Motorway Traffic Risks Identification Model - MyTRIM Methodology and Application

THÈSE N^{O} 4998 (2011)

PRÉSENTÉE LE 16 NOVEMBRE 2011

À LA FACULTÉ ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT

LABORATOIRE DES VOIES DE CIRCULATION

PROGRAMME DOCTORAL EN ENVIRONNEMENT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Minh Hai PHAM

acceptée sur proposition du jury:

Prof. B. Merminod, président du jury
Prof. A.-G. Dumont, Prof. E. Chung, directeurs de thèse
Prof. J. Barceló, rapporteur
Prof. M. Bierlaire, rapporteur
Prof. S. Hoogendoorn, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

Abstract

Road traffic crashes are becoming increasing concerns in many countries. In Europe, many efforts have been devoted to improve road traffic safety yet the important target of halving the number of yearly road deaths in 2010 could not be achieved in many European countries. Among different road types, motorways are safe by design yet crashes if occur would be severe due to high speed practiced. If motorway traffic crash risk could be identified, lives could be saved and severity could be reduced.

For this objective, the current thesis aims to establish a methodology for developing models capable of identifying real-time traffic crash risk on motorways. A real-time MotorwaY Traffic Risk Identification Model (MyTRIM) is developed for a study site on motorway A1 in Switzerland. MyTRIM is tested, validated with real data.

Three types of historical data altogether available at the study site are used for developing MyTRIM. The data include individual vehicle traffic data collected from double loop traffic detectors, meteorological data from meteorological station located near the study site, and a crash database containing crashes recorded by the police. Based on crash time, pre-crash data representing traffic and meteorological conditions leading to crashes are extracted. Similarly, non-crash data representing traffic and meteorological conditions that are unrelated with crashes are also extracted. As crashes are rare events, a methodology for sampling non-crash data comparable with pre-crash data is developed using clustering – classification basis: non-crash data are clustered into groups; pre-crash data are classified into obtained groups; pre-crash and non-crash data within one group are similar and therefore, comparable. Each group is called a *traffic regime*.

Under each traffic regime, a regime-based Risk Identification Model (RIM) is developed to differentiate pre-crash and non-crash data. Given a new datum, regime-based RIM must be able to classify the datum into pre-crash or non-crash. As a result of the model development, variables which are important for the differentiation are also identified. These important variables can be potential for implementing countermeasures to prevent the risk from ending up with a crash. MyTRIM is developed based on the outputs from regime-based RIM. MyTRIM memorizes the latest risk evolution to predict whether there is crash risk in the coming time interval. Regime-based RIM and MyTRIM are tested and validated using real data. Results show that regime-based RIM and MyTRIM perform with high accuracy.

The output of MyTRIM can be useful for traffic operators as an input for actively managing the traffic. The developed methodology can be applied for any motorway traffic sections where similar data are available.

Keywords: motorway traffic safety, accident prevention, real-time risk identification, individual vehicle data, meteorological data, imbalanced data sets, rare events, MyTRIM.

Résumé

Les accidents de la circulation sont une préoccupation de plus en plus croissante dans de nombreux pays. En Europe, malgré que de nombreux efforts aient été consacrés à améliorer la sécurité routière, l'objectif de réduire de moitié le nombre de morts annuel sur les routes en 2010 ne pourrait être atteint dans plusieurs pays. Parmi les différents types de routes, les autoroutes sont les plus sures grâce à leur conception. Cependant, les accidents qui s'y produisent sont sévères due à la grande vitesse pratiquée. Si le risque d'accident du trafic autoroutier pouvait être identifié, des vies pourraient être sauvées et la gravité de ces accidents pourrait être réduite.

Dans cette perspective, cette thèse vise à établir une méthodologie pour développer des modèles capables d'identifier en temps réel le risque d'accident de la circulation sur les autoroutes. Un modèle d'identification des risques de circulation en temps réel sur les autoroutes (MyTRIM) est développé pour un site d'étude sur l'autoroute A1 en Suisse. MyTRIM est testé et validé avec des données réelles.

Trois types de données historiques disponibles sur le site d'étude sont utilisés pour développer MyTRIM. Il s'agit des données de trafic des véhicules individuels recueillies par des capteurs de trafic à double boucle, des données météorologiques de la station météorologique située à proximité du site d'étude, et d'une base de données d'accidents contenant les accidents enregistrés par la police. Basées sur le temps d'apparition des accidents, les données pré-accidentelles représentant les conditions de circulation et météorologiques conduisant à des accidents sont extraites. De même, les données représentant un trafic et des conditions météorologiques sans rapport avec l'apparition d'un accident sont extraites. Comme les accidents sont des événements rares, une méthodologie pour échantillonner les données non-accidentelles par rapport à des données pré-accidentelles est développée, basée sur un processus regroupement - classification: les données non-accidentelles sont regroupées; les données pré-accidentelles sont ensuite classées dans ces groupes définis; les données pré-accidentelles et les données non-accidentelles au sein d'un groupe sont similaires et, par conséquent, comparables. Chaque groupe est appelé régime de circulation.

Dans chaque régime de circulation, un modèle d'identification des risques (RIM) est développé pour différencier les données pré-accidentelles et non-accidentelles. Pour chaque nouvelle donnée, le RIM doit être capable de classer celle-ci en donnée pré-accidentelle ou non-accidentelle. Grâce à l'élaboration du modèle, les variables importantes pour la différenciation peuvent également être identifiées. Ces variables présentent un potentiel pour la mise en œuvre de contre-mesures afin de prévenir qu'un risque se transforme en accident. MyTRIM est développé sur la base des données d'output des RIM. MyTRIM mémorise l'évolution du risque pour prédire si un risque d'accident dans un intervalle de temps futur existe. Les RIM basés sur les régimes de circulation et le MyTRIM sont testés et validés à l'aide de données réelles. Les résultats montrent que les RIM et le MyTRIM fonctionnent avec une grande précision.

Les résultats de MyTRIM peuvent être utilisés par les gestionnaires de la circulation comme une donnée d'entrée pour la gestion active du trafic. La méthodologie développée peut être appliquée pour toutes les sections d'autoroutes présentant des données similaires.

