponds to a perfect classifier and a value of

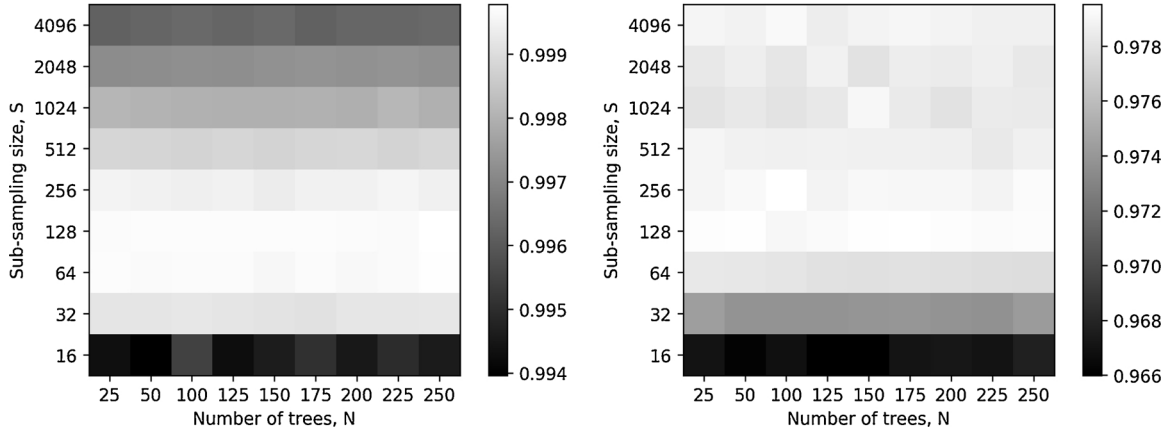

Fig. 4. AUC for the models of the sections 1 (left) and 5 (right) evaluated over the test set as a function of the design parameters S and N. The values shown correspond to the average between 100 replications.

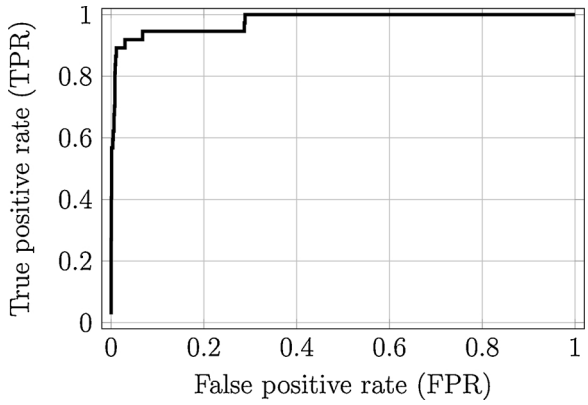

Fig. 5. ROC curve for the model of the section 5 evaluated with the parameters $N = 100$ and $S = 256$. Variations on R allow us to move along the ROC curve, low values of R correspond to the bottom left part of the ROC curve and high values correspond to the top right part.

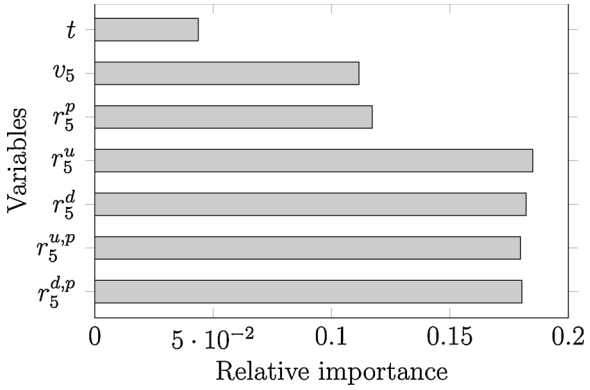

Fig. 6. Relative importance of the input variables of the model corresponding to the section 5.

Table 4
AUC values for the obtained models, using $N = 100$ and $S = 256$.

