



Queensland University of Technology
Brisbane Australia

This may be the author's version of a work that was submitted/accepted for publication in the following source:

[Hamzehei, Asso, Chung, Edward, & Miska, Marc](#)
(2013)

Pre-crash and non-crash traffic flow trends analysis on motorways.
In O'Keefe, B (Ed.) *Australasian Transport Research Forum 2013 Proceedings*.

Australasian Transport Research Forum, Australia, pp. 1-14.

This file was downloaded from: <https://eprints.qut.edu.au/63219/>

© Copyright 2013 Queensland University of Technology and the Authors

This work is covered by copyright. Unless the document is being made available under a Creative Commons Licence, you must assume that re-use is limited to personal use and that permission from the copyright owner must be obtained for all other uses. If the document is available under a Creative Commons License (or other specified license) then refer to the Licence for details of permitted re-use. It is a condition of access that users recognise and abide by the legal requirements associated with these rights. If you believe that this work infringes copyright please provide details by email to qut.copyright@qut.edu.au

Notice: *Please note that this document may not be the Version of Record (i.e. published version) of the work. Author manuscript versions (as Submitted for peer review or as Accepted for publication after peer review) can be identified by an absence of publisher branding and/or typeset appearance. If there is any doubt, please refer to the published source.*

<http://atrconference2013.com/>

Pre-Crash and Non-Crash Traffic Flow Trends Analysis on Motorways

Asso Hamzehei, Edward Chung, and Marc Miska

Smart Transport Research Centre, Queensland University of Technology, Australia

a.hamzehei@qut.edu.au

Abstract

Crashes that occur on motorways contribute to a significant proportion (40-50%) of non-recurrent motorway congestions. Hence, reducing the frequency of crashes assists in addressing congestion issues (Meyer, 2008). Crash likelihood estimation studies commonly focus on traffic conditions in a short time window around the time of a crash while longer-term pre-crash traffic flow trends are neglected. In this paper we will show, through data mining techniques that a relationship between pre-crash traffic flow patterns and crash occurrence on motorways exists. We will compare them with normal traffic trends and show this knowledge has the potential to improve the accuracy of existing 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 to 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. Using the K-Means clustering method with Euclidean distance function allowed the crashes to be clustered. Then, normal situation data was extracted based on the time distribution of crashes and were clustered to compare with the "high risk" clusters. Five major trends have been found in the clustering results for both high risk and normal conditions. The study discovered traffic regimes had differences in the speed trends. Based on these findings, crash likelihood estimation models can be fine-tuned based on the monitored traffic conditions with a sliding window of 30 minutes to increase accuracy of the results and minimize false alarms.

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

1. 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 as arteries do in the human bodies' blood vessel 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 on 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 a number of different kinds of specialised sensors such as cameras, magnetic sensors, infrared sensors, microwave sensors, laser sensors, and inductive loop detectors (Lawrence A. Klein, 2006). 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 observed from inductive loop detectors has been the data source for such studies. Therefore, traffic flow variables such as speed, volume, and occupancy and their variances are analysed to discover their relationships with crash occurrence. Data limitation and/or methodological shortcomings resulted in contradictory findings from different studies and sometimes incompatible conclusions. In crash estimation studies, present conditions are compared to normal traffic conditions to examine crash likelihood and develop traffic safety indicators. Therefore, in addition to crashes corresponding traffic flow data, traffic flow data from non-crash situations should be extracted in order to distinguish a non-crash situation from a risky situation in real-time.

In this study, loop detector data of the study area and crash dataset with detailed information about crashes are used. The first objective of this paper is to find risky and normal dominant traffic flow patterns. The second objective is investigating similarity between risky and normal traffic regimes. In this regard, speed is selected as the main factor to observe traffic conditions. A half hour time window immediately prior to crash occurrence is selected. So, for each crash one speed series is available. The speed series contain information of pre-crash traffic conditions. The series are clustered using a non-hierarchical clustering algorithm (K-Means) to cluster different pre-crash speed series. Normal condition data is extracted as 30 minutes speed series and clustered to discover normal traffic flow patterns. Then, both the risky and normal traffic clusters are compared in terms of similarity between them.

2. 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 conditions. For example Martin (2002) 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. (2004) 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 (Golob & Recker, 2004; Golob, Recker, & Alvarez, 2004).

Zheng (2012) shows that Crash Occurrence Likelihood (COL) is not the same in different traffic conditions. The risk of crash occurrence was less for free flow condition while transition and congestion traffic condition 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, Jang, Skabardonis, & Kang, 2012). Yeo et al (2012) 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.

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(2006) and Pande (2006) 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. (2005) defined a non-crash situation as a 5 minute time period, half an hour before an accident occurrence. Whereas, Pham (2011) clustered all the non-crash traffic flow data in order to identify traffic regimes and considered the traffic regimes as the non-crash situations (M. H. Pham, Bhaskar, Chung, & Dumont, 2010; M.H. Pham, El Faouzi, & Dumont, 2011) .

Although some research has been conducted on crashes in accordance with traffic states, this area of research still requires further investigation. In the literature of traffic conditions associated with crashes, traffic flow is commonly considered only around the time of crash occurrence. 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, crash likelihood estimation studies generally focus on traffic conditions in a short time window and longer-term pre-crash traffic flow trends are neglected. 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.