Mots-clés: sécurité du trafic autoroutier, prévention des accidents, identification des risques en temps réel, prédiction du risque d'accident, données de véhicule individuel, données météorologiques, déséquilibre des classes, événements rares, MyTRIM.

Acknowledgements

I take this opportunity to thank those who made this dissertation possible.

First and foremost, I would like to express my sincere thanks to my supervisors Professor André-Gilles Dumont and Professor Edward Chung. Professor Dumont has accepted me as a member of LAVOC and presented his constant confidence in me. Professor Dumont has provided his full support for me to conduct research and to complete my PhD. Professor Chung has offered me the chance to become his PhD student and worked with me deeply in the topic. Discussions with Professor Chung not only brought new understandings but also opened new challenges that I needed to face with. His precious advices and guidance, his patience and his availability to me are invaluable and gave me lots of encourages.

I would like to thank my colleague and friend, Dr. Ashish Bhaskar, who shared with me all his ideas and experiences of doing PhD. He frequently gave me friendly and priceless advices and comments.

I would like to express my thanks to jury members, Prof. Bertrand Merminod, Prof. Jaime Barceló, Prof. Michel Bierlaire and Prof. Serge Hoogendoorn, for their constructive comments and suggestions which are helpful to improve the quality of the report.

I wish to express my gratefulness to people who discussed with me and gave me advices regarding my thesis. They are Dr. Nour-Eddin Al Faouzi, Dr. Olivier de Mouzon, Prof. Masao Kuwahara, Prof. Michel Bierlaire, Prof. Nikolaos Geroliminis, Dr. Marc Miska, Dr. Jean-Marie Fuerbringer, Prof. Takashi Oguchi and Dr. Hiroshi Warita.

I am thankful to the members and ex-members of LAVOC for their support and help. To mention a few, Jean-Jacques Hefti, Emmanuel Bert, Patrick Rychen, Nicolas Bueche, Margarita Rodriget, Chiara Paderno, Jean-Wilfrid Fils-Aimé, Michel Pitet, Jean-Claude Reymond, Charles Gilliard and Dominique Corday.

I also would like to thank the community of Vietnamese students in Lausanne, VnLausanne, for their support and encouragements.

Finally, I would like to express my special gratitude to my family: my wife Minh Ha, my daughter Minh An, my parents, my sisters and brother for their constant and unconditional support, affection, understanding, and encouragement.

(Minh Hai, PHAM)

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

This chapter presents the motivation, problem statement research objectives, research questions, research scope of the current research. Thereafter, the approach to fulfill the research objectives is presented. Expected contributions, both theoretical and practical, are discussed.

1.1. Research Motivation

In recent years much attention has been devoted to road traffic safety in most countries. According to the World Health Organization - WHO (2004), road traffic crashes kill 1.2 million people every year, an average of 3'200 per day. Road traffic crashes also injure between 20 and 50 million people a year and have been ranked the 11th major cause of death, accounting for 2.1% of all deaths globally.

The European Transport Safety Council - ETSC (2006) estimates that more than 41'000 people die every year from road crashes in the European Union (EU). In 2001, the European Commission fixed the road safety target of halving the number of yearly road deaths by 2010. However, most of member countries couldn't achieve this goal.

Unlike other road types, motorways are the safest roads by design. Yet in 2006 at least 3'270 people were killed on the motorway network in the EU, representing about 8% of the total number of road deaths (ETSC, 2008). As motorway crashes can be rather severe, improving motorway safety would not only reduce the number of deaths but would also moderate crash severity.

Unlike other road types, motorways are the safest roads by design. Yet in 2006 at least 3'270 people were killed on the motorway network in the EU, representing about 8% of the total number of road deaths (ETSC, 2008). As motorway crashes once occur will be severe, improving motorway safety would not only reduce the number of deaths but also moderate crash severity.

To that extent, the research reported in this thesis aims at devising a methodology for the development of real-time motorway traffic risk identification models. If risky traffic situations could be recognized, preventive measures could probably be taken to clear or reduce the risk before crashes occur.

1.2. Problem Statement

Crash evolution can be divided into three phases: before, during and after the crash. Once a crash occurs, it is necessary to provide urgent health care services to the injured parties, to prevent secondary collisions, as well as to guarantee traffic fluidity as much as possible. During crash occurrence, the involved individuals, especially the drivers, need to react accordingly to the situation to reduce its severity. In this phrase, passive equipment can help to improve the security of the involve individuals. Incident detection is useful when the crash has already occurred so that urgent services and manual traffic control can be brought to the site. However, all the actions undertaken during or after the crash are reactive, aiming to reduce the crash severity with its consequences. In order to prevent collisions, traffic evolution before the crash should be evaluated and identified.

In practice, as illustrated for a crash occurring at around 21:00 in Figure 1-1, with constant traffic surveillance, incident detection tries to detect collisions as soon as they occur (i.e. accidents can only be detected in the aftermath).If there is a way to identify real-time crash risks, high traffic risk before the crash should provide an indication that this risk needs to be reduced to as a preventive measure, to avoid further accident occurrence. Although incident detection systems have been successfully developed and implemented, identifying real-time traffic crash risk remains an interesting yet incompletely characterized topic. Motivated by this fact, this thesis addresses the real-time crash risk identification problem.

