Traffic Safety Risks Trends and Patterns Analysis on Motorways

Asso Hamzehei (Corresponding author)
PhD student, Smart Transport Research Centre,
School of Civil Engineering and Build Environment
Queensland University of Technology, Gardens Point Campus
Brisbane, QLD 4001
Australia
a.hamzehei@qut.edu.au

Professor Edward Chung
Director of Smart Transport Research Centre,
School of Civil Engineering and Build Environment
Queensland University of Technology, Gardens Point Campus
Brisbane, QLD 4001
Australia
edward.chung@qut.edu.au

Dr Marc Miska
Smart Transport Research Centre,
School of Civil Engineering and Build Environment
Queensland University of Technology, Gardens Point Campus
Brisbane, QLD 4001
Australia
marc.miska@qut.edu.au

Word Count
Text: 5750
Figures and Tables (7 x 250): 1750
Total: 7500

Abstract

Crashes that occur on motorways contribute to a significant proportion (40-50%) of non-recurrent motorway congestions. Hence, reducing the frequency of crashes assist in addressing congestion issues (Meyer, 2008). Analysing traffic conditions and discovering risky traffic trends and patterns are essential basics in crash likelihood estimations studies and still require more attention and investigation. In this paper we will show, through data mining techniques, that there is a relationship between pre-crash traffic flow patterns and crash occurrence on motorways, compare them with normal traffic trends, and that this knowledge has the potentiality to improve the accuracy of existing crash likelihood estimation models, and opens the path for new development approaches. The data for the analysis was extracted from records collected between 2007 and 2009 on the Shibuya and Shinjuku lines of the Tokyo Metropolitan Expressway in Japan. The dataset includes a total of 824 rear-end and sideswipe crashes that have been matched with crashes corresponding traffic flow data using an incident detection algorithm. Traffic trends (traffic speed time series) revealed that crashes can be clustered with regards to the dominant traffic patterns prior to the crash occurrence. K-Means clustering algorithm applied to determine dominant pre-crash traffic patterns. In the first phase of this research, traffic regimes identified by analysing crashes and normal traffic situations using half an hour speed in upstream locations of crashes. Then, the second phase investigated the different combination of speed risk indicators to distinguish crashes from normal traffic situations more precisely. Five major trends have been found in the first phase of this paper for both high risk and normal conditions. The study discovered traffic regimes had differences in the speed trends. Moreover, the second phase explains that spatiotemporal difference of speed is a better risk indicator among different combinations of speed related risk indicators. Based on these findings, crash likelihood estimation models can be fine-tuned to increase accuracy of estimations and minimize false alarms.

Keywords- Traffic Flow Regimes; Traffic Flow Trends; Motorway Crashes; Risky and Normal Traffic; Clustering;

INTRODUCTION

Crashes can occur on any part of a road network. However, among different types of roads, motorways (also referred as expressways, highways, and freeways) have received more attention from governments and researchers. Motorways play an important role in the traffic networks. Motorways transport a huge number of passengers and goods between and within cities. The economies of countries depend heavily on the flow of cars in motorways with less congestion and high speed. So, a crash on a motorway could have adverse effects on both the health of people and can be detrimental to the economies. In this regard, authorities have tried to better control the motorways' traffic. Many motorways are equipped with different kinds of specialised sensors such as cameras, magnetic, infrared, microwave, laser, Bluetooth, and inductive loop detectors sensors (*1; 2*). In addition to these sensing technologies, there have been many traffic and transportation systems developed for monitoring vehicles, network traffic flows, transport infrastructure, and transport operators. The large volumes of data gathered from flow of vehicles have provided the opportunity for authorities and researchers to analyse this data and find new ways to reduce the motorway traffic risks factors as well as speed harmonisation and congestion reduction.

There is a necessity for suitable techniques to extract knowledge from large and multi-dimensional road traffic flow data. In this regard, data mining has become an active research area. Data mining, generally referred to as knowledge discovery in database (KDD), is a combination of statistical and Artificial Intelligence (AI) techniques for extraction of patterns and knowledge stored in massive databases and data repositories.

Crash related studies have been aiming to reveal influential factors that impact on motorway crashes. Traffic flow data (speed, volume, and occupancy and their variances) observed from inductive loop detectors has been the data source for such studies. Data limitation and/or methodological shortcomings resulted in contradictory findings from different studies and sometimes incompatible conclusions (*3*). In crash likelihood estimation studies, present conditions are compared to normal traffic conditions to examine crash likelihood and develop traffic safety indicators. A part of crashes caused mostly by traffic flow and traffic conditions prior to crash occurrence (risky traffic conditions). Detecting such a risky traffic conditions make it possible to avoid crashes occurrence or to reduce their severity (*4-10*).

The objectives of this study are determining risky and normal dominant traffic trends and patterns; identifying traffic regimes, investigating similarity between risky and normal traffic trends, and finding a risk indicator that distinguish crashes from normal traffic condition more precisely. In this regard, speed is selected as the main factor to observe traffic condition. A half an hour time window immediately prior to crash occurrence is selected from upstream and downstream of crash locations. The traffic situations are clustered using a non-hierarchical clustering algorithm (K-Means). In the phase one of the paper, we identified traffic regimes using upstream speed observed for 30 minutes. In the second phase, pre-crash situations clustered on three different combination of speed risk indicators and searched for unique pre-crash traffic conditions (clusters) that are not common in the normal situations.

The rest of the paper is structured as follows: second section presents a brief state of the art. Study site and data sources are explained in the third section. Methodology is presented in the section four. The section five includes the results of the study. Conclusions are given in the last section.

BACKGROUND

Studies on motorway crashes can be divided into aggregate and disaggregate studies. Aggregate studies use traffic flow data aggregated hourly or longer while disaggregate studies use minutely traffic flow data. Disaggregate studies which were mainly conducted prior to 2002, discovered a relationship between crashes and traffic conditions. For example Martin (*11*) examined the effect of traffic flow on crashes. He discovered severe crash rates are higher in light traffic conditions and crashes occur more frequently on 3 lane than 2 lane motorways.

However, in more recent disaggregate studies, Golob et al. (*12*) developed a tool to monitor traffic safety by assessing traffic flow changes in real-time. They demonstrated 21 traffic flow regimes at three different times of day and their corresponding weather conditions. As a part of their conclusion, they found that congestion strongly influences traffic safety (*12; 13*).

Zheng (*3*) shows that Crash Occurrence Likelihood (COL) is not the same in different traffic conditions. The risk of crash occurrence was less for free flow conditions while transition and congestion traffic conditions received higher COL, respectively. Zheng applied the Logit model to study the relationship between the traffic condition and crash occurrence.