3. Study site

The study sites are two Tokyo inner city expressways of 24 kilometres length in total which included 3180 crashes over two years (2007-09). 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. The accuracy of time of crashes is checked and adjusted by using incident detection algorithms.

4. Methodology

The objective of this research is to understand traffic patterns that end-up with a crash, by finding dominant normal and risky traffic patterns, exploring possible relationships between pre-crash traffic flow patterns and crash occurrence on motorways, increasing the accuracy of crash likelihood estimations, categorising crashes according to their pre-crash traffic flow trends, and investigating the similarity between risky and normal traffic regimes.

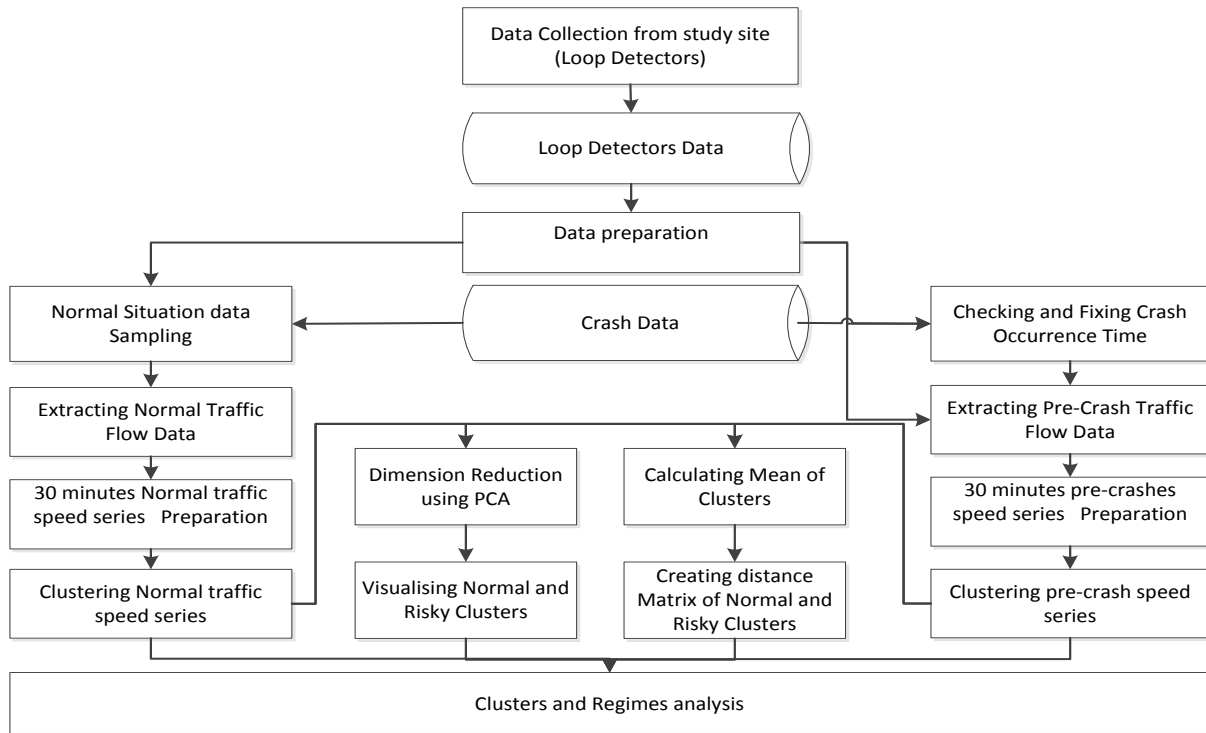


Figure 1 Methodology of the study

The methodology proposed is shown in Figure 1, and will be used to discover and analyse dominant normal and risky traffic patterns. 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 speed series for half an hour before crashes are prepared for analysis. The speed series of crashes are clustered using the K-Means algorithm to discover the existence of dominant trends prior to crash occurrence on motorways. 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. At this point, the risky clusters are obtained. The next step includes finding normal traffic patterns to be investigated along with risky traffic patterns. In this regard, normal traffic data is sampled and extracted. Thirty minute speed series of normal traffic situations are clustered. Consequently, in the last step, both normal and risky traffic clusters will be analysed and compared in terms of similarity.

4.1 Traffic situation

The traffic situation is the state of the traffic that is being measured by traffic flow detectors. The aim of the traffic situation is to explain the safety level of the traffic condition in a specific road section and time period. This research divides traffic situations into pre-crash situations and non-crash situations. As Figure 2 shows, the crash period includes a period of time before and after crash occurrence and post-crash is a period of time that the traffic state is coming back to normal following a crash.

4.1.1 Pre-Crash Situation

A pre-crash situation refers to the traffic state in a period of time prior to a crash in the crash location. ; In this research, the period of time is set to 30 minutes. The traffic condition in this period of time is considered as a risky state.

4.1.2 Non-crash situation

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 (figure 2).