Section	AUC
1	0.9982
2	0.9977
3	0.9987
4	0.9942
5	0.9797

0.5 corresponds to a random classifier. The ROC curve illustrates the performance of a classifier in terms of true positive rate (TPR) and false positive rate (FPR) (Hastie et al., 2009). The TPR is defined as the fraction of correctly classified positives out of the total positives and the FPR as the fraction of incorrectly classified positives out of the total negatives. Note that the TPR is also known as the *recall* and the FPR as the *fall-out* or *false alarm ratio*. In the context of this work, a true positive rate is the incident detection rate and false positive is a false alarm ratio. The ROC curve shows TPR and FPR values for all possible thresholds (see Fig. 5), the threshold is a function of the ratio of outliers, R. Low values of R correspond to TPR and FPR near 0 and high values of R correspond to TPR and FPR near 1. It is necessary to emphasize that labeled data, i.e., report containing information about incidents, is exclusively necessary to measure the detection performance of the model, but not for training the model. Fortunately, the value of these parameters, N and S, are not very critical when optimizing the detection performance, as it will be shown later, and some default parameters could be used as suggested in Liu et al. (2008) without a significant loss of detection performance.

Fig. 4 shows the AUC for the model corresponding to the freeway sections number 1 (left) and 5 (right) as a function of the design parameters, N and S. The value of the AUC shown in the figure corresponds to the average between 100 replications in order to account for the randomness present in the method for building the isolation forests. As it is shown in Fig. 4, the AUC value is not very sensitive to the design parameters, the same conclusion remains valid in the models of the rest of freeway sections. The suggested values in Liu et al. (2008) turn out to deliver good performance in this empirical study. Note that for the section number 5 the AUC values for the different values of N and S are contained in the interval from 0.965 to 0.980. Similar results are obtained for the section number 1.

Fig. 5 shows the ROC curve for the model corresponding to the freeway section 5 with the parameters $N = 100$ and $S = 256$. As shown in this figure, it is possible to detect more than 90% of incidents while having a false alarm rate of about 1.5%. We recall that the value of the ratio of outliers, R, is the parameter that balances the trade-off between true and false positive rate, i.e., variations on R allow us to move along the ROC curve. The AUC values for the models corresponding to each freeway section, with the parameters $N = 100$ and $S = 256$, are shown in Table 4.

The values for the AUC are similar or even superior to the obtained in other supervised classification approaches where information about speed, occupancy and volume is used. For example, the paper Wang et al. (2008) applies partial least squares regression to the I-880 data set obtaining an AUC value of around 0.94.

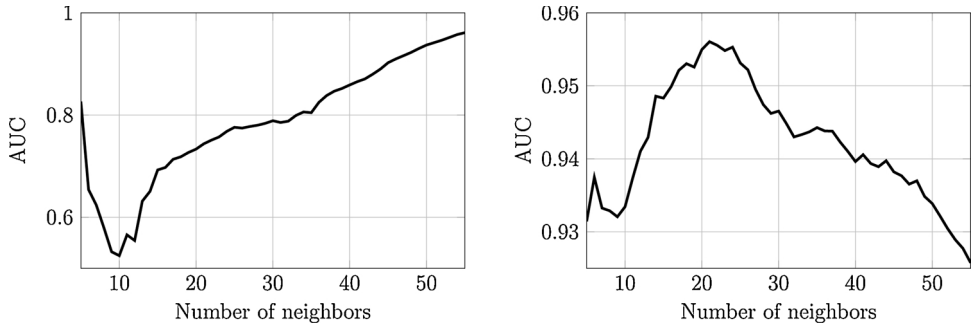

Fig. 7. AUC versus number of neighbors for the local outlier factor (LOF) model when applied to the freeway sections 1 (left) and 5 (right).

Appendix A shows the velocity versus time of the day for the test set, indicating the anomalies detected by the proposed AID method and the traffic incidents that were reported during this period of time. The value of R was set to 0.005 in the threshold used in this test.