The most influential factor on motorway crash occurrence is traffic states. Yeo et al (*14*) investigated the involvement of motorway crashes in four traffic states: Free Flow (FF), Back of Queue (BQ), Bottleneck Front (BN), and Congestion (CT). Traffic data is being measured for upstream and downstream detectors of crashes in order to specify the traffic states. By plotting the speed of downstream and upstream stations of a crash they segmented the crashes into the four defined traffic states. (*15*) divided freeway traffic flow into different states and investigated the safety performance regarding each state. They utilised occupancy to identify traffic states then impact of traffic flow parameters on crash occurrence evaluated in the identified traffic states.

In addition, another aspect of crash studies is studying the normal situations and mapping the crashes into the recognised regimes based on normal traffic situations. However, safety studies introduced a different definition for the normal situation. Abdel-Aty (*16*) and Pande (*17*)chose random traffic flow data from non-crash times. However, many studies defined the non-crash situation as the equivalent time and day of other weeks of each crash. It means if a crash occurred on Wednesday at 1pm, a non-crash situation for this case is other Wednesdays traffic situations at 1pm. Oh et al. (*18*) defined a non-crash situation as a 5 minute time period, half an hour before an accident occurrence. Whereas, Pham (*19*) clustered all the non-crash traffic flow data in order to identify traffic regimes and considered the traffic regimes as the non-crash situations (*4; 20; 21*).

Furthermore, Hamzehei et al (*22*) analysed 1 hour upstream speed series of pre-crash traffic situations for rearend and sideswipe crashes. They clustered one hour pre-crash speed series and found 11 dominant speed trends for crashes and categorized them into five traffic regimes. These traffic regimes were free flow, congestion, transition from free flow to congestion, transition from congestion to free flow, and unstable traffic situations. The authors extended their studies in another work by adding

downstream location in the study. Also, they considered normal situations to compare dominant risky traffic patterns with normal traffic patterns (*23*).

Although some research has been conducted on crashes in accordance with traffic states, this area of research still requires further investigation. These studies have tried to find relationships between traffic flow variables or traffic conditions and crashes just before crash occurrence (a 5 minute time window prior to the crash). In other words, the majority of literature has focused on the impacts of traffic characteristics on crash occurrence or just a particular traffic condition. There is lack of thorough research on traffic conditions that resulted in crashes. Moreover, a risk indicator is required that explains crashes more clearly and gives a unique risky pattern that is detectable from normal situations. In addition, non-crash situations have been sampled either randomly or from equivalent previous weekdays. The chosen samples are not a comprehensive representation of all the traffic situations. The defined methodologies for choosing a sample of non-crash traffic situation require further investigation to make sure they are a suitable representative of real normal conditions. As a result, in this study we aim to fill the current gap in the study of traffic condition of crashes.

STUDY SITE

The study sites are two Tokyo inner city expressways of 24 kilometres length in total which included 3180 crashes over two years (Dec. 2007- Nov. 2009). There are 201 loop detectors spread along the study site and data is available for this two year period. The data includes average speed, volume, and occupancy aggregated over the lanes into five minutes intervals. The crash dataset includes reported crash time, location of the crash, type of the crash, number of cars involved in the crash, and type of cars. In this study, we consider rear-end and sideswipe crashes. Therefore the number of vehicles involved in a crash is two or more vehicles. The severity of crashes is not considered because of crash data limitations. The accuracy of time of crashes is checked and adjusted by using incident detection algorithms.

METHODOLOGY

The objectives of this research are to understand traffic patterns and conditions that end-up with a crash by finding dominant risky traffic patterns, exploring possible relationships between pre-crash traffic conditions and crash occurrence on motorways, categorising crashes according to their pre-crash traffic flow trends, investigating the similarity between risky and normal traffic conditions, and finding risk indicators that make crashes distinct from normal situations in order to increase the accuracy of crash likelihood estimations models.

The proposed methodology (Figure 1) will be used to discover and analyse dominant risky traffic patterns in terms of different traffic risk indicators (speed and spatiotemporal difference of speed in upstream and downstream of crashes). The skeleton of the methodology is shown in Figure 1. First, loop detector data is collected from the study site. The data requires major pre-processing. Second, among the crashes, rear-end and sideswipe crashes are selected. The next step is to check the accuracy of the reported time of crash occurrence. The extracted pre-crash traffic flow data is pre-processed and the traffic speed for half an hour before crashes from closest upstream and downstream detectors are prepared for analysis. From this point, the research is divided into two phases. First, identifying risky and normal traffic trends and determining traffic regimes considering 30 minutes speed series

from upstream detector. Second, pre-crash situations clustered on 3 different combinations of risk indicators shown in Table one. Non-crash situations mapped into obtained clusters to analyse similarity of pre-crash clusters with non-crash situations. One of the objectives of this phase is to find special pre-crash clusters that do not have similar non-crash situations. The other objective is testing different risk indicators and investigating which combination of them cluster crashes in a better way that crashes be more distinguishable from non-crash situations. Furthermore, the clusters profiles are examined to check for further differences between clusters in terms of the time of crash occurrence, crash bound, and the day of the week. The K-Means algorithm applied for clustering traffic situations(*24; 25*).