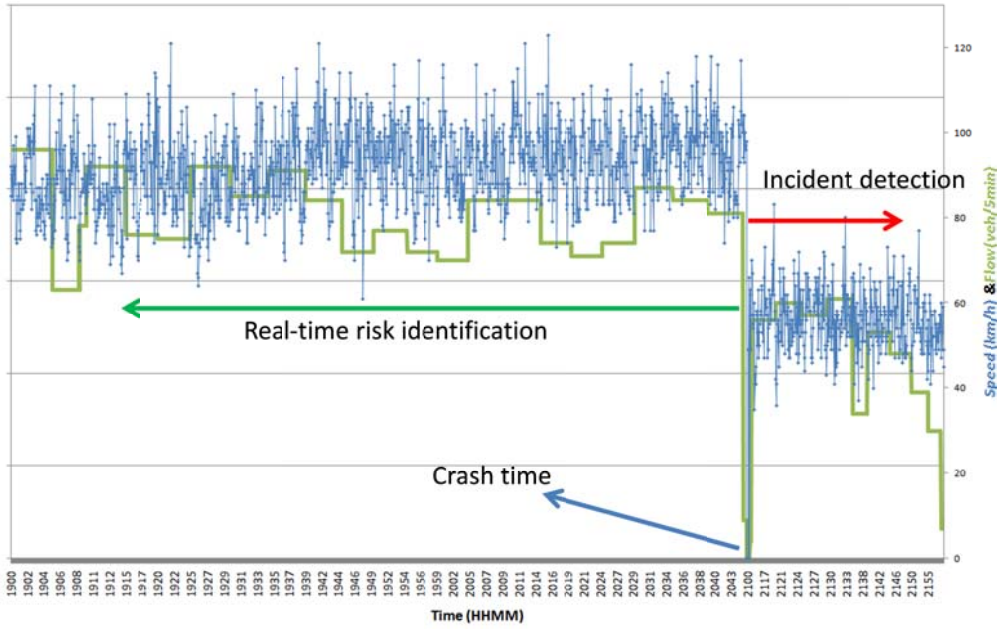

Figure 1-1: Distinction between real-time risk identification vs. incident detection

1.3. Research Objectives

The current research aims to address the real-time crash risk identification problem with the main objective of developing a methodology for identifying such risk on the motorways. To this end, the following instrumental tasks will be undertaken:

1) Understand the state of the art of similar studies (i.e. studies on motorway traffic safety using traffic flow data to develop real-time tools for monitoring motorway traffic crash risk).
2) Understand the needs and requirements for developing real-time tools for identifying motorway traffic crash risks under the Swiss motorway network conditions.
3) Establish a methodology for real-time traffic risk identification. The methodology should be flexible in order to be applied on road conditions that are similar to the Swiss motorway network ones.
4) Apply the methodology to develop a model capable of identifying traffic crash risk for a study site in Switzerland.
5) Provide an application framework where real-time crash occurrence can be identified in advance.

1.4. Research Questions

Traffic crashes are rare events on motorways. Usually, there is an evident causality in the dynamic of accidents. The effect of the causality is the crash itself whereas the cause is usually unknown. As the cause always occurs prior to the effect, to prevent the effect, it is necessary to identify the cause. Therefore, the cause must be found in the traffic evolution prior to the collision. One approach is to differentiate between a series of traffic situations before the collisions with series of traffic situations occurring under similar conditions that do not end up with a crash.

Therefore, the current thesis attempts to find answers to following questions:

1) What is the information required for representing traffic situations?
2) As crashes are rare events, which traffic situations that do not end up with collisions are comparable to the ones ending up in a crash?
3) Which method can be chosen to develop models for differentiating traffic situations that do not end up with collisions from the ones that do result in collisions?
4) Having identified crash risk evolution, how to develop a tool that can predict traffic crash risk with high accuracy?

1.5. Research Scope

To address the questions presented above and based on the available conditions in Switzerland, the following areas have to be investigated.

1.5.1. Types of Crashes

The types of crashes mentioned in the current research encompass events in which there is a collision between two or more vehicles or between one vehicle and stationary objects.

Motorways are by default designed to prevent certain types of collisions. For example, head-on crashes can never occur on motorways as there is a separator between the two traffic directions. Besides, there is a low occurrence probability for crashes that are usually observed at intersections of normal roads such as T-style ones (head of a vehicle with side of another vehicle) as there are no crossing at level with any other road.

There are collisions that can indeed occur on motorways yet it is almost impossible to automatically prevent for they are caused by reckless driving under the influence of alcohol and drugs or by technical problems. We will not consider these crashes in this current research.

Thereby, the category of collisions of interest in the current research, are the ones related to the traffic itself, termed *traffic-induced crashes*. The drivers who are involved in these particular collisions are therefore normal people under normal behavior at the moment of the crashes. It is important to note that a healthy driver under normal conditions should not be influenced neither by alcohol and drugs, nor by fatigue or mental/physical illness.

1.5.2. Study site

Once the methodology for identifying traffic crash risk is established, study sites are necessary for testing the methodology. Ultimately, study sites should be on motorways. According to the Organization for European Economic Cooperation (OECD, 2004), a motorway is defined as a *road specially designed and built for motor traffic, which does not serve properties bordering on it, and which:*

a) is provided, except at special points or temporarily, with separate carriageways for the two directions of traffic, separated from each other, either by a dividing strip not intended for traffic, or exceptionally by other means;
b) does not cross at level with any road, railway or tramway track, or footpath;
c) is specially sign-posted as a motorway and is reserved for specific categories of road motor vehicles.

Entry and exit lanes of motorways are included irrespectively of the location of the sign-posts. Urban motorways are also included.

Road sections to be selected should be "2x2" motorway road sections (i.e. two lanes per direction). This selection is based on the fact that "2x2" motorway road sections contribute up to 73.68% of the length of the Swiss motorway network (see (FEDRO, 2009)). The study site of the current research is presented in details in section 4.4.

1.5.3. Applicability of Results

One target of the current research is to provide traffic operators with a tool to monitor traffic in real-time. Any methodology or model to be developed must be applicable with real-time data.

1.6. Approach

The approach of the current research is presented and summarized in Figure 1-2.

With the motivations explained in section 1.1, and the problem statement and objectives presented in sections 1.2 and 1.3, respectively, the next step would be to focus on the state of the art by reviewing similar studies in the literature with criticism, which would provide an overview on how to move on from here: what should be accomplished and what should not be accomplished in the current analysis. Based on the critical review, the methodology for identification of real-time traffic crash risks is devised to fill in certain gaps in the state of the art. Once the methodology is established, it is applied in a real-life case study under Swiss conditions. Data and study sites are introduced before a model capable of identifying real-time traffic crash risks is developed, evaluated and validated with comparison to models developed with existing methodologies in the literature. The outcome of the model development process should include not only the model itself but also the causes of crashes. From these causes, appropriate preventive measures can be proposed to reduce collision risks.

