Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Symposium of Transport Simulation (ISTS'18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)

Anomaly Detection for Instantaneous Driving Speed Distribution Obtained from ETC 2.0 Data in Japan

Wataru Yamamoto[a], Yuki Yamamoto[a], Rina Takayama[b] and Makoto Tsukai[a] *

[a]*Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima 739-8527, Japan*
[b]*Hiroshima city goverment, 1-6-34 Kokutaijicyo, Naka-ku, Hiroshima 730-8586, Japan*

Abstract

The rapid improvement in traffic monitoring devices to record the traffic state has been developed from road side loop coils to probe car recorders and then to GPS records with on-site data uploading. This study purposed a method to detect the anomaly driving speed decrease obtained from ETC 2.0 dataset in Japan. For the method, decision tree analysis and the statistical test for instantaneous driving speed distribution is used. On the basis of the distribution, an anomaly index for each section in a link was calculated to quantify the characteristics of the road-side section. The estimation results of accident analysis using the anomaly index showed that the driving speed or its variance difference of the continuous section affects accidents more than the state of driving speed anomaly.

Keywords: decision tree; statistical test; anomaly index; driving speed decrease; traffic accident

1. Introduction

Due to the rapid improvement in traffic monitoring system, devices to record the traffic state has changed from road side loop coils to probe car recorders and then to GPS records with on-site data uploading. The updating in monitoring devices also leads to the change in the dataset in traffic state, especially in terms of spatial or temporal

* Corresponding author. Tel.: +81-82-424-7849; fax: +81-82-424-7849.
E-mail address: wataruy9@hiroshima-u.ac.jp

observation format. For example, the loop coils can record traffic status at the fixed site, while probe cars can collect the traffic state along the driving route like a float on the water flow. Comparing to those conventional systems, GPS traffic monitoring to record the traffic state along the driving route is characterized by a huge number of cars with the recording devices. Such the huge information about traffic status by GPS system, however, requires a novel data processing approach to extract a valuable implication.

GPS records obtained from the driving car include the trip trajectory between the origin to the destination, and the records can be used to make OD table or to provide the link travel time. Considering the conventional data usage approaches, OD estimation naturally requires a large number of samples for each OD, while the link travel time provision do not require the complete OD trajectory, but require the partial trajectory enough to cover the links on the network, with filling out the time domain for dynamic monitoring to traffic state. The GPS observation can collect the instantaneous speed at the recorded timing with its location, however, was merely used for traffic management or planning. The analysis for the instantaneous driving speed distribution observed for short section would be useful to clarify the characteristics of driving speed distribution along the routes.

A driving speed is thought one of the causes of traffic accidents. For example, the level of average and variance of driving speed for each section at accident is applied as the factor. In addition, the difference of continuous driving speed is used. However, we do not know which one of above the traffic states has more influence on the accident. To clarify the relationship between occurrence of the accident and the traffic states helps to discuss a countermeasure of the traffic accident.

This study purposes to detect the anomaly driving speed decrease to cause traffic accidents obtained from ETC 2.0 dataset in Japan. Anomaly index is calculated for the dataset, on the basis of rational data classification into ordinary and anomaly subsets. For this classification, decision tree analysis and the statistical test for instantaneous driving speed distribution is used. This method can find the relationship between the proposed anomaly index and the potential risk for traffic accident, by estimating the traffic accident risk model including the anomaly index. The estimated model can clarify the significant factor to cause accident with the assumption of served conditions of traffic state.