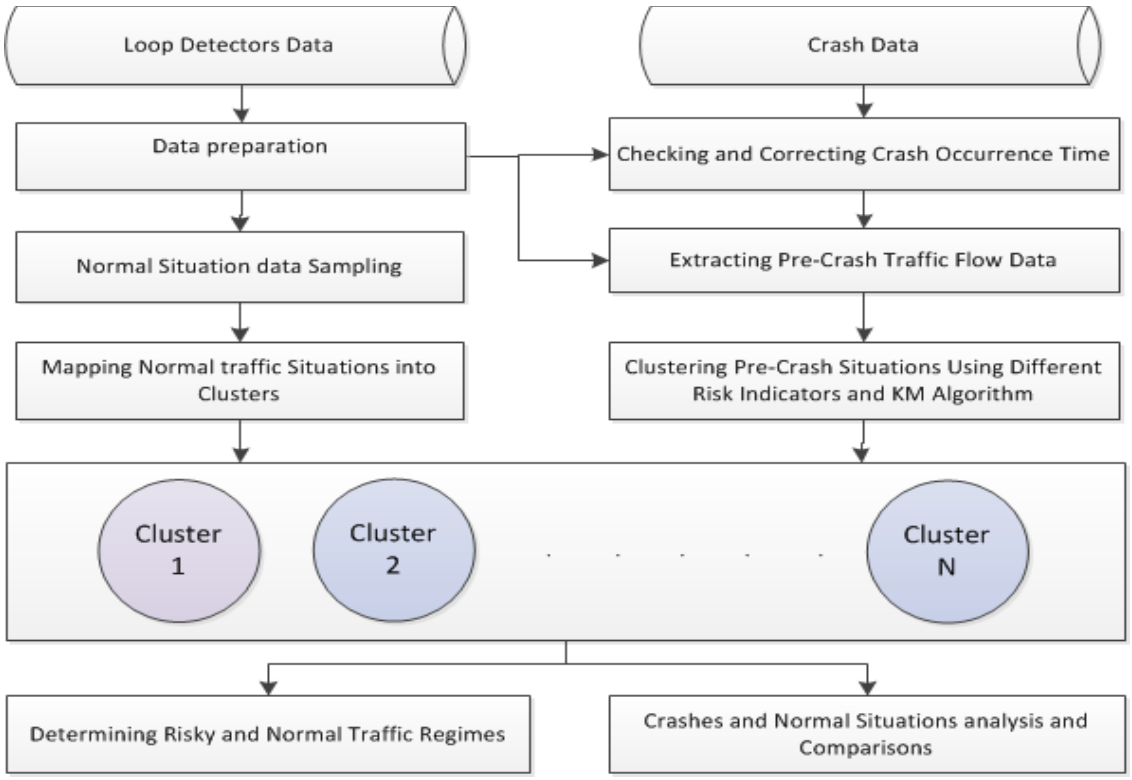


FIGURE 1 Methodology of the study

Crashes can occur due to unstable or risky traffic situations. Therefore, any variation in traffic flow variables can reveal the cause and mechanism of crash occurrence. In this regard, speed is selected to study the dynamics and changes of traffic conditions. As the objective is to discover dominant risky traffic speed patterns, the time window should be long enough to observe traffic speed fluctuations over time. The observation time period that traffic speed might have had an influence on the crash occurrence is set to half an hour. It means, for each crash, 30 minutes of traffic data prior to the crash occurrence from selected loop detectors will be extracted. However, the challenge might be why 30 minutes? Why not 45 or 60 minutes? In a previous study, the authors applied speed series with a 60 minute time window (*22*). Shortening the time window causes a few of the clusters to merge. For example, crashes in long congestion (1 hour) will be merged with the ones under shorter congestion (30 minutes). Although, shortening the time frame sacrifices some information about pre-crash traffic speed dynamics, it increases homogeneity of clusters. Short timeframes become important when normal situations are taken into account.

Data Preparation and Pre-processing

Loop detectors data randomly contain noises that may result in unreasonable values for speed, volume, and occupancy. Moreover, they might be out of order and not measure the traffic flow values. In the noisy cases, traffic flow values can be evaluated and discarded when the values for volume, occupancy, and speed are not reasonable. For example, there is a non-zero value for speed but the flow or occupancy is zero. Additionally, crashes should be checked as to whether the corresponding traffic data is available.

Moreover, crashes are reported and recorded by humans and the reported crash time might not be accurate. Crash studies require precise time of crash occurrence. Incident detection algorithms can be applied to check the accuracy of crashes reported time and find the exact crash occurrence time based on the traffic flow data (*26; 27*). In this study, an incident detection algorithm introduced by Guiyan et al(*27*) are used with some adjustments for correcting crash occurrence time.

Pre-crash and Non-Crash Traffic Situations

The traffic situation is the state of the traffic that is being measured by loop detectors. This research divides traffic situations into pre-crash and non-crash situations. A pre-crash situation refers to the traffic flow in a period of time prior to a crash in the crash location. In this research, the period of time for a traffic situation is set to 30 minutes. The traffic condition in this period of time is considered as a risky state. In addition, a non-crash situation is defined as any traffic period that does not have overlap with crash periods. In other words, non-crash situations are all traffic periods except pre-crash and post-crash periods until traffic is coming back to normal state.

Normal Situation Sampling

In crash studies, in addition to crashes corresponding traffic flow data, non-crash data is being used. Crashes are rare events on motorways and there is an imbalance between the non-crash and pre-crash situations. For example, using all the non-crash situations in crash prediction models will cause bias in the predicted value. Using all the non-crash data, the models would estimate the real time data as a non-crash situation due to over fitting models to non-crash situations. On the other hand, the non-crash situations have a large population and are not easy to handle, especially in terms of running time order. Previous studies selected the non-crash cases randomly or from equivalent time of previous weekdays of crashes.

In this study, non-crash situations from each route are sampled based on the time distribution of crashes in that specific route:

$$NTS_A = \sum_{H=0}^{23} \alpha C_H NTS_H$$

NTS_A	Non-crash Traffic Situation in route A
α	Ratio of non-crash traffic situations selection
C_H	Ratio of crashes at hour H in route A
NTS_H	Total number of non-crash situations in route A at hour H

After specifying the non-crash samples, the 30 minutes speed time series for each non-crash situation will be extracted and will be ready for the next step which is clustering pre-crash and non-crash speed time series.

Figure 2 shows the crashes distribution over daily hours in both inbound and outbound routes of Route 3 Shibuya line and Route 4 Shinjuku line. The number of extracted non-crash situations (4120) is 5 times of the number of pre-crash situations (824). These 4120 non-crash situations are distributed among daily hours for crashes as shown in Figure 2. For example, in Route 3 Shibuya line-inbound, 297 crashes have occurred with the distribution shown. Therefore, 297*5=1485 non-crash situations are extracted in that route during two years of available data in this study.