4.3. Variable importance

The input variables usually have different influence over the output of the model, and a measure of this importance provides some insights into the data. In the context of the random forest algorithm for supervised learning (Hastie et al., 2009), the variable importance indicates the reduction in the cost function during the training process associated with a given variable. The cost function measures the mismatch between the labeled data and the output of the model. A large value indicates that this variable has a large influence on the output of the model. Unfortunately, the measurement of the variable importance cannot be applied to the isolation forest algorithm due to its unsupervised nature. However, it is possible to generate a labeled set of data by considering the input variables and the outputs generated by the isolation forest algorithm to train a classification model using random forest algorithm. In this way, it is possible to measure indirectly the variable importance in the proposed AID method. Fig. 6 shows the relative importance (normalized variable importance) of the input variables of the model corresponding to the freeway section 5. In this model, the variables with higher relative importance are the relative change of velocity with respect to the upstream and downstream sections at the present and previous time intervals. The variable with less relative importance is the time of the day, this is partly expected since the probability of occurrence of an incident is independent of the time of day. It is interesting to note that some papers do not include the variable time of the day in the analysis to detect incidents, see for example, Parsa et al. (2020).

The procedure presented above was used to select the set of input variables used in this work. Initially, a larger set of input variables was considered but eventually the variables with lower importance were dropped. Some examples of the input variables that were dropped are relative change of velocity with respect to other sections in addition to downstream and upstream, day of the week, and number of detected vehicles. An alternative approach to avoid this process of featuring engineering (selecting the adequate input variables) is to use deep learning methods for anomaly detection, see Chalapathy and Chawla (2019), but this comes at the expense of having to tune a large number of parameters.

4.4. Comparison with other approaches

Most of the literature in AID methods adopt a supervised approach requiring incident information that rarely has sufficient accuracy. Some examples of such methods are Chen and Wang (2009), Wang et al. (2013) and Parsa et al. (2020). On the other hand, applications of unsupervised approaches to AID in freeways are very scarce. Most of the unsupervised approaches are based on building a model able to predict the behavior under nominal condition and identify incident as mismatches between the prediction and the measurement. An additional caveat of this approach is the large number of parameters that have to be tuned for the success of the implementation. Some papers where this approach has been adopted are Chandola et al. (2009), Mercader and Haddad (2018) and Chakraborty et al. (2019).

However, the use of unsupervised approach without predicting normal conditions has been proposed for data quality control or data cleaning, see, Chen et al. (2010). In addition, this approach typically has a relative small number of free parameters. This process aims to identify and remove the anomalous data from the data set. Traditionally, this anomalous data may be of two kinds: (i) anomalies caused by equipment failures, and (ii) incidents or anomalous traffic events. The first kind, anomalies due to equipment failures, is not a big concern for Bluetooth detectors since the monitoring system usually provides a variable indicating the status of each sensor. Under this premise, the anomalies detection approaches presented in Chen et al. (2010) could be used for incident detection.

We compare the proposed AID approach to the anomaly detection method included in Chen et al. (2010). In particular, it is considered a well-known density-based approach called local outlier factor (LOF) that offers superior performance compared to other methods included in Chen et al. (2010) and does not require knowledge about the probability density function of the data set. The general principle behind density-based approaches is that instances in low density regions can be regarded as anomalies. The reader is referred to Chen et al. (2010) for a more detailed exposition of this method. Fig. 7 shows AUC versus number of neighbors, that is the main design parameter in LOF, for the LOF model applied to the freeway sections number 1 and 5. This figure shows values of AUC in the range from 0.52 to 0.96 for section 1 and from 0.93 to 0.96 for section 5. These results can be compared to the one presented in Fig. 4 that shows AUC values in the range from 0.994 to 0.999 for section 1 and from 0.965 to 0.980 for section 5. It is observed that LOF has higher sensitivity to the design parameter (number of neighbors) while having a lower detection performance than the isolation forest algorithm used in the proposed AID methodology.

5. Conclusions and future work

This paper has presented an unsupervised (model-based) anomaly detection approach for AID based on Bluetooth tracking data. The presented AID method is easily implementable in practice. The calibration of the model is relatively simple and does not require labeled data (incident information), which usually is not reliable. Additionally, Bluetooth sensors offer a very cost-effective solution against other measuring methods like inductive loop detectors. It is worth to emphasize that the methodology proposed here could be applied to traffic data collected by other sensing technologies that provide similar type of data collected from Bluetooth sensor. The motivation of the proposed approach is that most of the previous AID methods require information

about occupancy or density, but Bluetooth sensors are not able to provide such information. Therefore, tailored AID method for Bluetooth sensors is needed.