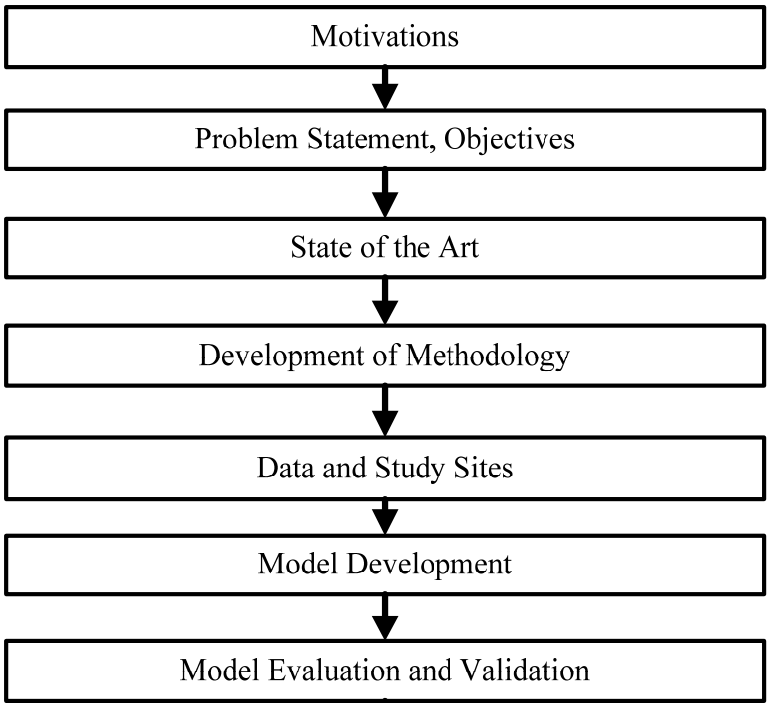

Figure 1-2: Research Workflow

1.7. Contributions

The planned contributions of the current research are both theoretical and practical.

1.7.1. Theoretical Relevance

1) A methodology for sampling traffic situations that do not result in collisions and that are comparable with the ones resulting in collisions.
2) A methodology for identifying real-time traffic crash risk on motorways.

1.7.2. Practical Relevance

1) A framework for predicting near future traffic crash risk in real-time.
2) A model developed using the methodology for study sites in Switzerland.

1.8. Organization of the Dissertation

The summary of the different chapters is presented in Table 1-1. Chapter 1 introduces the motivation, the problems, the objectives, the scope, the approach as well as the potential contribution of the current research. Chapter 2 focuses on reviewing relevant existing studies. Chapter 3 begins with the presentation of the methodology development, as well as with the theoretical fundamental of the methodology. With

the developed methodology, chapter 4 presents data and study sites used for a case study in the current research. Chapter 5 presents the concepts of traffic situations and their definition with respect to the selected study site. Chapter 6 focuses on data sampling issue and provides analyses on traffic regimes obtained. Chapter 7 discusses the application of the developed model in a real-time framework with results and analyses. Ultimately, the conclusions as well as potential application and future extensions are discussed in chapter 8.

Table 1-1: Organization of the current dissertation

Chapters	Content
Chapter 1	Introduction to the research problems and objectives.
Chapter 2	Literature review of relevant studies.
Chapter 3	Methodology of the current research.
Chapter 4	Introduction to data and study site as well as data related issues.
Chapter 5	Definition of Traffic Situations
Chapter 6	Data Sampling and Traffic Regimes
Chapter 7	Risk Identification
Chapter 8	Conclusions

Chapter 2 State of the Art

This chapter summarizes the studies previously undertaken that are relevant to the problems addressed in the current research. Here, we will concentrate on road traffic safety studies in general, with a particular attention on motorway traffic safety studies that utilize traffic flow data. Critical reviews to such studies are also discussed, in order to identify the gaps in the state of the art, which provide a starting point for the definition of the development of the methodology of the current research.

2.1. Overview

Road traffic safety is a branch of traffic engineering aiming to reduce the number of deaths and injuries as well as property damages that are the consequence of collisions of vehicles traveling on public roads. The main road traffic danger is represented collisions between a vehicle with other vehicles, pedestrians, and moving obstacles, such as animals or stationary obstructions (e.g. trees or utility poles). Therefore, the main objective of road traffic safety is to reduce the number and the severity of such crashes.

In term of terminology, it is worth noting that there are organizations such as the NHTSA (1997), RoadPeace (2011), "a CRASH is NO accident" (2011), cautioning not to use the term *accident* and using other names, such as *crash or collision* instead. The idea behind this renaming is to draw the attention to the fact that crashes *are not acts of fate but are predictable and preventable* (NHTSA, 1997), whereas accidents are events which are out of human control . Thereby, the term *accident* is not used in the current research specifically because we aim to reduce the number and the severity of road traffic collisions.

Different efforts were dedicated to uncover the factors that cause crashes. Rumar (1985) using British and American crash reports, found that 57% of car crashes were due solely to driver factors, 3% solely to roadway factors, 2% solely to vehicle factors, 27% to combined roadway and driver factors, 6% to combined vehicle and driver factors, 1% to combined roadway and vehicle factors, and 3% to combined roadway, driver, and vehicle factors. These data suggests that, except collisions due solely to vehicle factors, roadway factors or to both, driver factors are wholly or partly involved in 94% out of the total crashes analyzed.

Crash risk interventions can be categorized in many forms, depending on the driver, the vehicle, and roadway factors. For example, drivers can be divided into different groups depending on age, type of vehicles, etc. such that the specific or typical problems of each group can be better addressed. In Switzerland for instance, young drivers from 18 to 24 years of age are recognized by the BPU (2011) as the group of drivers highly influenced by alcohol and having the highest rate of fatalities and severe injuries (10.5 every 100'000 inhabitants during weekdays and 17.9 every 100'000 inhabitants during weekends). Therefore, legal measures adopted against young drivers are usually strict. One of such measures is related to application of the "trial driving license" for a period of three years after the success of driving examination, called probationary period. The permanent driving license can only be obtained after three years if no traffic code violation by the driver is observed (SAN, 2011).