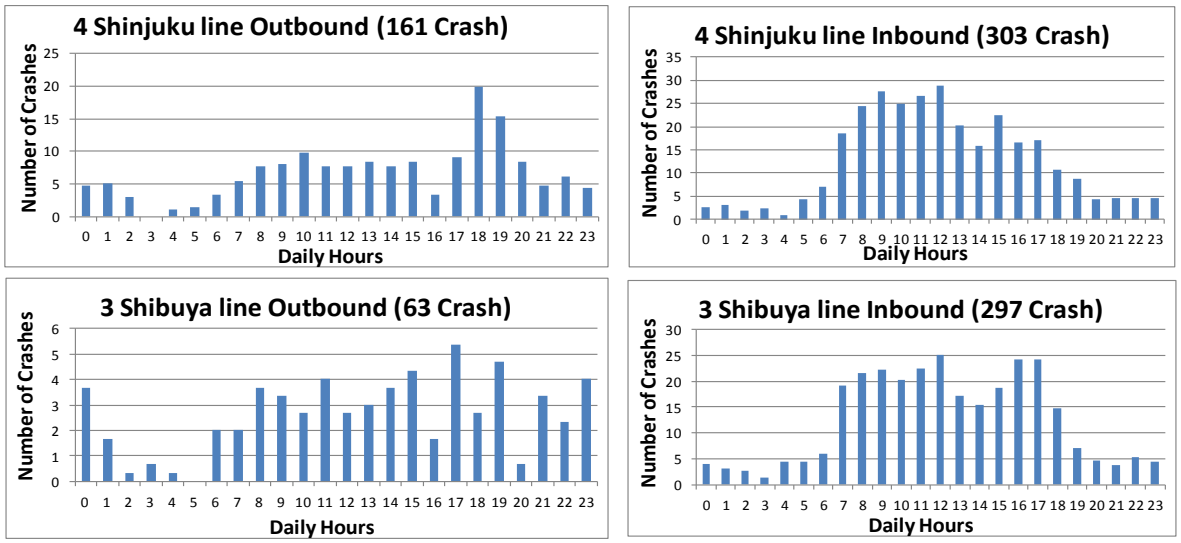


FIGURE 2 Crashes distribution at daily hours in inbound and outbound of both Shibuya and Shinjuku expressway

Risk Indicators Used in the Current Study

In this study, 30 minutes speeds (6 time frames) of upstream and downstream detectors of crashes are selected for clustering and analysis. These indicators (variables) are upstream speed of timeframe 1 to 6 (TST1-6), downstream speed of timeframe1 to 6 (DST1-6), upstream temporal speed difference of timeframes 1-2 and 2-3 (UTSDT12, UTSDT23), downstream temporal speed difference of timeframes 1-2 and 2-3 (DTSDT21, DTSDT23), upstream-downstream spatial speed difference of timeframe 1 and timeframe 2(UDSSDT1, UDSSDT2). The following table shows the usage of variable in different clustering rounds.

TABLE 1 Speed related risk indicators used in the current study

Rounds\Risk indicators	UST1	UST2	UST3	UST4	UST5	UST6	DST1	DST2	DST3	DST4	DST5	DST6	UTSDT12	UTSDT23	DTSDT21	DTSDT23	UDSSDT1	UDSSDT2
Traffic Regimes	×	×	×	×	×	×												
1st Clustering Round	×	×	×	×	×	×												
2nd Clustering Round	×	×	×	×	×	×	×	×	×	×	×	×						
3rd Clustering Round													×	×	×	×	×	×

Traffic Situations Clustering and Traffic Regimes

This research exploits the K-means clustering method to cluster traffic situations. K-means clustering is a method of clustering which aims to partition N observations into K clusters in which each observation belongs to the cluster with the nearest mean. Normal evaluation of a proper K is to minimize the inner-cluster variation and maximize the among-cluster variation. K-means clustering is sensitive to outliers; therefore outliers must be deleted before running the clustering algorithm on the data(*28; 29*). Several distance functions can be used with K-Means clustering to calculate the distance between objects. The suitable distance function in this study is the Euclidean distance function. Basically, it is the geometric distance in the multidimensional space. The following equation depicts the distance between two vectors of x and y:

$$\text{Distance}(x,y) = \sqrt{\sum(xi - yi)^2} \quad (28)$$

The obtained clusters represent different groups of risky traffic patterns. Dominant trends are frequent traffic trends which have been observed between many of speed time series. In order to recognise such trends, clusters should have a considerable number of members to be regarded as dominant trends. Despite the advantages of K-Means clustering algorithm, it cannot detect the suitable number of clusters and it should be one of the starting parameters of the KM. In this study the Dunn Index and Silhouette value is applied to determine the suitable number of clusters.

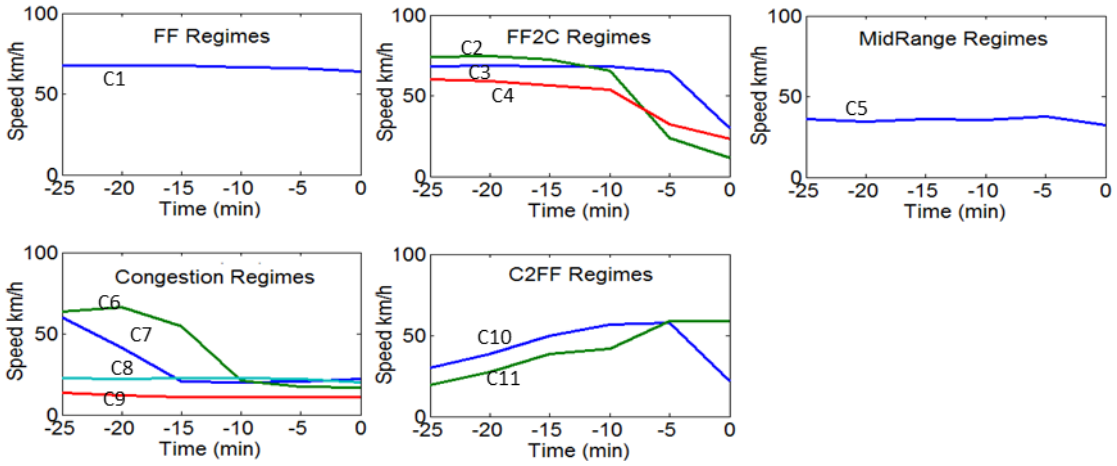
Clusters and Regimes analysis

This section of the methodology includes two phases. The first phase analyses risky and normal clusters and identify traffic regimes using 30 minutes speed series of upstream detector measured speed. Clusters can be compared by their similarities using a distance function. However, the second phase is attempting to cluster pre-crash situations on different combinations of risk indicators in order to achieve more distinguishable risky clusters from normal situations. Table 1 shows the risk indicators used in each clustering round. The first clustering round takes just 30 minutes measured speed by upstream detector into account. Traffic data are aggregated into 5 minutes level, so 30 minutes is 6 timeframe (6 speed value). The second clustering round is using 30 minutes measured speed of closest upstream and downstream. The Third clustering round is using spatiotemporal difference of measured speed at upstream and downstream locations. In each clustering round, normal situations are mapped into the clusters. Then, for each cluster we have a set of normal situations that have been the most similar ones to that cluster. For further analysis, the cluster centres for crashes and normal situations are calculated and compared.