Application of the developed AID method has delivered satisfactory performance level, measured quantitatively in term of AUC, reaching similar values to state-of-the-art AID methods based on supervised classification approach and relying on speed, volume and occupancy data. A higher performance for the presented AID method is expected if information about volume and occupancy would also be used.

It is also expected that other advanced methods like unsupervised deep learning, see Chalapathy and Chawla (2019), Kwon et al. (2019), may achieve the same or higher detection performance than the proposed method. However, this is at the expense of using a more complex model (large number of parameters) than the proposed in this work.

Finally, a caveat of the proposed AID method is that it is able to identify anomalous traffic conditions, but it is not able to distinguish the mechanism that generated these conditions, e.g., traffic accidents, maintenance works, or anomalous traffic patterns. Future works could explore the process of identification of anomalies and posterior classification on data streams by applying novel techniques based on semi-supervised learning (Mu et al., 2017; Zhu et al., 2018b).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author's contribution

The authors equally contributed to the context of the paper.

Appendix A. Application of the proposed AID to the test set

Figs. A.8–A.11 show the velocity versus time of the day for the considered freeway sections along some of the days included in the test set. The red shadow areas correspond to traffic incidents that were reported on the incident report provided by the traffic control center. Finally, the traffic conditions detected as anomalous by the designed AID methods are displayed using red dots.

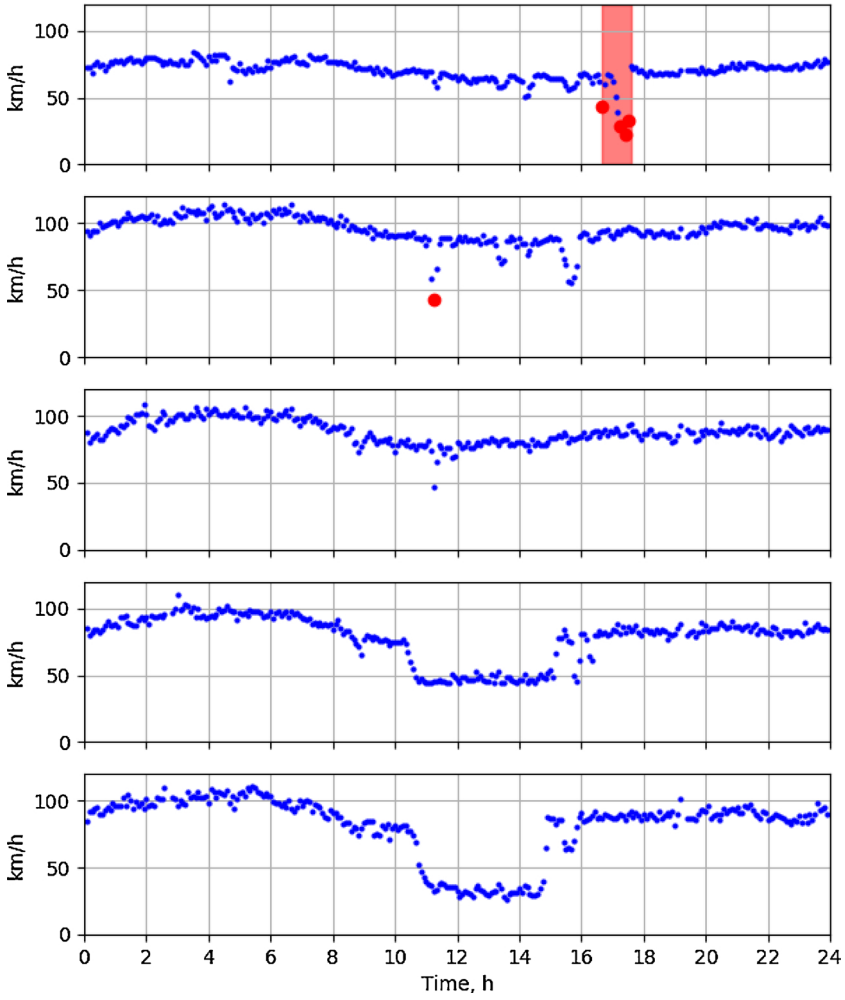

Fig. A.8. Section 1 to 5 (from bottom to top) during the day 1 of the test set. Red points indicate detections of anomalous traffic conditions. Red shadow area indicates traffic incident according to the incident report. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