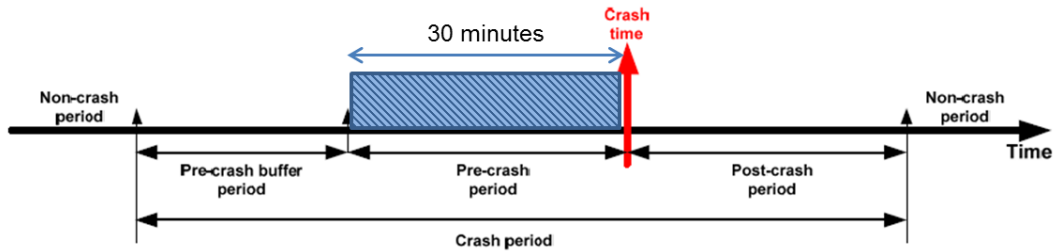


Figure 2 Crash and non-crash time periods

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 flow 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 minutes time window (Hamzehei, Chung, & Miska, 2013). 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.

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 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, accidents should be checked as to whether the traffic data is available. The crashes where their corresponding traffic data is unavailable or noisy would be discarded.

4.2 Crash Occurrence Time Accuracy

As crashes are reported and recorded by humans, the crash occurrence time might not be accurate. Crash studies require precise time of crash occurrence. For instance, in crash likelihood estimation studies, traffic flow data is used to develop models for predicting crashes. Without knowing the crash time, models cannot be trained to estimate crash likelihood. Incident detection algorithms can help to check the accuracy of crashes and find the exact crash occurrence time based on the traffic flow data. When an incident happens, the road partially or fully becomes blocked. In any kind of road with different numbers of lanes, there are traffic pattern changes for upstream and downstream road sections of the incident location. This post incident trend is expected to continue for a period of time until the road is cleared. In the upstream of the incident, fewer cars can pass the incident location. Therefore, it is expected that the occupancy for the upstream section increases rapidly, while speed and flow decrease. On the other hand, flow and occupancy decrease and speed increases in the downstream section of the road (Guiyan, Shifeng, Qi, Ande, & Hui, 2010; Jin, 2009). This concept is used in the current study to check and adjust accuracy of crashes reported time.

4.3 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

For example, if the number of non-crash situations per crash is chosen as 5 (α), route A contains 10% of crashes that happened at 8am, and non-crash situations contains 1000 sample, then we would have ($5 * 10\% * 1000$) 500 non-crash situations for 8am for route A of the whole study period. The same methodology will be applied for other hours for all the routes.

After specifying the non-crash samples, the 30 minute 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.

4.4 Speed Time-Series Clustering and Traffic Regimes

This research exploits the K-means clustering method to cluster extracted 30 minute pre-crash speed series. 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 (Han & Kamber, 2006; Witten & Frank, 2005). 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:

$$\text{Distance}(x,y) = \sqrt{\sum (x_i - y_i)^2} \text{ (Han \& Kamber, 2006)}$$

The obtained clusters represent different groups of risky traffic patterns. Dominant trends are frequent fluctuation trends which have been observed between many speed time series. In order to recognise such trends, clusters should have a considerable number of members to be regarded as dominant trends. The number of clusters is set by Dunn index. The obtained clusters are analysed further by creating their profiles to investigate the common similarities inside each cluster.

4.5 Clusters and Regimes analysis

This section of the methodology analyses risky and normal clusters with each other. Clusters can be compared by their similarities using a distance function. In both risky and normal categories, the mean of speed series in each cluster will be calculated. Distance of the mean of each cluster will be calculated from other risky and normal clusters. It will result in a matrix that contains all the distances between clusters.

Visualising clusters will help in finding meaningful differences between risky and normal clusters. As each time series has 6 timeframes, the clusters have 6 dimensions. In this study we use PCA to represent clusters by their first two principal components in a 2D diagram. In other words, the six dimensions will be reduced to two dimensions

5. Results

5.1 Non-crash Traffic Situation Sampling

As discussed in the methodology, there is imbalanced data among risky and normal situations. To overcome this problem, the normal data is sampled. Figure 3 shows the crash 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 3. For example, in Route 3 Shibuya line-inbound, 297 crashes have occurred with the shown distribution. Therefore, $297 \times 5 = 1485$ non-crash situations are extracted in that route during two years of available data in this study.