2. Related work

Anomaly detection has attracted research interest not only for communication networks and social networks, but also for probe data. Qi and Mohamed (2015) explored a viability of a proactive real-time traffic monitoring strategy to simultaneously evaluate traffic states and safety. The objective of their study was to improve a system performance of urban expressways by reducing congestion and accident risk. In this study, Microwave Vehicle Detection System deployed on the expressway network in Orlando was utilized for analysis. Franco et al. (2018) developed accident prediction models for a stretch of an urban expressway at Autopista Central in Santiago, Chile. A process includes a random forest procedure to identify a strongest precursors of accidents, and a calibration/estimation of Support Vector Machine and Logistic regression classification models. Quan et al. (2014) proposed a congestion detection and notification scheme using VANETs for urban expressways. VANETs was a spatial–temporal effectiveness model to disseminate a congestion area and survival time of the congestion. Simulations through Trans Modeler indicated that the scheme ensures the accuracy of the estimated congestion level. Consequently, the scheme can provide effective references for driving path-planning or on-site route shift. Qinghua et al. (2014) presented an automatic incident detection methodology that integrates moving average model with stationary wavelet decomposition, which layer coefficients are extracted from a difference between an upstream and downstream occupancy. The proposed methodology was validated with the data obtained in Tokyo Expressway by ultrasonic sensors. Experimental results showed that the proposed model can distinguish traffic accident from other condition such as cyclic traffic congestion. Asakura et al. (2015) proposed two incident detection methods using probe vehicle data. The first method aggregates the properties of probe vehicles, and the second method combines probe vehicle trajectories with a shock wave after an incident. These methods could effectively predict the time and place of road congestion caused by an incident. Lishuai et al. (2016) presented a novel approach to apply data mining in flight data. In this approach, they applied a Gaussian Mixture Model (GMM) detected the flights with unusual data patterns. Kinoshita et al. (2015) proposed a procedure to detect traffic incidents from probe data by identifying ordinary congestion under daily traffic conditions from unusual congestion. Zhenhua et al. (2016) employed dictionary-based compression algorithm to identify the features of both spatial and temporal patterns in a multi-dimensional traffic-related data to reveal a characteristics of regional traffic flow patterns in urban road networks. In proposed methods, several indices of anomaly were

formulated. In the proposed model, supervised learning was made to give the indicator of anomaly. On the other hand, when the training data is not available, anomaly in the dataset should be detected by referring to the statistical characteristics of the given dataset. This study proposes a procedure to give anomaly index based on non-supervised learning. In order to ensure the stability of anomaly condition in driving speed, instead, machine learning approach is used.

One of machine learning which performs conditional classification is decision tree analysis. Soyoung et al. (2016) studied to review, validate, specify, and prioritize Korea's strategic policies for pedestrian safety enhancement using the decision tree method to model pedestrian injury severities. The results showed that pedestrian age and walking area were primary conditions to cause pedestrian fatalities and severe injuries. Aurenice et al. (2017) identified some rules for detecting traffic accidents from accident records by decision tree algorithm. Jinxian et al. (2015) developed a regression tree-based model to predict subway incident delays, which are major negative impacts on the passengers.

In this study, decision tree analysis and a statistical test for instantaneous driving speed distribution is used to classify the instantaneous driving speed distribution into ordinary and anomaly subsets. And anomaly indices of each section are calculated on the basis of classified subsets in order to clarify whether the accidents occurrence is conditioned by the anomaly index or not.

3. Theoretical background of decision tree analysis

A driving speed decrease observed in driving trajectory is detected by decision tree analysis with statistical test. The theory of decision tree analysis and statistical test is summarized as follows.

3.1. Decision tree analysis

Decision tree analysis give a tree structure to divide the whole samples into several subgroups with the branching variables. Branching variable at the node in the tree is obtained from several candidate variables. In the decision tree a branching variable is called a node, and a line connecting the branching condition with the follows branching condition is called a link, and each of end node in the tree is called a leaf. Depending on a parameter setting of the algorithm, a three composed of some branches to give significantly different distribution of objective variable is obtained. In this study, we use a binary tree with two or less branches. Many algorithms in the decision tree analysis such as C 5.0, ID 3, CHAID, CART. In this study, we use CART provided as rpart in R. By following the links from the root to each leaf, we can obtain the combination of the branching conditions i.e. a set of conditional variables with driving speed distribution of each leaf. CART generates a tree model by the following three steps.1):tree growing 2):tree pruning 3):selection of the optimum tree. The first step is to divide the sample into two groups and set the variable giving the most significant subgroups in terms of the difference in objective variable. Such the division is recursively applied for each subgroup to meet the criteria of subgroup size, tree depth, or significant level of subgroup difference. The second step prunes unimportant branches from the tree generated in the first step to improve the accuracy of the analysis. In the third step, trees are evaluated to find an optimal one by using the cross validation method.