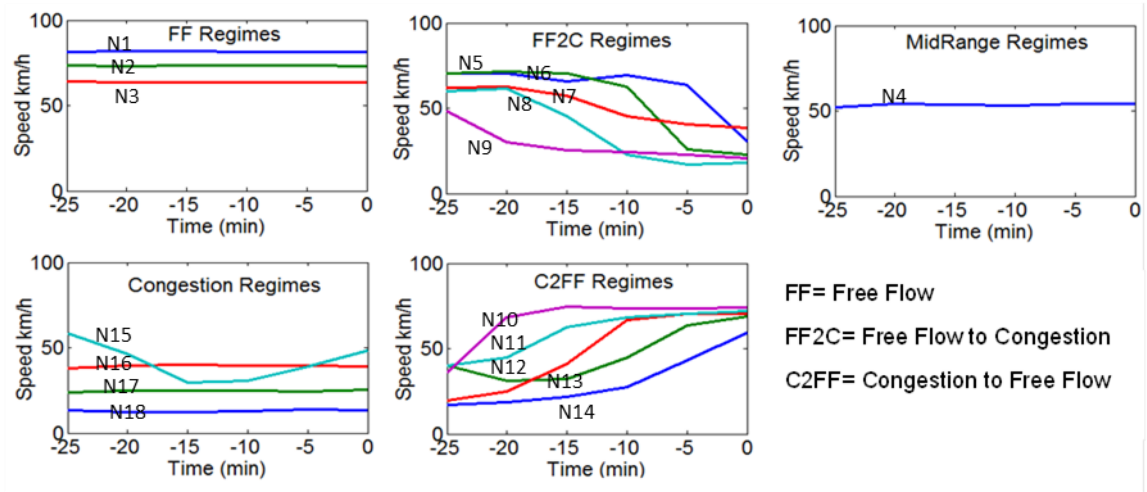
RESULTS

Pre-crash and Non-crash Traffic Situation Clustering and Traffic Regimes

Both pre-crash and non-crash traffic speed series are clustered for 2 to 20 numbers of clusters. Eleven clusters was the most appropriate number of clusters for the pre-crash speed series and eighteen clusters for non-crash speed series, respectively. Figure 3 shows pre-crash and non-crash speed series clusters. Each line in Figure 3 represents one cluster that is the cluster center. Figure 3-a shows the cluster centers for pre-crash cluster and Figure 3-b depicts non-crash cluster centers.



(a)



(b)

C= Crash clusters, N= Normal cluster

FIGURE 3 Average of speed time series of pre-crash and non-crash clusters categorized by traffic regimes

Among the clustering results from both pre-crash and non-crash clusters, five different traffic regimes are recognizable: situations where traffic was in the free flow state during the half hour prior to crashes; situations where traffic was in the free flow but changed to the congestion state; situations where traffic speed was around 50 Km/h (MidRange); situations where traffic was in the congestion condition during the 30 minutes observation window; and situations where traffic was in the congestion state but changed to the free flow state. The following explains the observed traffic regimes in the clustering results:

- FREE FLOW REGIME: this traffic regime contains 23% of crashes (cluster 1) and 70% of non-crash situations (clusters 1, 2, and 3).These four clusters have the same pattern but different in their range of speed. The speed has been constant during the 30 minutes observation for majority of the situations but varied for different situations from 60 to 90 km/h. Traffic speed for pre-crash situations (cluster 1) varies from 60 to 85 km/h. In the non-crash situations, the cluster 2 and 3 speeds are in the range of pre-crash situations but cluster 1 speed is above the speed of pre-crash situations. Free Flow regime contains 238 crashes out of 824 which

means that 29% of crashes have occurred in the free flow state. Among the weekdays, Saturday received more crashes.

- TRANSITION FREE FLOW TO CONGESTION REGIME: this regime contains traffic situations that traffic has turned from a free flow state into a congestion state. The main factor of crash occurrence is congestion in the downstream of the crash location. While traffic is in Free Flow state in upstream and suddenly downstream turns to congestion condition, traffic in the upstream faces a fast deceleration. This fast deceleration is recognized as the influential factor in crash occurrence in this traffic regime. There are three clusters (2, 3, and 4) in pre-crash situations which contain 28% of crashes. Also, there are five clusters (5, 6, 7, 8, and 9) in non-crash situations which contain 4% of total traffic situations. Moreover, the peak hour for these crashes was at 6am and 3pm and the weekday profile reveals that Sunday has received double the number of crashes than other weekdays while distribution of crashes on other weekdays is almost at the same level.
- MIDRANGE TRAFFIC REGIME: this regime refers to a traffic state that is between Free Flow and congestion state and speed is around 50 Km/h. Cluster 5 in pre-crash situations and cluster 4 in non-crash situations contains midrange traffic situations. Figure 3a and 3b show that speed in midrange clusters are different for pre-crash and non-crash situations. The non-crash midrange regime has a 50 Km/h speed average while the respective cluster in pre-crash midrange regime has a 40 Km/h speed average.
- CONGESTION: this regime refers to a situation where the traffic state is in a congestion situation. This regime is the biggest among the pre-crash situations having 34% of all pre-crash situations and four clusters (6, 7, 8, and 9) that belong to the congestion regime. Also, four clusters 15, 16, 17 and 18 of non-crash situations belong to this traffic regime by having 17% percent of all non-crash situations. Cluster 6 and 7 of pre-crash situations are carrying crashes that traffic has been in free flow condition until 15 to 10 minutes before the crash time. The rest of the clusters in both pre-crash and non-crash situations have been in a congestion condition during the 30 minute observation window. Fatigue and tiredness of drivers during too much deceleration and acceleration may be one of the possible reasons for crashes in this regime. Among the weekdays, Friday received the most number of crashes in the Congestion regime. Moreover, peak times for crashes in this regime are 12pm and 6pm
- TRANSITION CONGESTION TO FREE FLOW: this regime contains traffic situations that traffic has turned from the congestion state into the free flow state. The main factor of crash occurrence is fluctuation of traffic speed during traffic returning to the free flow state from a congestion state. There are two clusters (10 and 11) in pre-crash situation pertaining to 4% of all crashes. Also, there are five clusters (10, 11, 12, 13, and 14) in non-crash situations which contain 4% of total traffic situations. Moreover, the peak hour for these crashes was at 6am and 3pm and weekday profile reveals that Sunday has received double the number of crashes than other weekdays whilst distribution of crashes on other weekdays are almost at the same level.