The interventions related to roadway or vehicle factors mainly aims to improve the design and maintenance. In Europe, there has been a Campaign for Safe Road Design since 2009, calling to *make safe road design a Europe transport priority* (EuroRAP, 2009). This European campaign is an expansion of the UK Campaign for Safe Road Design and targets to cut Europe's toll of road deaths and serious injuries by a third by improving road design (EuroRAP, 2009). For better addressing roadway factors,

roads are classified into three main categories: built-up areas, non-built-up areas, and motorways. In many countries, such as in the UK (DfT, 2009) or in Switzerland (BPU, 2011), the highest number of casualties is observed in built-up areas, which is much more than on motorways. Therefore, a higher number of preventive measures are applied in built-up areas, especially on shared space locations, where vulnerable road users, such as pedestrians and bicyclists, can be found. Vehicle safety is also an effective approach to improve traffic safety by reducing the chances of driver's errors (called *active safety*) or by improving safety equipment and vehicle design in order to reduce crash severity once crash occurs (called *passive safety*). Examples of active safety are the intelligent systems added in vehicles to assist drivers such as: automatic breaking systems, adaptive cruise control, pre-crash system, etc. Examples of passive safety equipment are: seat belts and air bags.

In this chapter, driver factors will be more summarized in section 2.2. The particularity of motorway traffic safety is discussed in section 2.3. Studies close to the current research are presented in section 2.4. Ultimately, section 2.5 summarizes the literature and the gaps to be filled in of the state of the art.

2.2. Driver Factors

Although various efforts undertaken prove that traffic safety is improved, the concept of problem solving does not exist yet. Even when safe roads and safe vehicles are designed and implemented, the number of crashes can only be reduced, but does not disappear. The reason being is that driver factors are vital, and it is exactly the driver who decides and performs all driving tasks. As early as in 1939, Farmer and Chambers introduced the term *accident proneness* to indicate *a personal idiosyncrasy predisposing the individual who possesses it in a marked degree to a relatively high accident rate*. This means that a number of drivers have to be responsible for causing accidents because they have certain harmful personal characteristics. Later on, Rumar (1985) found that drivers are responsible for 94% of crashes. Yet even considering such a result, it is always possible to claim that drivers could have better reacted or could have been more prepared to avoid collisions due to road or vehicle problems. In order to improve safety, driver factor approaches emphasize safety education and motivation of persons aiming to modify unsafe driving behaviors.

However, driver blaming is not supported by many research groups. People following systems theory consider that collisions arise from interaction among humans, machines, and environment. Collisions are seen as a failure of the whole traffic system rather than a failure of the driver. Under normal conditions, it is assumed that there is the harmony between humans and environment, such that the chance of collision is naturally low. For this reason, drivers are victims of crashes, since the traffic system is too complex for the driver's capacity to process information.

According to Parker et al. (1995), there is *a three-fold typology of aberrant driving behaviors*. The three types are: lapses, errors, and violations. Lapses are behaviors related to driving skills when the driver is absent-minded, with consequences mainly for the perpetrator, posing no threat to other road users. Lapses are related to technical mastering, such as mistaking the breaking with the gas pedal, turning on the wrong turning signal between left and right turns, etc. Errors are more dangerous than lapses, as errors are the failing result of planned actions aimed at reaching a goal. Examples of errors include: changing lanes for passing without seeing the coming vehicle on the opposite one, failing to recognize stop signs at intersection, etc. It is believed that lapses and errors are influenced by cognitive factors such as attention and habits. Both lapses and errors depend on driving skills and are observed more frequently in older drivers, or in young drivers who just obtained their license. The third type of driving behavior, which is the most dangerous one, is related to violations. Violating actions are committed intentionally, in the

knowledge that one is engaging in a potentially dangerous and often illegal behavior. Examples of violations include: speeding, close following, overtaking on the inside, texting while driving, etc.

To participate in road traffic, drivers are required to have sufficient driving skills including both physical and mental capabilities. Therefore, lapses and errors are influenced by perceptual, intentional, or judgmental processes. To avoid lapses and errors, the drivers can improve their limitations in observation as well as technical mastery. Vehicles and roads can also be improved to facilitate driver's behaviors. On the contrary, violations are believed to be based on motivational and/or social factors. General solutions to diminish violations deal with changing people's beliefs and/or motives for avoiding violations such as attitude campaigns, police surveillance, speed cameras, or influencing the driver subconsciously through *smart-design* of the road environment.

2.3. Motorway Traffic Safety

2.3.1. Introduction

Researches in motorway traffic safety can be broadly divided into groups of studies based on the objects of the improvement, such as infrastructure, vehicles, drivers, etc. Improving traffic safety by improving traffic flow is also part of these studies, and aims at predicting the future trend of traffic risks by analyzing past crashes and suggesting countermeasures when risks are identified, in order to prevent potential future ones. As the traffic flow is the main object of our study, we classify our study into the group of motorway traffic safety studies using traffic flow data.

According to Golob et al., (2004), consecutively by Pande, (2005), studies on motorway traffic safety using traffic flow data can be divided into two groups, depending on the units of analyses undertaken: aggregate studies and disaggregate studies. Aggregation studies consider counts of crashes or crash rates for specific time periods (typically months or years), as well as for specific spaces (specific roads or networks), as units of analysis and use statistical distribution parameters of traffic flow for that specific time and space. For disaggregate studies, the units of analysis are the crashes themselves and traffic flow is represented by parameters of at the time and place of each crash. As our study examines each crash to verify the difference between traffic conditions leading to the crash in crash case and non-crash traffic conditions, it can be classified into the group of disaggregate studies. This classification is illustrated in Figure 2-1, where traffic safety studies using traffic flow data (Traffic Flow related studies) are partitioned into the two smaller groups: aggregation studies and disaggregation studies.