3.2. Statistical test

A determination of the driving speed decrease requires the following two statistic steps to find the conditions and then to test the significance of conditions. However, to test the raw GPS speed is difficult because the conditional distribution of speed is merely different from the whole data distribution, and because conventional statistical test is too sensitive in large sample dataset. This is a common problem in applying normal distribution hypothesis test to big data. For example, suppose that a normal distribution is assumed for distribution of all data and leaves, and a normal test is performed. The normal test statistic z is expressed by the following equation (1), assuming that the number of samples of all the data is N, the number of samples of the leaf is n_i, the average and variance of all data is μ, σ, and the average and variance of leaf i areμ_i, σ_i. When N is very large, z easily exceeds the conventional significance level such as 0.05 or 0.01. Similarly, when calculating the difference between the two distributions based on the goodness-of-fitness test. Let N_i (i:1~r) be the frequency of observations in each data range, and n_i for the same in each leaf. The

test statistic χ^2 is given by equation (2). Alternatively, the test statistic D of the Kolmogorov-Smirnov test, which is a kind of nonparametric test to detect the difference in sample distribution, is equation (3). Where, $F_n(x)$ is an empirical distribution function of a leaf, and $F(x)$ is an empirical distribution function of whole data. Note that in any of the above tests, the test statistic is affected by the number of samples. In the above statistical test, the significance level is set when N and n_i are relatively small. Therefore, if the number of samples is large as much, a significant difference will be easily detected between all distributions with ordinal the significance level.

$$z = \frac{\mu - \mu_i}{\sqrt{\frac{\sigma^2}{N} + \frac{\sigma_i^2}{n_i}}} \tag{1}$$

$$\chi^2 = \sum_{i=1}^{r} \left\{ \frac{(n_i - np_i)^2}{np_i} \right\} \tag{2}$$

$$D = \frac{Nn_i}{N + n_i} \sup \left| F_n(x) - F(x) \right| \tag{3}$$

In order to avoid above shortcomings the following data processing and statistical test are applied. In this study, the Cochran-Armitage test to find a tendency of linearity of categorical frequency and the ordered categories, is used, since the test does not include the number of samples in the test statistic. The test statistics is obtained as the function of the slope when n_k is regressed to ordered categorical variable where n_k is a corresponding sample number of each categorical variable. Note that k is the number of categories. This test is applicable only when linearity exists in n_0 and n_i. Note that n_0 is the number of whole sample and n_i is the number of each leaf sample. In the following analysis, whole samples are divided by the ranges calculated by the percentile value (hereinafter referred to as %ile) of whole data. The advantage of the above non-linear transformation is to enable to detect the small decrease but frequency occurring phenomena in driving speed distribution. Then, obtain ordered category variable and the number of samples corresponding to each range. As the result, the expected distribution of whole data becomes uniform distribution. A problem in our dataset is that the speed data can only be obtained as an integer, so then threshold values between %ile categories given also as integers. In such data, different %ile categories can fall within a same integer range if division is too fine. While the division is made rough, an ability to detect the speed decrease would be low. Suppose the sum of squared residuals from linear function by categorical variable of whole data and each leaf i are S_0, S_i respectively. The determination coefficient for whole sample or for each leaf i are R_0^2, R_i^2, respectively. The variances of whole sample and each leaf are given equation (4). Then, the variance s of both data is obtained in equation (5). From the above, the test statistics is given by the following equation (6), assuming that the slope of whole sample n_0 is β_0, and the slope of each leaf is β_k. If there is a significant difference between the distribution of whole data and the distribution of leaves t value is negative (the slope of the leaf is negative). The significant driving speed decrease is obtained with the significant level at 1%.

$$s_0^2 = s_0(1 - R_0^2), \qquad s_i^2 = s_i(1 - R_i^2) \tag{4}$$

$$s = \sqrt{\frac{s_0 + s_i}{n_0 + n_i - 4}} \tag{5}$$

$$t = \frac{\beta_0 - \beta_i}{s\sqrt{\frac{1}{n_0} + \frac{1}{n_i}}} \tag{6}$$

3.3. Anomaly indices

The samples with anomaly label satisfies the condition that the speed decreases detected by the decision tree analysis and the statistical test. The rest of samples are labeled with ordinary. By using these data following three indices are 1) an anomaly score, 2) a driving speed effect size, and 3) a variance difference are calculated as indices indicating the level of anomaly.

1) anomaly score

The anomaly score is calculated by logarithm of the ratio of cumulative probability density calculated to test between the ordinary and anomaly driving speed distribution function of x_i. Due to the sample limitation i indicates the section of highway. In order to clarify the implication of index, we set x as speed limit of the section. This anomaly score represents the deviation between an ordinary driving speed distribution and anomaly one with respect to the speed limit. Therefore, a section with small anomaly score can be regarded as an easy to drive.

$$a(x_i) = \log \frac{\int_0^{x_i} p(x_i)dx_i}{\int_0^{x_i} p(x_i')dx_i'} \tag{7}$$

Where, $a(x_i)$ is the anomaly score, i is the section, $p(x_i)$ is the probability density of the anomaly, and $p(x_i')$ is the probability density of the ordinary.