TABLE 2 Pre-crash clustering and non-crash situations mapped into the clusters- numbers and percentages of members in each cluster

Cluster Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1st Clustering Round – Upstream speed (6 timeframes= 30 minutes)											
PCS involvement	108	**133**	62	114	26	86	32	106	87	50	20
NCS Mapped into Clusters	2296	**20**	42	805	21	247	45	154	244	150	96
PCS involvement percent	13.1	**16.1**	7.5	13.8	3.2	10.4	3.9	12.9	10.6	6.1	2.4
NCS members percent	55.7	**0.5**	1.0	19.5	0.5	6.0	1.1	3.7	5.9	3.6	2.3
2nd Clustering Round – Upstream and Downstream speed (6 timeframes= 30 minutes)											
PCS involvement	116	**103**	27	43	72	76	36	**48**	113	152	38
NCS Mapped into Clusters	2383	**29**	37	70	149	247	37	**22**	281	677	188
PCS involvement percent	14.1	**12.5**	3.3	5.2	8.7	9.2	4.4	**5.8**	13.7	18.4	4.6
NCS members percent	57.8	**0.7**	0.9	1.7	3.6	6.0	0.9	**0.5**	6.8	16.4	4.6
3rd Clustering Round – Spatiotemporal difference of speed (2 timeframes= 10 minutes)											
PCS involvement	296	**37**	48	**31**	**67**	15	88	**69**	23	**64**	86
NCS Mapped into Clusters	3065	**7**	107	**2**	**8**	42	450	**23**	20	**4**	392
PCS involvement percent	35.9	**4.5**	5.8	**3.8**	**8.1**	1.8	10.7	**8.4**	2.8	**7.8**	10.4
NCS members percent	74.4	**0.2**	2.6	**0.0**	**0.2**	1.0	10.9	**0.6**	0.5	**0.1**	9.5

PCS= Pre-Crash Situation, NCS= Non-Crash Situation

TABLE 3 Distance of pre-crash clusters centres from the mapped non-crash centers (normalized between 0 to 1)

Crash\Normal	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
C1	0.00	0.52	0.27	0.55	0.31	0.34	0.13	0.53	0.49	0.42	0.20
C2	0.58	0.08	0.58	0.87	0.33	0.75	0.67	0.97	0.57	0.58	0.51
C3	0.29	0.44	0.10	0.54	0.36	0.43	0.33	0.64	0.59	0.57	0.36
C4	0.55	0.81	0.51	0.13	0.62	0.65	0.44	0.38	0.82	0.59	0.74
C5	0.41	0.39	0.50	0.52	0.15	0.67	0.41	0.62	0.42	0.29	0.46
C6	0.39	0.67	0.44	0.62	0.59	0.22	0.37	0.58	0.87	0.50	0.54
C7	0.19	0.64	0.33	0.38	0.39	0.40	0.04	0.33	0.57	0.38	0.40
C8	0.60	1.00	0.70	0.34	0.73	0.71	0.45	0.06	0.87	0.54	0.80
C9	0.55	0.57	0.64	0.69	0.33	0.88	0.57	0.77	0.16	0.57	0.52
C10	0.44	0.54	0.57	0.46	0.32	0.64	0.39	0.51	0.57	0.14	0.55
C11	0.20	0.45	0.36	0.72	0.28	0.48	0.34	0.72	0.37	0.52	0.04

C1= pre-Crash traffic situations in cluster1

N1 = Normal traffic situations mapped into Cluster1

Speed Related Risk Indicators and Traffic Situations Clustering

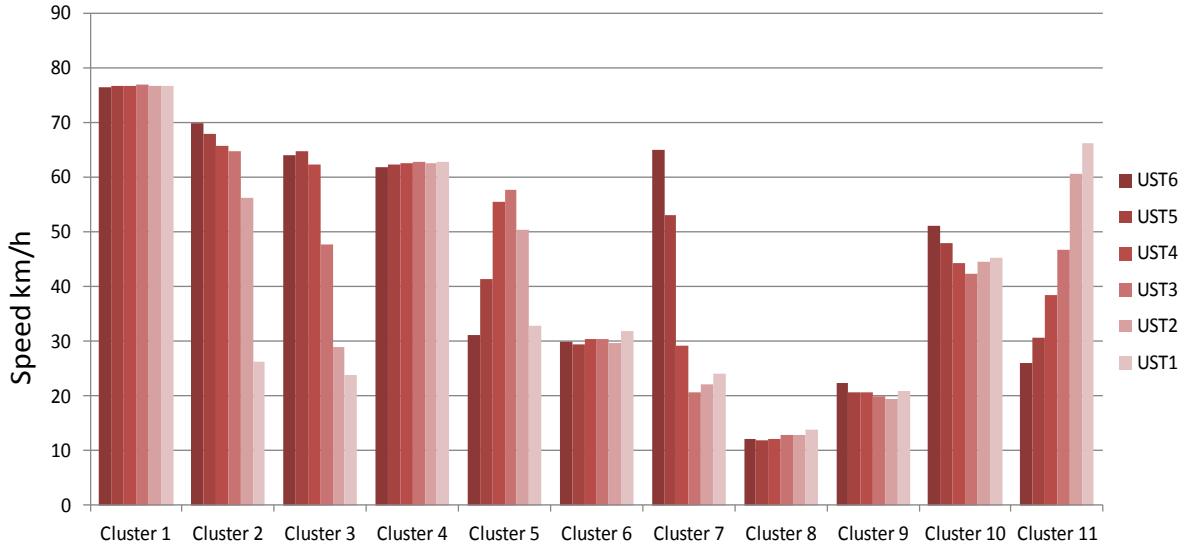
As mentioned in the methodology, the objective of this phase (second phase) is clustering on different combinations of risk indicators to achieve unique and specific risky clusters where there are no or few normal situations that can be paired with them. Table 1 shows the risk indicators that applied in the three different clustering rounds. All the rounds are clustered on 11 numbers of clusters. K-Means by default choose K (K is number of clusters) initial seeds randomly among the available points of clustering which results in having clusters in a random order in each round. The points

closest to the centre of clusters in phase one are selected as the initial seeds for the clustering rounds one to three. The reason for selecting initial seeds is that we want clusters in the same order as they are in the identified regimes in the phase one.