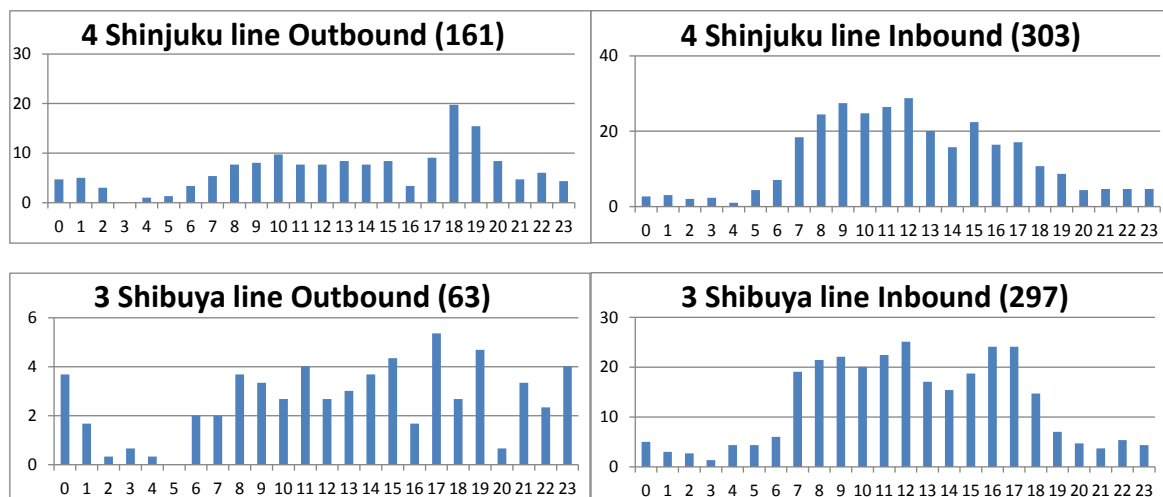


Figure 3 Crash distribution in daily hours

5.2 Pre-crash and Non-crash Traffic Speed Series Clustering

K-Means clustering with Euclidean distance function are the final candidates for clustering the pre-crash and non-crash traffic speed series. 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 is applied to find the suitable number of clusters. Both pre-crash and non-crash traffic speed series are clustered for 2 to 30 numbers of clusters. The pre-crash speed series received the highest value of Dunn Index for 11 clusters and non-crash speed series received the highest value of Dunn Index for 18 clusters, respectively. Figure 4 shows pre-crash speed series clusters and Figure 5 shows non-crash speed series clusters.



Figure 4 Pre-crash traffic speed series clusters

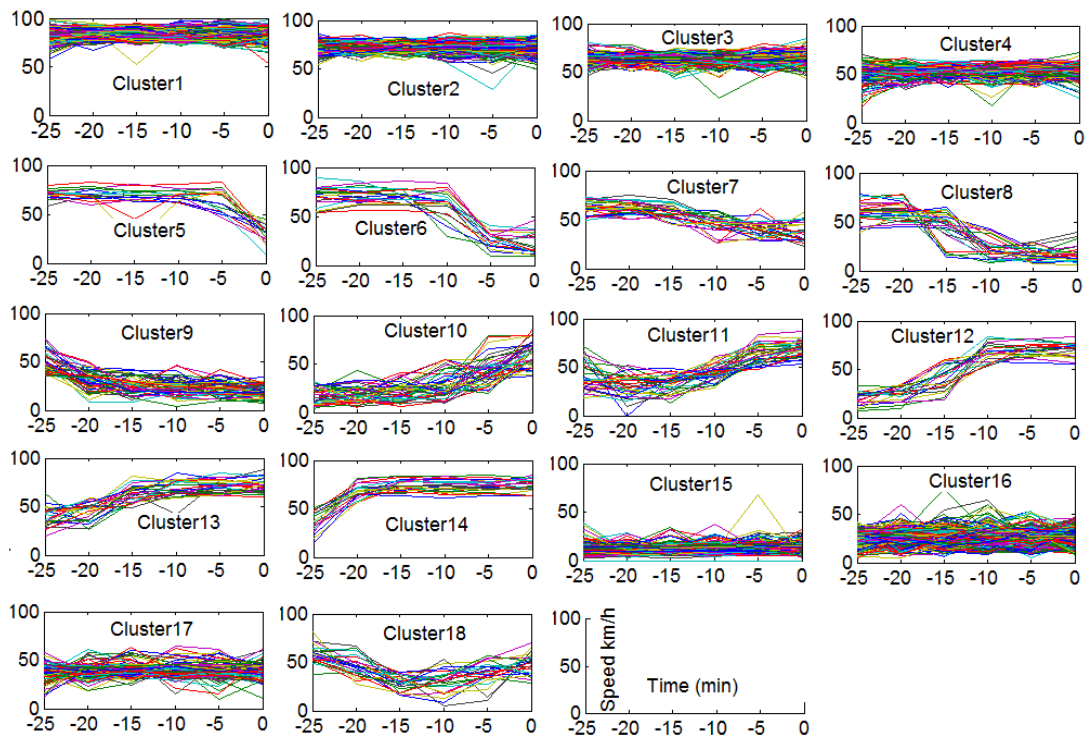


Figure 5 Non-crash traffic speed series clusters

Among the clustering results from both pre-crash and non-crash clusters, five different traffic regimes are recognizable: situations where traffic was in free flow state during the half hour prior to crashes; situations where traffic was in free flow but changed to congestion; situations where traffic speed was around 50 Km/h (MidRange); situations where traffic was in congestion condition during the 30 minutes observation window; and situations where traffic was in congestion but changed to free flow. Table 1 shows pre-crash and non-crash speed time series clusters, their number of members, and the traffic regimes they belong to. The following table explains the observed traffic regimes in the clustering results.