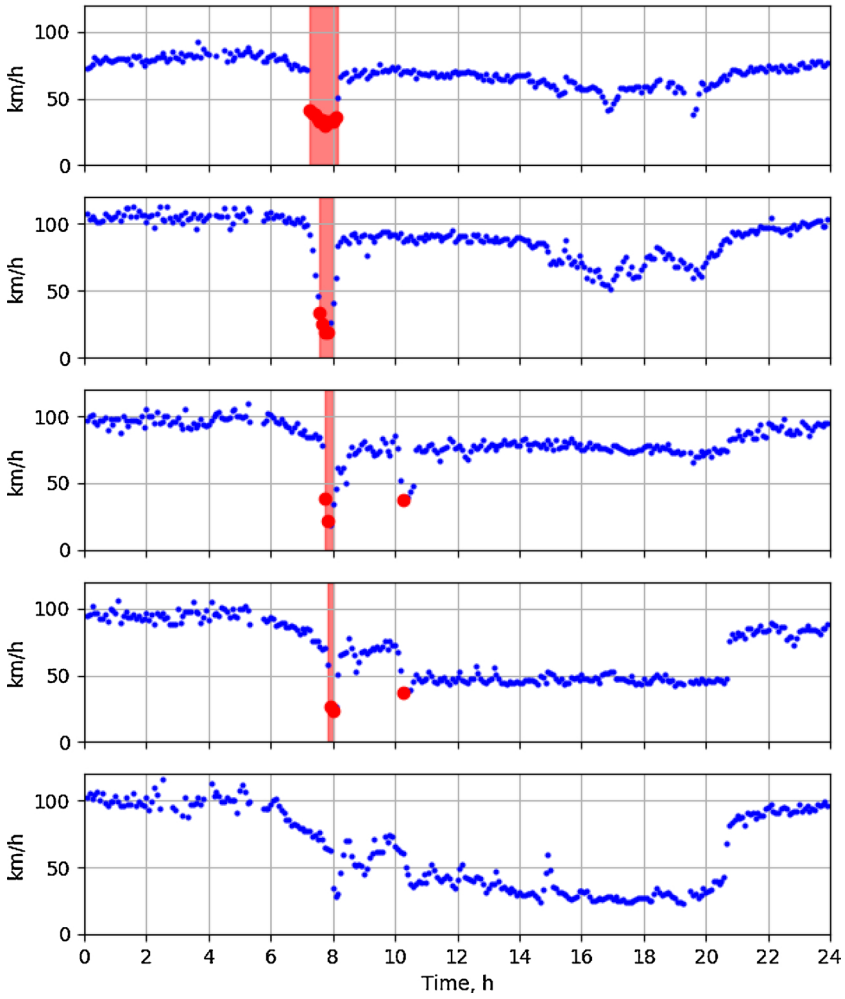

Fig. A.9. Section 1 to 5 (from bottom to top) during the day 4 of the test set. Red points indicate detections of anomalous traffic conditions. Red shadow areas indicate traffic incident according to the incident report. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