2) driving speed effect size

The driving speed effect size indicates the difference between the average value of two consecutive sections. Since the effect size represents a relative speed change among the consecutive sections. The small value of the effect size can be regarded as an easy to drive.

$$\delta_i = \frac{|\mu_i - \mu_{i-1}|}{S_p} \tag{8}$$

$$S_p = \sqrt{\frac{n_i S_i^2 + n_{i-1} S_{i-1}^2}{n_i + n_{i-1}}} \tag{9}$$

Where, δ is the effect size, μ is the average value, S_p is the pooled standard deviation for two sections, i is the section, n is the number of samples, and S is the standard deviation.

3) variance difference

The variance difference between two consecutive sections is calculated. The variance difference represents a change in the relative speed variation, and the small value of variance difference can be regarded as an easy to drive.

$$\Delta \mathrm{var} = \left|\sigma_i^2 - \sigma_{i-1}^2\right| \tag{10}$$

Where, Δvar is the variance difference, i is the section, and σ is the standard deviation.

4. Results and discussion

4.1. Data

A target area of this study is the section of about 130 km from Hatsukaichi Junction(JCT) to Kasaoka Interchange(IC) of Sanyo Expressway, observed from April to December in 2015. We used a traffic counter data, ETC 2.0 data and a road gradient data. The traffic counter is a traffic flow measuring device that accumulates the number of passing vehicles in each device by using the geomagnetic sensor. Traffic counter is installed with approximately one IC section for each direction. The traffic counter data used in this study is aggregated value per one hour. ETC 2.0 is a system with several services added to the conventional automatic toll collection (ETC) service in Japan. The system transmits information from the vehicle to the antenna on the road side. ETC 2.0 data records driving trajectory at intervals of about 200 m (100 m interval for some devices). The road gradient was calculated based on the latitude, longitude, and altitude recorded at 10 m intervals recorded. The road gradient G_i is calculated from a vertical distance V_i between a start and end points and a horizontal distance H for each section i. A positive value indicates an upward gradient, and a negative value indicates a downward gradient.

If a penetration rate of ETC2.0 probe vehicles is sufficiently high, we could use ETC2.0 dataset for each section in a link with relatively short time window. However, the number of sample is sometimes missing in short time window under the current with low penetration rate of probes. Therefore, the anomaly indices are calculated for each section of the link, without setting the time window.

4.2. Labeling of a driving speed decrease

In this study, the objective variable of the decision tree analysis is the instantaneous speed observed in the ETC2.0 data. We excluded the data of the night time when the traffic volume was very low, so then analyzed the data from 7 am to 7 pm. In addition, the analysis target is ordinary vehicles because of following two reasons. At first, the tendency of speed is different for each type of vehicle (heavy and normal vehicle). Secondly, the available number of samples is small for heavy vehicle labeled anomaly. The following six types of explanatory variables were set. 1):traffic volume 2):heavy vehicle ratio (HVR) 3):holiday dummy 4):rain dummy 5):kp (km post is the length from origin of the road per one km) 6):time in a day. The number of records used for the analysis was about 12.8 million records in nine months. The results of the decision tree analysis are shown in Fig. 1. When the t value calculated by the equation (6) is larger than the value of 1% significance with 8 degree of freedom and the sign of t is negative, the leaf is determined the speed decrease. In this analysis, since the number of categories k is set to 10, the degree of freedom can be considered to be 8. In the six leaves indicated by the red frame are detected as the occurrence of the speed decrease. Based on this results, about 2.3 million records out of about 12.8 million records are classified into speed decrease.