Table 1 Distribution of crashes inside clusters and regimes of both pre-crash and non-crash situations

	FF			Transitoin-FF2C					MidRang e	Congestion				T-C2FF				
Pre-crash Clusters	C1			C2	C3	C4	C5	C6	C7	C8	C9	C10		C11				
Cluster members	195			165	25	37	88	45	22	123	90	20		14				
	23.6%			20%	3%	4.4%	10.6%	5.5%	2.7%	15%	16%	2.5%		1.7%				
Regimes members	195			227			88	280				34						
percentage	23%			28%			11%	34%				4%						
Non-crash Clusters	C1	C2	C3	C5	C6	C7	C8	C9	C4	C15	C16	C17	C18	C10	C11	C12	C13	C14
Cluster members	888	1415	597	17	19	27	27	52	201	191	393	122	25	42	34	21	28	21
	22%	34%	14.5%	0.4%	0.4%	0.7%	0.7%	1.3%	4.9%	4.6	9.5%	3.0%	0.6	1.0%	0.8%	0.5%	0.7%	0.5%
Regimes members	2900			142					201	731				146				
percentage	70%			4%					5%	17%				%4				

FF: Free Flow

T-FF2C: Transition-Free Flow to Congestion

T-C2FF: Transition-Congestion to Free Flow

- 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 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. 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. Table 1 shows that Free Flow regime contains 238 crashes out of 824 which means 29% of crashes have occurred in free flow state. Among the weekdays, Saturday has 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 (3, 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. The cluster 5 in pre-crash situations and the cluster 4 in non-crash situations contain midrange traffic situations. Figure 4 and 5 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.

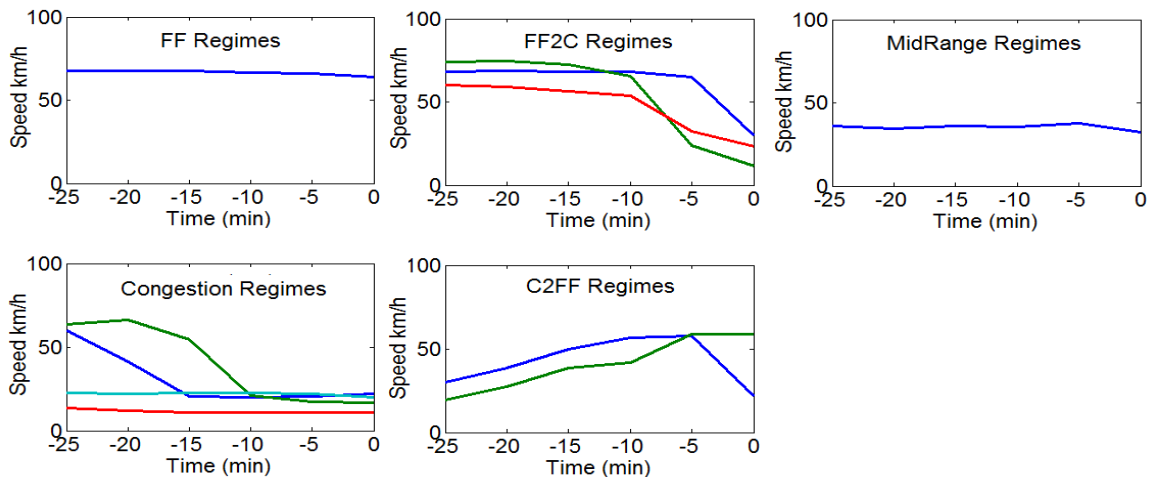


Figure 6 Average of speed time series of pre-crash clusters categorised by traffic regimes

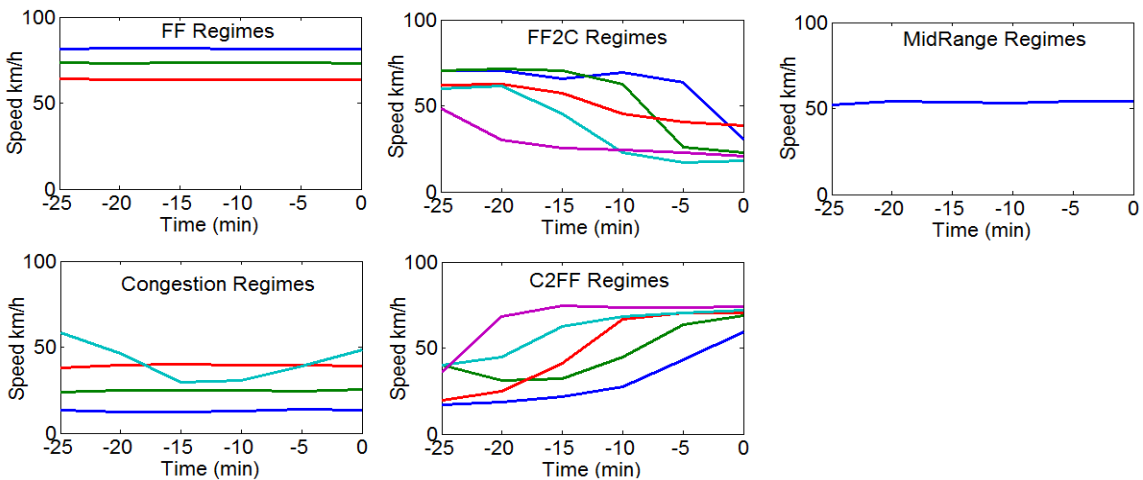


Figure 7 Average of speed time series of non-crash clusters categorised by traffic regimes

- **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 minutes 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 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 congestion state into free flow state. The main factor of crash occurrence is fluctuation of traffic speed during returning of traffic into the free flow state. While traffic is in congestion state. There are two clusters (10 and 11) in pre-crash situations with population of 4% of 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 while distribution of crashes in other weekday is almost in a same level.