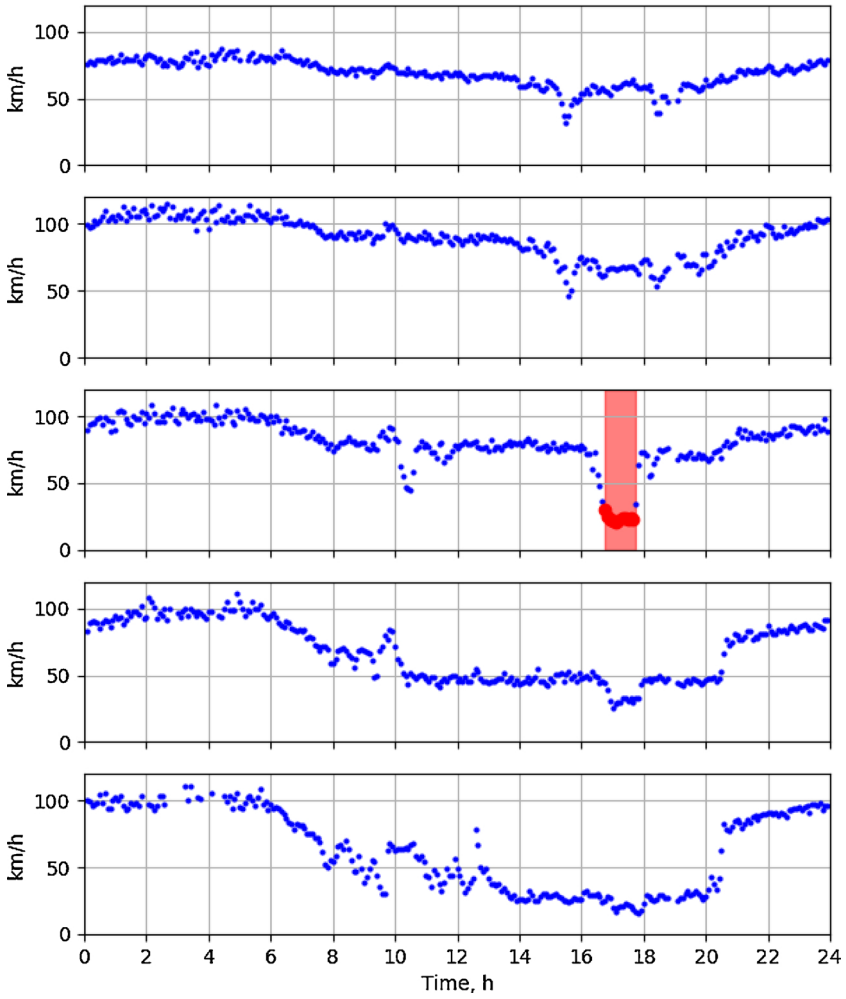

Fig. A.10. Section 1 to 5 (from bottom to top) during the day 5 of the test set. Red points indicate detections of anomalous traffic conditions. Red shadow areas indicate traffic incident according to the incident report. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