Infrastructure related studies focus on improving traffic safety by improving the infrastructure. Stine et al. (2010), for example, investigated the safety of highway medians through iterative simulations of off-road median encroachments. Another example is given by the study of Donnell and Mason (2006), who investigate median design policies for high-speed divided highways to assess existing median barrier warrant criteria.

Vehicle and driver related studies investigate traffic safety issues that are applicable both for motorways and non-motorways. For vehicle related studies, Blum and Eskandarian (2006), reviewed the research that has been conducted on intelligent speed adaptation, and presented possible strategies to maximize both effectiveness and acceptability to mitigate deleterious effects on roadway safety. Umedu et al. (2010) focused on the inter-vehicular communication, namely on the distributed detection of dangerous vehicles on roads and highways, and proposed a dangerous-vehicle-detection protocol to detect drivers who violate

speed limits. Klingender et al. (2009) concentrated on the different effects of adapting the maximum weight and dimensions of heavy commercial vehicles allowed by the European Commission by providing an overview of the in-depth safety analysis of heavy commercial vehicles on European roads. Regarding driver related studies, a large effort has been devoted to discover the influence of driver state on driving (Arnedt et al., 2001; Rzepecki-Smith et al., 2010); (Mulder et al., 2008).

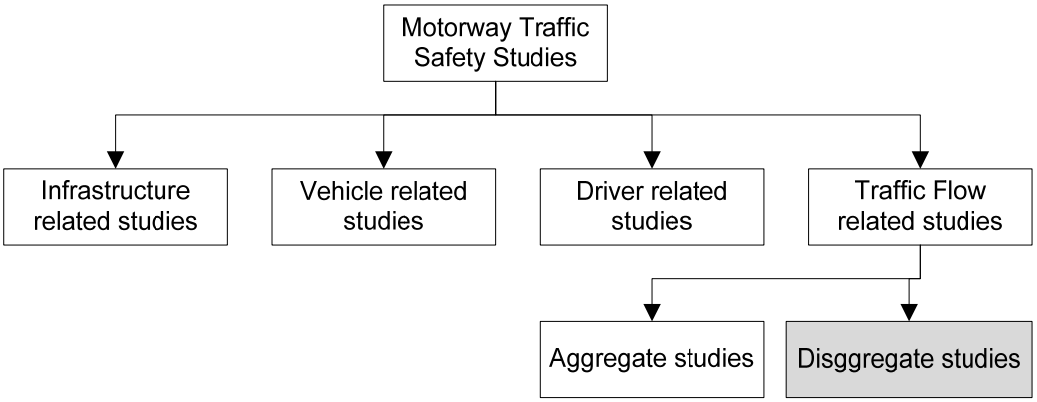

Figure 2-1: Research area of the current study

It is worth noting that the classification presented in Figure 2-1 is rather relative, as traffic is a complex system. To improve its safety the interaction between different groups is necessary and unavoidable.

2.3.1. Aggregate Studies

During the past decades a large effort has been devoted to aggregate studies on motorways traffic safety. To date, aggregate studies are still widely used to find and to analyze the relationship between crash rate/frequency and other factors such as the congestions, infrastructure, weather effect, etc.

Wang et al. (2009) investigated the relationship between traffic congestion and road accidents to clarify the speculation that there may be an inverse relationship between the latter and road accidents. The study aimed to explore the impact of traffic congestion on the frequency of road accidents using a spatial analysis approach, while controlling for other relevant factors that may affect road accidents. The results suggest that traffic congestion has little or no impact on their frequency on the M25 motorway in the UK.

Aguero-Valverde and Jovanis (2006; 2008) focused on spatial analyses to produce spatial models. The authors tried to explore the effect of spatial correlation in models of road crash frequency at the segment level. Different segment neighboring structures were tested to establish the most appropriate one in the context of modeling crash frequency in road networks. A full Bayesian hierarchical approach was used with conditional autoregressive effects for the spatial correlation terms. Analyses of crash, traffic, and roadway inventory data from a rural county in Pennsylvania indicate the importance of including spatial correlation in road crash models. The models with spatial correlation show a significantly better fit than the Poisson lognormal model with heterogeneity only. Parameters significantly different from zero included annual average daily traffic (AADT) and shoulder widths of less than 4 ft, and between 6 and 10. One important result relates to the potential of spatial correlation to reduce the bias associated with model misspecifications.

Recently, Lord and Mannering (2010) provided a detailed review of the key issues associated with crash-frequency data, as well as reporting the strengths and weaknesses of the various methodological approaches that researchers have used to address these problems. The authors concluded that while the steady march of methodological innovation (including recent applications of random parameter and finite mixture models) has substantially improved our understanding of the factors that affect crash-frequencies, it is the prospect of combining evolving methodologies with far more detailed vehicle crash data that holds the greatest promise for the future.

In general, aggregate studies, although they bring more insights of crash frequency, do not propose any real-time models for identifying instant traffic crash risks.

2.3.2. Disaggregate Studies

Disaggregate studies are relatively new, and are made possible by the availability of data being collected in support of intelligent transportation systems developments. Motorway traffic management centers routinely archive traffic flow data from sensor devices such as inductive loop detectors. Different disaggregate data types can be obtained from such sensor devices depending on the types of sensors and on the aggregation level of collected traffic data. For example, single inductive loop detectors collect traffic data for intervals of several seconds, and provide the vehicle count and the occupancy during the intervals. Double inductive loop detectors can provide more detailed data than vehicle counts and occupancies. The average speed during the time interval is the extra information. Moreover, double loop detectors can also provide individual vehicle data including: speed, headway, and time gap, length of each vehicle. It is worth noting that there is no time interval in collecting individual vehicle data (i.e. data are extracted whenever there is a vehicle passing by the detectors)

There are six research groups, that we are aware of, who have undertaken disaggregate studies and have reported positive results in the literature. Most of the studies aim to develop real-time models for identifying traffic risk using disaggregate traffic flow data. The original idea of these studies is to compare non-crash traffic conditions (i.e. traffic conditions when no crash is recorded) and traffic conditions where traffic crash risk is high. In reality, it is difficult to determine which traffic conditions are highly risky. Therefore, crash databases are used to extract traffic conditions leading to crashes (*pre-crash traffic conditions)*. Pre-crash traffic conditions are highly risky because they end up in crashes. By developing models capable of differentiating pre-crash and non-crash conditions, it is expected that a new real-time traffic condition can be collected and classified into non-crash or pre-crash conditions. Therefore, preventive actions can be implemented, if the traffic condition is in its "pre-crash" state, to avoid potential crashes.