4.3. Relationship occurring traffic accident and anomaly indices

Estimate the occurrence factors of accidents per 1 km section by a logistic regression model using several anomaly indies proposed above (the anomaly score, the effect size, the variance difference). The objective variable is whether an accident occurred or not. The frequency of occurrences of accidents in each section is considered as a weight at the estimation. Explanatory variables are two types, such as referring to the own section and continuous section for the target section. The effect size and the variance difference is calculated for “whole”, “ordinary” and “anomaly”, and whole means all vehicle types including heavy and normal vehicles. The estimation results of the model are shown in Table 1. Model 1 is the estimation result including the explanatory variable of the two types simultaneously, and model 2 is the estimation result indicating only for the continuous section. Comparing model 1 and model 2 AIC (smaller the better), model 2 is smaller and the fit is better. Therefore, model 3 is obtained as a result of selecting the combination of the explanatory variables to give minimum AIC by stepwise method starting from model 2. In model 3, the effect size of whole and ordinary are positive, and the variance difference is negative, respectively. Considering the significant variables and signs, the factors to cause accident seems that the difference in average speed between the upstream section is large and the difference of variance from the upstream is small. In other words, the results

shown that accident tends to occur in the state that the inter-vehicle distance is large or tight state. These results also show that the accident occurrence is more affected by the difference in traffic state of the continuous section than that of the own section.

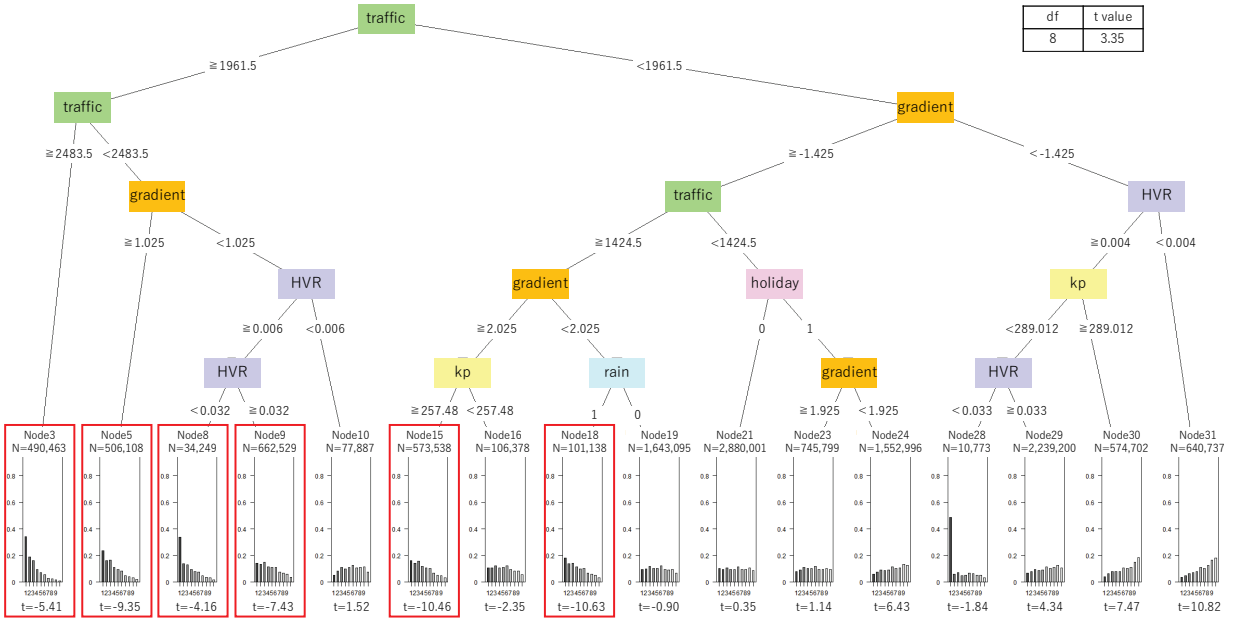

Fig. 1. The result of decision tree analysis and statistical test.

Table 1. The estimation results.

Explanatory variable		Model1			Model2			Model3		
		Coef		Std. Error	Coef		Std. Error	Coef		Std. Error
own	Anomaly score	-0.226		0.585		—			—	
	Average speed	0.021		0.065		—			—	
	Variance	-0.004		0.006		—			—	
continuous	Effect size _ whole	7.245	*	2.995	7.112	**	2.679	6.823	**	2.624
	Effect size _ ordinary	3.247		2.647	3.133		2.696	4.355	.	2.373
	Effect size _ anomaly	1.554		1.881	1.512		1.794		—	
	Δvar _ whole	-0.029	**	0.009	-0.030	**	0.008	-0.029	**	0.008
	Δvar _ ordinary	0.656	.	0.371	0.547	.	0.316		—	
intercept		0.009		6.583	1.051	*	0.410	1.162	**	0.389
AIC		190.100			184.810			183.530		
R-squared		0.169			0.164			0.159		
Likelihood Ratio Test		23.730	**		23.020	**		22.300	**	
the number of sample		103								