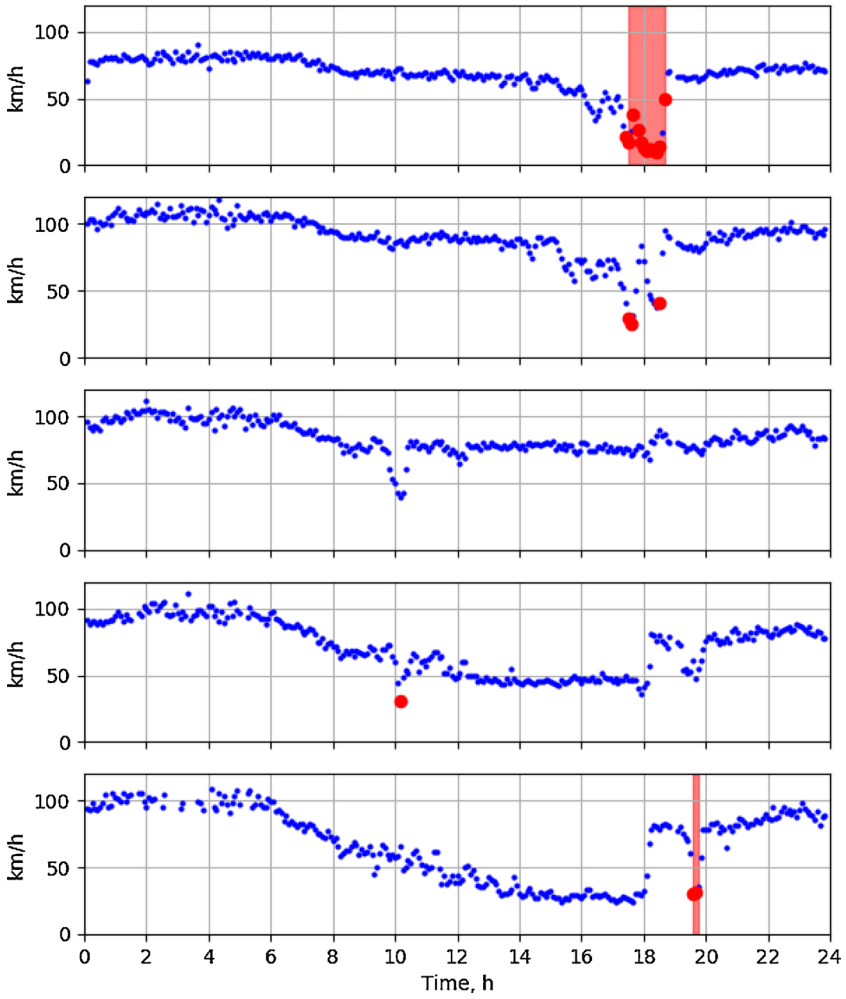

Fig. A.11. Section 1 to 5 (from bottom to top) during the day 7 of the test set. Red points indicate detections of anomalous traffic conditions. Red shadow areas indicate traffic incident according to the incident report. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Aggarwal, C.C., 2015. Data Mining. Springer.

Anbaroglu, B., Heydecker, B., Cheng, T., 2014. Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. Transp. Res. Part C: Emerg. Technol. 48, 47–65.

Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., Ciuffo, B., Djukic, T., Hoogendoorn, S., Marzano, V., et al., 2016. Towards a generic benchmarking platform for origin-destination flows estimation/updating algorithms: Design, demonstration and validation. Transp. Res. Part C: Emerg. Technol. 66, 79–98.

Bhaskar, A., Chung, E., 2013. Fundamental understanding on the use of bluetooth scanner as a complementary transport data. Transp. Res. Part C: Emerg. Technol. 37, 42–72.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.

Chakraborty, P., Hegde, C., Sharma, A., 2019. Data-driven parallelizable traffic incident detection using spatio-temporally denoised robust thresholds. Transp. Res. Part C: Emerg. Technol. 105, 81–99.

Chalapathy, R., Chawla, S., 2019. Deep Learning for Anomaly Detection: A Survey. arXiv:1901.03407.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: a survey. ACM Comput. Surv. 41 (3), 15.

Chen, S., Wang, W., 2009. Decision tree learning for freeway automatic incident detection. Expert Syst. Appl. 36 (2), 4101–4105.

Chen, S., Wang, W., van Zuylen, H., 2010. A comparison of outlier detection algorithms for its data. Expert Syst. Appl. 37 (2), 1169–1178.

Gu, Y., Qian, Z.S., Chen, F., 2016. From twitter to detector: real-time traffic incident detection using social media data. Transp. Res. Part C: Emerg. Technol. 67, 321–342.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer New York Springer Series in Statistics.

Hellinga, B., Knapp, G., 2000. Automatic vehicle identification technology-based freeway incident detection. Transp. Res. Rec.: J. Transp. Res. Board (1727), 142–153.

Hossain, M., Abdel-Aty, M., Quddus, M.A., Muromachi, Y., Sadeek, S.N., 2019. Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. Accid. Anal. Prev. 124, 66–84.

Khoury, J.A., Haas, C.T., Mahmassani, H., Logman, H., Rioux, T., 2003. Performance comparison of automatic vehicle identification and inductive loop traffic detectors for incident detection. J. Transp. Eng. 129 (6), 600–607.

Kwon, D., Kim, H., Kim, J., Suh, S.C., Kim, I., Kim, K.J., 2019. A survey of deep learning-based network anomaly detection. Cluster Comput. 1–13.

Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. Accid. Anal. Prev. 135, 105371.