In this section, we summarize the work of each research group and highlight their contribution.

2.3.2.1. Studies by Oh et al.

2.3.2.1.1. Objectives

Study reported by Oh et al (2001) is regarded as one of the first disaggregate studies. The authors aim to quantify the measure of accident likelihood using real-time traffic data from inductive loop detectors based on the concept that *disruptive* traffic conditions contribute to traffic accidents, and can be represented by temporal and spatial variations in traffic parameters. The study demonstrates the potential

capacity of identifying traffic conditions that lead to accidents from real-time traffic data. The authors defined an indicator described by the traffic dynamics index that classifies traffic conditions into two patterns: *normal* traffic condition and *disruptive* traffic condition as illustrated in Figure 2-2. The authors expected that the indicators would be stable under normal conditions and would become unstable under disruptive conditions.

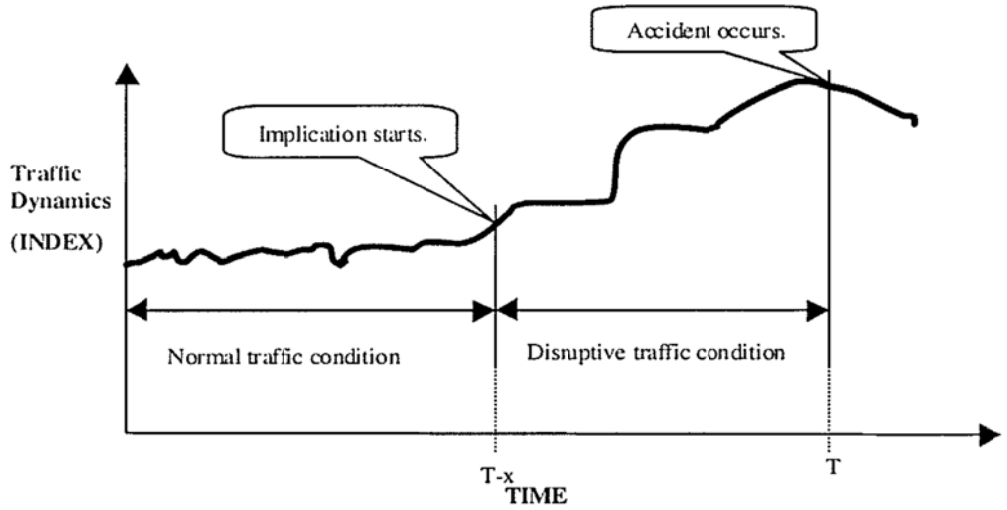

Figure 2-2: Implication of accident by traffic dynamics. Source: (Oh, Oh, Ritchie and Chang, 2001)

2.3.2.1.2. Data Used and Study Site

To find the indicator, traffic data for a 9.2-mile long stretch of the I880 freeway in Hayward, California (February 16 to March 19, 1993) was used. The data was collected from the 17 and from the18 loop detector sections for northbound and southbound lanes respectively, and contain 10-second aggregated information including traffic flow, occupancy, and speed during two periods from 5AM to 10AM, and from 2PM to 7PM. During the collection period, up to 4 probe vehicles traveled on the study section and the incident database was constructed based on reports submitted by probe vehicle drivers.

2.3.2.1.3. Methodology

The first step is to define *normal* and *disruptive* traffic conditions. For each accident, two 5-minute periods before the accident are used. The 5-minute period at 30 minutes before the accident is considered as a *normal* condition and the 5-minute period right before the accident is treated as a *disruptive* condition. The 5-minute traffic data representing normal and disruptive traffic conditions were generated by re-aggregating raw 10-seconds data. The generated normal and disruptive data contain six parameters, namely the mean and the standard deviation of occupancy, flow and speed that were considered as indicator candidates.

Using the normal and the disruptive data sets collected from 52 accidents, the authors used *t*-statistics and found that the 5-minute standard deviation of speed was the most significant in differentiating the normal and disruptive traffic conditions and was therefore selected as the indicator. The authors used the indicator (i.e. the 5-minute standard deviation) in developing a Bayesian model to estimate the probability

for a traffic condition (with its corresponding indicator) to become either normal (with the probability of 0.0) or disruptive (probability of 1.0). A probability threshold was necessary for classifying the latter.

2.3.2.1.4. Comments

The authors were successful in demonstrating the potential of the developed model to be applied in real-time with the definition of a probability threshold that classifies real-time traffic conditions into normal or disruptive. The authors also suggest that reducing accident likelihood is equivalent to reducing the speed variation of vehicles which can be undertaken through an information system, in order to suggest drivers to either slow down or speed up as part of the road infrastructure or via an in-vehicle system.

However, the following potential improvements can be undertaken:

More variables can be used (instead of only the standard deviation of speed). Traffic crash occurrences are the result of complex traffic evolution and depend on many factors.
The five-minute interval 30 minutes before crash might not be representative of normal traffic conditions.
Only the aspect of risk identification was scrutinized in the study. In term of crash prevention, identifying crash risks right before they occur doesn't really help to prevent them.

2.3.2.2. Studies by Lee et al.

2.3.2.2.1. Objectives

The first study of this group in this domain is presented in their 2002 article (Lee et al., 2002). Changes occurred in the later development of the study and the updates are presented in (Lee et al., 2003). In these studies, the authors define the concept of *crash precursors* to refer to *various traffic flow characteristics which lead to crash occurrence*. The objectives of these studies include a suggestion of the rational methods by which crash precursors included in the model can be determined on the basis of experimental results and on the performance test of the crash prediction model.