5.3 Clusters distances and comparison

To further analyse pre-crash and non-crash clusters, similarity of them are calculated using Euclidean distance function. In this regard, average of speed time series of clusters are chosen as representative of the 29 clusters. Appendix one shows the distance of clusters in a 29x29 matrix. Lower value of distance between two clusters reveals their higher similarity. The distance between pre-crash clusters with themselves and distance between non-crash clusters with themselves show the quality of speed time series clustering. To answer the research questions of this study, the pre-crash clusters are compared with the non-crash clusters in terms of their similarity. A desirable pre-crash cluster is a cluster that is far from the non-crash clusters that show the normal traffic patterns. Therefore, pre-crash clusters are compared with the non-crash clusters. Basically, clusters of a same regime from both pre-crash and non-crash clusters are expected to be more similar to each other rather than clusters in other regimes.

According to Appendix 1 and table one, clusters C2, C3, C4, C6, C10, and C11 have high distance from non-crash clusters. For example, the most similar cluster to C2 (165 member) is N5 (17member). Cluster C2 that contains 20% of crashes has a very small similar cluster (N5 with 0.4% population) among non-crash clusters. Clusters C10 and C11 are unique clusters that are far from all 18 non-crash clusters. However, clusters C1, C8, and C9 have a similar match among non-crash clusters. There similar matches are clusters N3, N15, and N16 respectively. Moreover, among non-crash clusters eight clusters (N1, N10, N11, N12, N13, N14, N17, and N18) discovered as unique non-crash clusters that no match found in the pre-crash clusters for them. The importance of the unique clusters is that they are more likely to be detected in Crash Likelihood Estimation studies.

Moreover, PCA applied on the speed time series to reduce the available six dimensions into 2 dimensions by choosing the first two principal components. The density and dispersion traffic situation are shown in the figures 8 and 9. Figure 8 shows the pre-crash speed time series on a 2D diagram where axis x is the first principal component and axis y shows the second principal component. The colors are showing different traffic regimes. In this visualization of speed time series, two regimes of transition from Congestion to Free Flow and MidRange are obviously not similar to the same regimes in non-crash data.

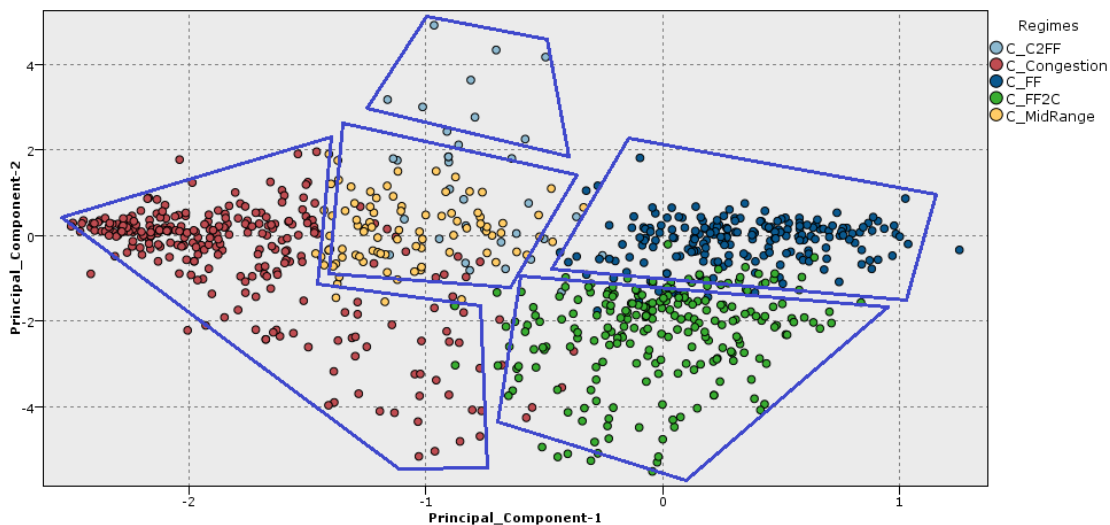


Figure 8 Pre-crash speed time series shown in a 2D diagram from first two principal components of data created by PCA