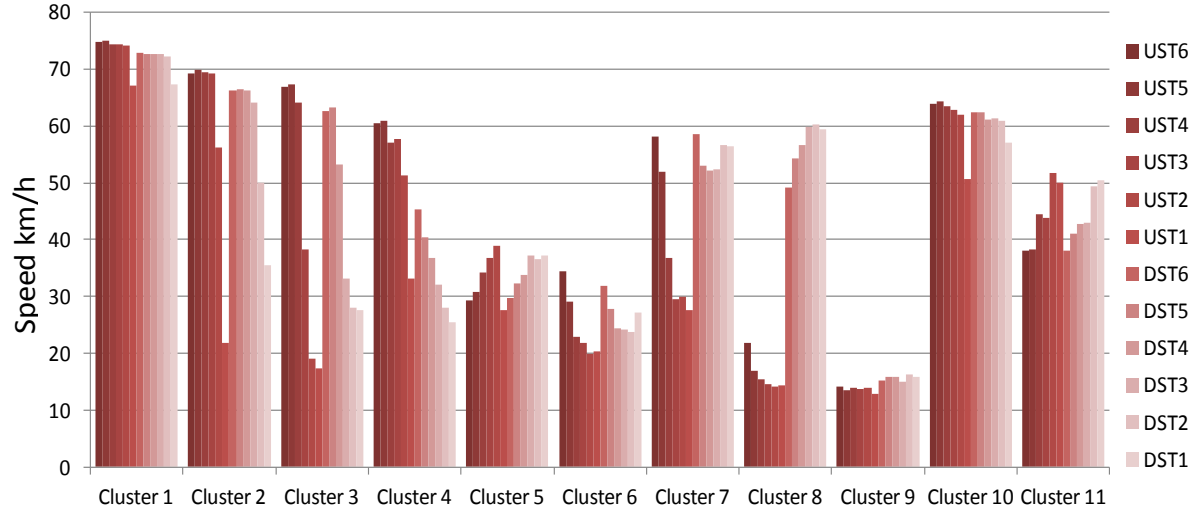
In Each Round of clustering, pre-crash situations are clustered. Then, normal situations are compared one by one with the cluster centers and assigned to the cluster with the least amount of distance. The assigned cluster is supposed to be the most similar cluster to that specific normal situation. Table 2 shows the pre-crash clusters members and number of normal situations assigned to each cluster. The results in Table 2 shows that spatiotemporal difference of speed is the most suitable risk indicators among the ones listed in Table 1. According to 3rd clustering round (clustering on spatiotemporal difference of speed), 32% of total crashes in clusters C2, C4, C5, C8, and C10 have received just 1% of normal situations assigned into them. It means these clusters can be considered as very risky traffic situations as there have not been many normal situations paired to them. These five clusters occur in very unstable traffic situations where traffic flow and speed have been fluctuating between the upstream and downstream of the crash location. Moreover, clusters C3, C6, and C9 that include 10.5% of all crashes have received 4% of normal situations. Table 3 shows the distance matrix of centroids of pre-crash clusters and centroids of assigned normal situations into the clusters. Distance of assigned normal situations into clusters C3, C6, and C9are more than 0.1 which means these 4% assigned normal situations cannot be counted as similar traffic situations to crashes. Therefore, 42.5% of clustered crashes have occurred in a special traffic situations. These crashes theoretically, are assumed to be detectable by crash likelihood estimation models as they are far from normal situations in terms of Euclidean distance. Cluster one represents crashes where traffic speed has been constant without fluctuation (Figure 4c). This cluster carries 35.9% of all crashes and 74.4% of normal situations. Pre-crash and non-crash situations inside this cluster are the most similar traffic situations among all other clusters and their paired normal situations. Also, cluster C7 has received 10.7% of crashes and 10.9% of normal situations respectively. Traffic in this cluster has been constant in both upstream and downstream but the speed level in upstream was 15 km/h higher from 5 to 10 minutes before crashes.

Figure 4 depicts mean of variables in the eleven clusters for the three clustering rounds. Figures 4a, 4b, and 4c represent the speed level of traffic situations in each cluster for clustering round one to three. The Figure 4c shows that the spatiotemporal difference of speed are high in the unique clusters like C2, C3, C4, C5, C6, C8, C9 , andC10. However, in the clusters C1, C7, and C11 that have received 95% of normal situations, the spatiotemporal difference of speed has been very low.

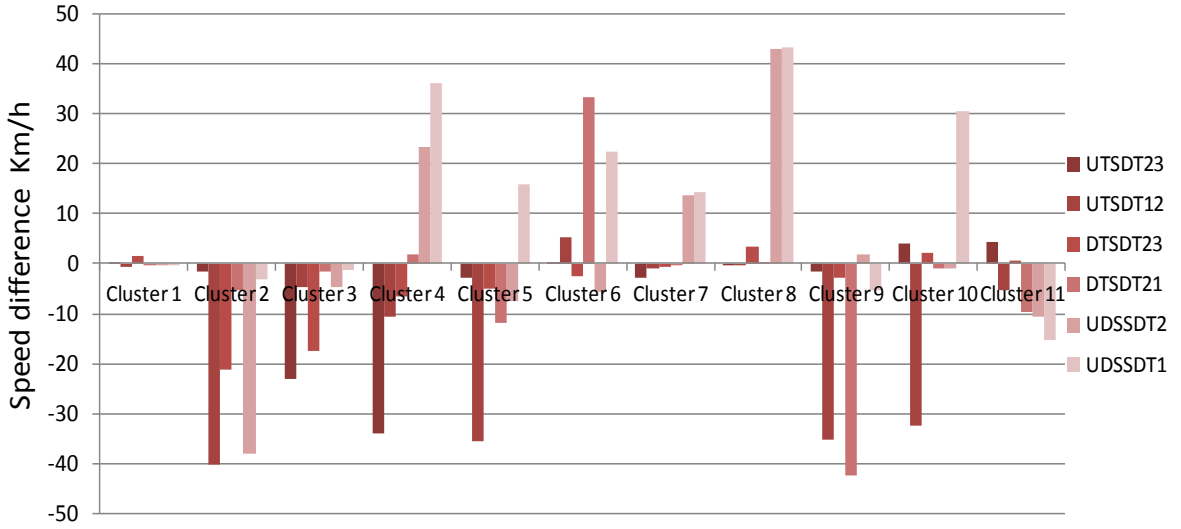
The results show that the spatiotemporal difference of speed outperforms other speed related risk indicators in terms of distinguishing risky situations from normal situations. On the other hand, other risk indicators that taken into account in this study was good explainers of the traffic regimes and different traffic conditions prior to crashes occurrence. Among crashes, 42.5% have been recognized as crashes that have occurred in very unique and high risk conditions.



(a) First clustering round on upstream speed of 6 timeframe(30 minutes)



(b) Second clustering round on 6 timeframes (30 minutes) of upstream and downstream speed



(C) Third clustering round on spatiotemporal difference of speed in upstream and downstream

FIGURE 4 Bar charts of variables values for the eleven clusters in the three clustering rounds

6. Conclusion

This paper studied pre-crash and non-crash traffic situations in two phases. First, pre-crash and non-crash traffic situations are clustered to find dominant risky and normal traffic trends and traffic regimes. In the second phase, pre-crash situations clustered on three different combinations of speed related risk indicators and normal situations assigned into pre-crash clusters to investigate similarity of extracted risky clusters with the normal situations. An ideal risky cluster is the one that has no overlap with normal situations. These risky patterns are important in Crash Occurrence Likelihood (COL) estimation studies and help to increase the accuracy of risk detection in real-time. In phase one, speed time series clustered using non-hierarchical clustering algorithm (K-Means) and Dunn index and Silhouette value are used to find the optimal number of clusters which was 11 clusters for risky situations and 18 clusters for normal situations. Among the both risky and normal traffic clusters, five major traffic regimes recognized: free flow, transition from free flow to congestion, midrange traffic, congestion, and transition from congestion to free flow. In phase two, results show that spatiotemporal difference of speed has been a better risk indicator for clustering pre-crash situations as 42.5% was almost unique in terms of comparing with normal situations. Building on the current study, future research can investigate other risk indicators other than speed on crash likelihood. In addition, the results of this paper can be used in crash estimation modeling and checking the accuracy of estimation models with and without having clustered risky and normal situations.