Li, X., Han, J., Lee, J.-G., Gonzalez, H., 2007. Traffic density-based discovery of hot routes in road networks. International Symposium on Spatial and Temporal Databases 441–459.

Liu, C., Zhao, M., Sharma, A., Sarkar, S., 2019. Traffic dynamics exploration and incident detection using spatiotemporal graphical modeling. J. Big Data Anal. Transp. 1 (1), 37–55.

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation forest. 2008 Eighth IEEE International Conference on Data Mining 413–422.

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2012. Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data 6 (1), 3.

Luvizon, D.C., Nassu, B.T., Minetto, R., 2016. A video-based system for vehicle speed measurement in urban roadways. IEEE Trans. Intell. Transp. Syst. 18 (6), 1393–1404.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.-Y., 2014. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. 16 (2), 865–873.

Margreiter, M., 2016. Automatic incident detection based on bluetooth detection in northern Bavaria. Transp. Res. Proc. 15, 525–536.

Mercader, P., Haddad, J., 2018. Automatic incident detection in freeways by using bluetooth based tracking. 7th Symposium of the European Association for Research in Transportation (hEART 2018).

Mercader, P., Uwayid, W., Haddad, J., 2020. Max-pressure traffic controller based on travel times: an experimental analysis. Transp. Res. Part C: Emerg. Technol. 110,

275–290.
Mu, X., Ting, K.M., Zhou, Z.-H., 2017. Classification under streaming emerging new classes: a solution using completely-random trees. IEEE Trans. Knowl. Data Eng. 29 (8), 1605–1618.
Park, H., Haghani, A., Samuel, S., Knodler, M.A., 2018. Real-time prediction and avoidance of secondary crashes under unexpected traffic congestion. Accid. Anal. Prev. 112, 39–49.
Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. 136, 105405.
Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 129, 202–210.
Restuccia, F., Ghosh, N., Bhattacharjee, S., Das, S.K., Melodia, T., 2017. Quality of information in mobile crowdsensing: survey and research challenges. ACM Trans. Sensor Netw. 13 (4), 34.
Sharifi, E., Hamedi, M., Haghani, A., Sadrsadat, H., 2011. Analysis of vehicle detection rate for bluetooth traffic sensors: a case study in Maryland and Delaware. Proceedings of the 18th World Congress on Intelligent Transport Systems 16–20.
Theofilatos, A., Chen, C., Antoniou, C., 2019. Comparing machine learning and deep learning methods for real-time crash prediction. Transp. Res. Rec. 2673 (8), 169–178.
Wang, J., Li, X., Liao, S.S., Hua, Z., 2013. A hybrid approach for automatic incident detection. IEEE Trans. Intell. Transp. Syst. 14 (3), 1176–1185.
Wang, J., Liu, B., Fu, T., Liu, S., Stipancic, J., 2019. Modeling when and where a secondary accident occurs. Accid. Anal. Prev. 130, 160–166.
Wang, W., Chen, S., Qu, G., 2008. Incident detection algorithm based on partial least squares regression. Transp. Res. Part C: Emerg. Technol. 16 (1), 54–70.
Yu, W., Park, S., Kim, D.S., Ko, S.-S., 2014. Arterial road incident detection based on time-moving average method in bluetooth-based wireless vehicle reidentification system. J. Transp. Eng. 141 (3), 04014084.
Yuan, J., Abdel-Aty, M., Wang, L., Lee, J., Yu, R., Wang, X., 2018. Utilizing bluetooth and adaptive signal control data for real-time safety analysis on urban arterials. Transp. Res. Part C: Emerg. Technol. 97, 114–127.
Zhan, X., Li, R., Ukkusuri, S.V., 2015. Lane-based real-time queue length estimation using license plate recognition data. Transp. Res. Part C: Emerg. Technol. 57, 85–102.
Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., Chen, C., et al., 2011. Data-driven intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. 12 (4), 1624–1639.
Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T., 2018a. Big data analytics in intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. 20 (1), 383–398.
Zhu, Y., Ting, K.M., Zhou, Z.-H., 2018b. Multi-label learning with emerging new labels. IEEE Trans. Knowl. Data Eng. 30 (10), 1901–1914.