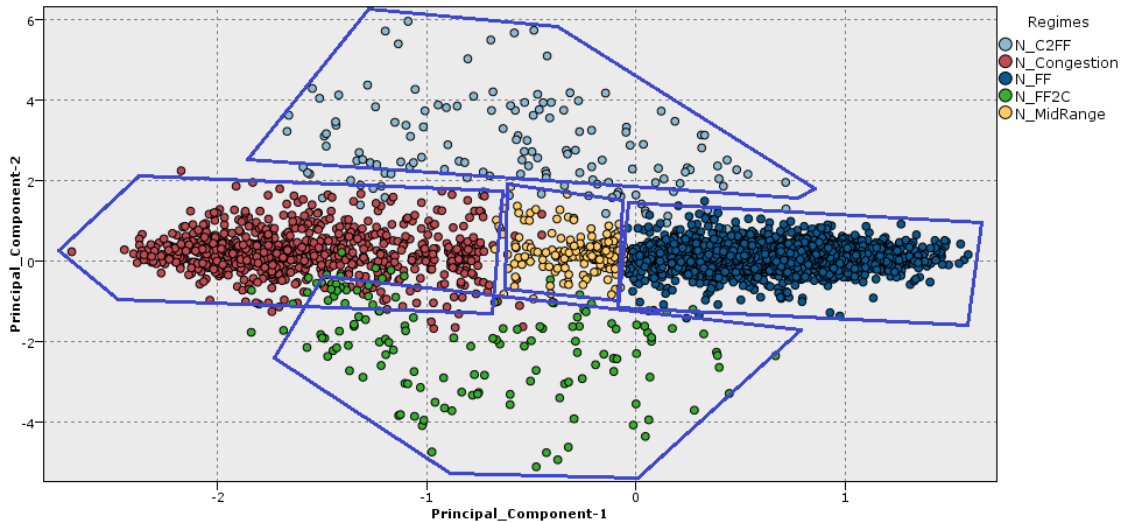


Figure 9 Non-crash speed time series shown in a 2D diagram from first two principal components of data created by PCA

6. Conclusion

This paper studied pre-crash and non-crash traffic situations to find dominant traffic trends in both situations. Also, pre-crash and non-crash situations investigated to discover risky trends that have less similarity to normal traffic trends. These risky patterns are important in Crash Occurrence Likelihood (COL) estimation studies and helps to increase the accuracy of risk detection in real-time. In this regard, speed time series clustered using non-hierarchical clustering algorithm (K-Means) and Dunn index is 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: situations where traffic was in free flow state during half an hour prior to crashes, situations where traffic was in free flow but changed to congestion, situations where traffic speed was around 50 Km/h (MidRange), situations where traffic was in congestion condition during the 30 minutes of observation window, and situations where traffic was in congestion condition but changed to free flow. The results show that clusters C2, C3, C4, C6, C10, and C11 have high distance from non-crash clusters which are more detectable in crash likelihood estimation studies. However, clusters C1, C8, and C9 have similar clusters in normal traffic clusters. Future works of the current study is taking other traffic related characteristics rather than speed into account such as traffic volume and road geometry. Moreover, upstream and downstream traffic flow data can be engaged together in traffic situations. 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.