(.) significant at the 90% level, (*)significant at the 95% level, (**)significant at the 99% level

5. Conclusion

In this study, we proposed the anomaly detection procedure for unsupervised data by instantaneous driving speed in order to find the driving speed decrease section using ETC 2.0 data in Japan. First, the decision tree analysis with the statistical test are applied to organize the condition of driving speed decrease. In the decision tree analysis, a data is divided in to a couple of subgroups with the most different distribution of driving speed, by the conditional variables. Then, we compared the original distribution with the distribution of each leave in the decision tree in order to determine the driving speed decrease. For the determination, the Cochran-Armitage test, which tests the tendency of categorical variables of two groups, was applied. The results of the test gave the subgroups with driving speed decrease and the group with no decrease. The anomaly label was given to data satisfying the condition of the speed decrease, and the ordinary label was given to the data which did not satisfy the condition. Next, anomaly indices of each section were calculated using the distribution of driving speed. The anomaly score, the driving speed effect size and the variance difference were applied as the anomaly indices.

The logistic regression model with the accident occurrence as the objective variable was estimated. The estimation results showed that the driving speed effect size and the variance difference have a significant influence on accidents occurrence rather than the anomaly score. Therefore, we can conclude that the driving speed or its variance difference of the continuous section affects accidents more than the state of driving speed anomaly. The limitations of current study are as follows. This study analyzed by a section, not considering a time window simultaneously because of the low penetration of ETC2.0. The future work should study how to set the time window that can withstand the analysis.

Acknowledgements

This study was supported by the Committee on Advanced Road Technology (CART), Ministry of Land, Infrastructure, Transport, and Tourism, Japan.

References

Asakura, Y., Kusakabe, T., Nguyen X., Ushiki, T., 2015. Incident Detection Methods Using Probe Vehicles with On-board GPS Equipment. Transportation Research Procedia Volume 6, 17–27.

Aurenice, F., Cira, P., Paulo, O., Ana, L., 2017. Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from Brazil. Case Studies on Transport Policy Volume 5, Issue 2, 200–207.

Dibakar, S., Priyanka, A., Albert, G., 2015. Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. Accident Analysis & Prevention. Volume 79, 133–144.

Franco, B., Leonardo, J., Francisco, B., Raul, P., 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transportation Research Part C: Emerging Technologies Volume 86, 202–219.

Jinxian, W., Yang, Z., Xiaobo, Q., Xuedong Y., 2015. Development of a maximum likelihood regression tree-based model for predicting subway incident delay. Transportation Research Part C: Emerging Technologies Volume 57, 30–41.

Kinoshita, A., Takasu, A., Jun, A., 2015. Real-time traffic incident detection using a probabilistic topic model. Information Systems Volume 54, 169–188.

Lei, L., Qian, W., Adel, W., 2105. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Research Part C: Emerging Technologies. Volume 55, 444–459.

Lishuai, L., R, Hansman., Rafael, P., Roy, W., 2016. Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring Transportation Research Part C: Emerging Technologies Volume 64, 45–57.

Qi, S., Mohamed, A., 2015. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transportation Research Part C: Emerging Technologies Volume 58, Part B, 380–394.

Qinghua, L., Edward, C., Liujia, Z., 2014. Fusing moving average model and stationary wavelet decomposition for automatic incident detection: case study of Tokyo Expressway. Journal of Traffic and Transportation Engineering (English Edition) Volume 1, Issue 6, 404–414.

Quan, Y., Zhihan, L., Jinglin, L., Junming, Z., Fangchun, Y., 2014. A traffic congestion detection and information dissemination scheme for urban expressways using vehicular networks. Transportation Research Part C: Emerging Technologies Volume 47, Part 2, 114–127.

Soyoung, J., Xiao, Q., Cheol, O.,2016. Improving strategic policies for pedestrian safety enhancement using classification tree modeling. Transportation Research Part A: Policy and Practice Volume 85, 53–64.

Zhenhua, Z., Qing, H., Hanghang, T., Jizhan, G., Xiaoling, L.,2016. Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network. Transportation Research Part C: Emerging Technologies Volume 71, 284–302.