References

[1] Lawrence A. Klein, M. K. M., David R.P. Gibson. Traffic Detector Handbook: Third Edition—Volume I.In*, No. 1*,
Turner-Fairbank Highway Research Center, Federal Highway Administration, McLean, VA, 2006. p. 288.

[2] Bhaskar, A., L. M. Kieu, M. Qu, A. Nantes, M. Miska, and E. Chung. On the use of Bluetooth MAC Scanners for live reporting of the transport network. Presented at 10th International Conference of Eastern Asia Society for Transportation Studies, Taipei, Taiwan, 2013.

[3] Zheng, Z. Empirical Analysis on Relationship between Traffic Conditions and Crash Occurrences. *Procedia - Social and Behavioral Sciences,* Vol. 43, No. 0, 2012, pp. 302-312.

[4] Pham, M. H., A. Bhaskar, E. Chung, and A. G. Dumont. Towards a pro-active model for identifying motorway traffic risks using individual vehicle data from double loop detectors.In *Road Transport Information and Control Conference and the ITS United Kingdom Members' Conference (RTIC 2010) - Better transport through technology, IET*, 2010. pp. 1-9.

[5] Abdel-Aty, M., N. Uddin, and A. Pande. Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1908, No. -1, 2005, pp. 51-58.

[6] Hourdos, J., V. Garg, P. Michalopoulos, and G. Davis. Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1968, No. -1, 2006, pp. 83-91.

[7] Lee, C., M. Abdel-Aty, and L. Hsia. Potential Real-Time Indicators of Sideswipe Crashes on Freeways. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1953, No. -1, 2006, pp. 41-49.

[8] Hossain, M., and Y. Muromachi. Understanding Crash Mechanisms and Selecting Interventions to Mitigate Real-Time Hazards on Urban Expressways. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2213, No. -1, 2011, pp. 53-62.

[9] Pande, A., A. Das, M. Abdel-Aty, and H. Hassan. Estimation of Real-Time Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2237, No. -1, 2011, pp. 60-66.
[10] Pande, A., and M. Abdel-Aty. Multiple-Model Framework for Assessment of Real-Time Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2019, No. -1, 2007, pp. 99-107.
[11] Martin, J.-L. Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention,* Vol. 34, No. 5, 2002, pp. 619-629.
[12] Golob, T. F., W. W. Recker, and V. M. Alvarez. Freeway safety as a function of traffic flow. *Accident Analysis & Prevention,* Vol. 36, No. 6, 2004, pp. 933-946.
[13] Golob, T. F., and W. W. Recker. A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A: Policy and Practice,* Vol. 38, No. 1, 2004, pp. 53-80.
[14] Yeo, H., K. Jang, A. Skabardonis, and S. Kang. Impact of traffic states on freeway crash involvement rates. *Accident Analysis & Prevention*, No. 0, 2012.
[15] Xu, C., P. Liu, W. Wang, and Z. Li. Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention,* Vol. 47, No. 0, 2012, pp. 162-171.
[16] Abdel-Aty, M., and A. Pande. ATMS implementation system for identifying traffic conditions leading to potential crashes. *Intelligent Transportation Systems, IEEE Transactions on,* Vol. 7, No. 1, 2006, pp. 78-91.
[17] Pande, A., and M. Abdel-Aty. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention,* Vol. 38, No. 5, 2006, pp. 936-948.
[18] Oh, J. S., C. Oh, S. G. Ritchie, and M. Chang. Real-time estimation of accident likelihood for safety enhancement. *Journal of Transportation Engineering-Asce,* Vol. 131, No. 5, 2005, pp. 358-363.
[19] Pham, M. H. Motorway Traffic Risks Identification Model - MyTRIM Methodology and Application.In *ENAC School of Architecture, Civil and Environmental Engineering, No. Doctorate*, École polytechnique fédérale de Lausanne, Lausanne 2011. p. 169.
[20] Pham, M. H., N. E. El Faouzi, and A. G. Dumont. Real-time identification of risk-prone traffic patterns taking into account weather conditions.In *90th annual meeting of Transportation Research Board, Washington, DC*, 2011.
[21] Pham, M. H., A. Bhaskar, E. Chung, and A. G. Dumont. Methodology for Developing Real-Time Motorway Traffic Risk Identification Models Using Individual-Vehicle Data.In *Transportation Research Board 90th Annual Meeting*, 2011.
[22] Hamzehei, A., E. Chung, and M. Miska. Pre-Crash Traffic Flow Trend Analysis on Motorways. Presented at OPTIMUM 2013 – International Symposium on Recent Advances in Transport Modelling, kingscliffe, Australia, 2013.
[23] Hamzehei, A., E. Chung, and M. Miska. Pre-Crash and Non-CrashTraffic Flow Trend Analysis On Motorways. Presented at Australasian Transport Research Forum 2013 Proceedings, Brisbane, Australia, 2013.
[24] Hamzehei, A., H. Farvaresh, M. Fathian, and M. Gholamian. A new methodology to study customer electrocardiogram using RFM analysis and clustering. *Management Science Letters,* Vol. 1, No. 2, 2011.
[25] Kieu, L. M., A. Bhaskar, and E. Chung. Mining temporal and spatial travel regularities for transit planning. Presented at Australasian Transport Research Forum 2013, Queensland University of Technology, Brisbane, QLD, 2013.
[26] Jin, J. P. Automatic incident detection based on fundamental diagrams of traffic flow.In, The University of Wisconsin - Madison, 2009. p. 153.
[27] Guiyan, J., N. Shifeng, L. Qi, C. Ande, and J. Hui. Automated incident detection algorithms for urban expressway.In *Advanced Computer Control (ICACC), 2010 2nd International Conference on, No. 3*, 2010. pp. 70-74.

[28] Han, J., and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann Publishers, 2006.
[29] Witten, I. H., and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.