Appendix 1

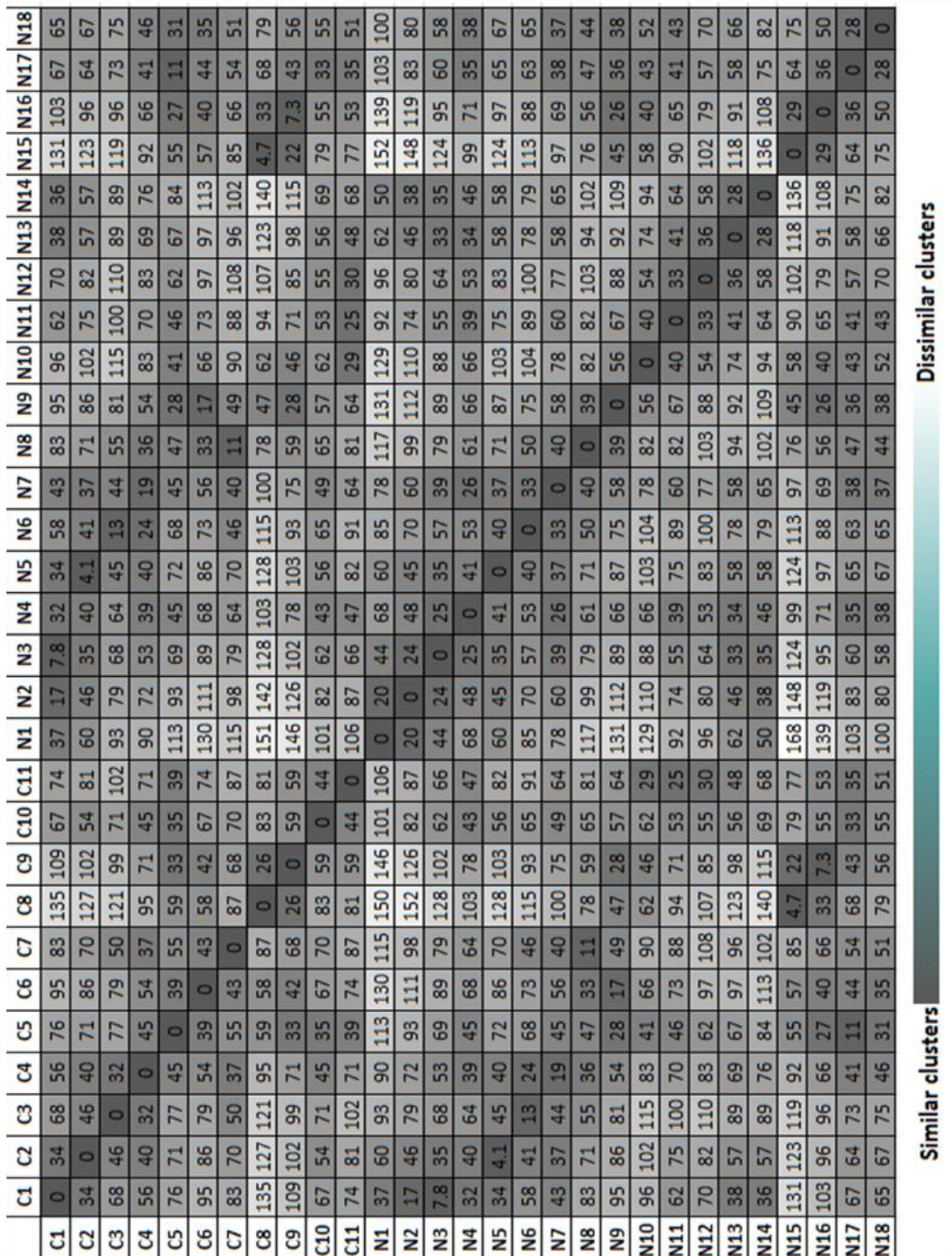


Figure 10 Matrix of distances between all clusters

References

- Abdel-Aty, M., & Pande, A. (2006). ATMS implementation system for identifying traffic conditions leading to potential crashes. *Intelligent Transportation Systems, IEEE Transactions on*, 7(1), 78-91.
- Chawla, N. (2005). Data Mining for Imbalanced Datasets: An Overview. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 853-867): Springer US.
- Golob, T. F., & Recker, W. W. (2004). A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A: Policy and Practice*, 38(1), 53-80. doi: 10.1016/j.tra.2003.08.002
- Golob, T. F., Recker, W. W., & Alvarez, V. M. (2004). Freeway safety as a function of traffic flow. *Accident Analysis & Prevention*, 36(6), 933-946. doi: 10.1016/j.aap.2003.09.006
- Guiyan, J., Shifeng, N., Qi, L., Ande, C., & Hui, J. (2010, 27-29 March 2010). *Automated incident detection algorithms for urban expressway*. Paper presented at the Advanced Computer Control (ICACC), 2010 2nd International Conference on.
- Hamzehei, A., Chung, E., & Miska, M. (2013). *Pre-Crash Traffic Flow Trend Analysis on Motorways*. Paper presented at the OPTIMUM 2013 – International Symposium on Recent Advances in Transport Modelling, Kingscliffe, Australia. <http://optimum2013.com/index.php/download-papers-attenders>
- Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques* (Second Edition ed.): Morgan Kaufmann Publishers.
- Jin, J. P. (2009). *Automatic incident detection based on fundamental diagrams of traffic flow*. The University of Wisconsin - Madison. Retrieved from <http://search.proquest.com/docview/305030731?accountid=13380> (3400010)
- Lawrence A. Klein, M. K. M., David R.P. Gibson. (2006). *Traffic Detector Handbook: Third Edition—Volume I* (Third ed., Vol. 1, pp. 288). McLean, VA: Turner-Fairbank Highway Research Center, Federal Highway Administration.
- Martin, J.-L. (2002). Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention*, 34(5), 619-629. doi: [http://dx.doi.org/10.1016/S0001-4575\(01\)00061-6](http://dx.doi.org/10.1016/S0001-4575(01)00061-6)
- Oh, J. S., Oh, C., Ritchie, S. G., & Chang, M. (2005). Real-time estimation of accident likelihood for safety enhancement. [Article]. *Journal of Transportation Engineering-Asce*, 131(5), 358-363. doi: 10.1061/(asce)0733-947x(2005)131:5(358)
- Pande, A., & Abdel-Aty, M. (2006). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, 38(5), 936-948. doi: 10.1016/j.aap.2006.03.004
- Pham, M. H. (2011). *Motorway Traffic Risks Identification Model - MyTRIM Methodology and Application*. Doctorate, École polytechnique fédérale de Lausanne, Lausanne (4998)
- Pham, M. H., Bhaskar, A., Chung, E., & Dumont, A. G. (2010, 25-27 May 2010). *Towards a pro-active model for identifying motorway traffic risks using individual vehicle data from double loop detectors*. Paper presented at the Road Transport Information and Control Conference and the ITS United Kingdom Members' Conference (RTIC 2010) - Better transport through technology, IET.
- Pham, M. H., El Faouzi, N. E., & Dumont, A. G. (2011). *Real-time identification of risk-prone traffic patterns taking into account weather conditions*. Paper presented at the 90th annual meeting of Transportation Research Board, Washington, DC.
- Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.
- Yeo, H., Jang, K., Skabardonis, A., & Kang, S. (2012). Impact of traffic states on freeway crash involvement rates. *Accident Analysis & Prevention*(0). doi: 10.1016/j.aap.2012.06.023
- Zheng, Z. (2012). Empirical Analysis on Relationship between Traffic Conditions and Crash Occurrences. *Procedia - Social and Behavioral Sciences*, 43(0), 302-312. doi: <http://dx.doi.org/10.1016/j.sbspro.2012.04.103>