2.3.2.2.2. Data Used and Study Site

The data used to calibrate the model include incident logs and traffic flow data extracted from loop detectors along a 10 km stretch of the Gardiner Expressway in Toronto, Canada, with a total of 38 loop detector stations. The data were collected for weekdays over a 13 month period, from January 1998 to January 1999, and matched with a total of 234 crashes on this section of the roadway during the study period.

2.3.2.2.3. Methodology

In one of their articles (Lee, Saccomanno and Hellinga, 2002) the authors found three crash precursors representing the traffic flow conditions prior to the crash occurrence. These precursors are represented by the average variation of speed on each lane (CVS1), by the average variation of speed difference across adjacent lanes (CVS2), and by traffic density (D). However, later study (Lee, Hellinga and Saccomanno, 2003) found that CVS2 did not have a direct impact on crash potential and hence was eliminated from

their consideration. In addition, the authors introduced in their latter study the parameter Q, which is the difference of speeds in the upstream and in the downstream ends of road sections.

For developing models, precursor data prior to 234 crashes were used as crash cases. The same number of non-crash cases contains the data collected at the same road sections for the same time periods under the same weather condition but in different days, when crashes did not occur. In this way, other traffic environmental factors such as road geometry, weather and typical traffic pattern are assumed to be controlled.

For each crash, the authors estimated the actual crash time from the analysis of changes in detector speed profiles using loop detectors upstream and downstream the crash site. The authors also undertook various analyses to choose the optimal observation time slice durations by comparing the difference between crash and non-crash cases. The optimal durations should maximize the differences in crash precursor values between the two. The authors found that 8, 3, and 2 minutes were the optimal values for the three precursors CVS1, D and Q, respectively.

In order to discriminate the crash and non-crash cases, the authors categorized precursor values by defining the several levels of precursors based on the distribution of normal traffic flow conditions in daily traffic with the number of categories and boundary values determined by calibrating different log-linear models, in order to choose the one that shows the best model performance.

2.3.2.2.4. Comments

In comparison with the study presented in section 2.3.2.1 by Ol et al., the authors used more variables (three precursors, namely, CVS1, D, and Q) in differentiating crash and non-crash cases. Here, the use of precursor Q provides an additional view of crash occurrences. Although using the same ratio of non-crash/crash cases (1:1), the authors selected randomly controlled non-crash cases which allow a higher chance of obtaining non-crash cases compared to crash cases. The authors also proposed to check the actual crash time using traffic data and test them to choose the optimal observation time slice durations.

However, there is still place for improvement as many decisions in the study were arbitrarily taken, especially for categorizing crash precursors. For the time being, the distributions of precursors were undertaken based on the dates arbitrarily selected. The hard boundaries based on unexhausted consideration of normal traffic flow conditions (on many different days) for each precursor might not be realistic enough, as traffic is a complex phenomenon and there are normally no boundaries in real life.

2.3.2.3. Studies by Golob et al.

2.3.2.3.1. Objectives

In one of their articles (Golob and Recker, 2004; Golob, Recker and Alvarez, 2004), the authors tried to relate crash characteristics with traffic flow conditions at the time of their occurrence. The ultimate goal of the study was to devise a safety performance measurement tool that can be used to measure the effects of changes in traffic flow patterns on traffic safety that could be used to predict future conditions, or to evaluate the effectiveness of advanced transportation management projects.

In a later work (Golob et al., 2008) the objective was to capture the relationships between traffic flow(as measured by an extensive set of statistical parameters) and the type of accidents occurring under different types of traffic flow conditions. The work builds upon some of their previous studies (Golob and Recker, 2004; Golob, Recker and Alvarez, 2004)

2.3.2.3.2. Data Used and Study Site

The authors use the database of crashes that occurred on mainline sections of the six major freeways in Orange County, California, during 1998. 1'192 out of the 9'341 crashes were selected as the corresponding valid loop detector data to perform the analysis (data for a full 30 min preceding the accident for three designated lanes at the nearest detector station was available). Crash characteristics include: the type of collision, the collision factor, the number of vehicles and other parties involved, the movement of each vehicle prior to collision, the location of the collision, the object struck by each vehicle, the number of injuries, and environmental conditions. The time of each crash is not known with precision.

Traffic flow information is represented by 30 seconds single loop detector data drawn from the Vehicle Detection System in approximately 8'000 locations on California freeways typically spaced one-third to one-half mile apart, providing volume (flow) and occupancy (the time percentage a vehicle is within the detection field of a loop) for each freeway lane at 30s intervals.

2.3.2.3.3. Methodology

In another study (Golob and Recker, 2004; Golob, Recker and Alvarez, 2004), the authors identify four blocks of three variables (one measure for each of the three lane type designations, left, interior, and right) as being potentially related to taxonomy of crash. The blocks indicate the prevailing traffic speed represented by the median flow/occupancy, the temporal variation of the prevailing speed represented by the difference between 90th and 50th percentiles of flow/occupancy, the mean traffic flow, and the temporal variation in the traffic flow. As the variables are highly correlated, Principle Component Analysis (PCA) is performed to extract a sufficient number of factors to identify independent traffic flow variables while simultaneously discarding as little information as possible in the original variables. The first six principle components were selected. For each principle component, the variable contributing the most to the component is selected to represent the component. Thereafter, the clustering method - K-means - is used to find homogenous groups of traffic flow conditions, which are called *traffic flow regimes*.

In the work of Golob and co-workers (Golob, Recker and Pavlis, 2008), traffic flow data is used as a predictor to calibrate and verify probabilistic models for freeway safety performance. Traffic data at the site of the crash 20 minutes prior to the crash was used to calculate four types of parameters: coefficients of variation, correlations of traffic conditions across lanes, autocorrelation (compared to previous 30 seconds intervals), and means and standard deviations of volume and speed , producing 36 parameters. A PCA process was undertaken to reduce the number of dimension from 36 parameters to 8 factors. Based on the *loadings* of each parameter, the most important ones for each factor were recognized. Each factor with its important parameters is interpreted to have a concrete meaning. Subsequently, the 8 factors were used in logistic regression models to capture the relationship between such factors, their second-degree interactions, and the probabilities of occurrence of an event. Four accident variables were analyzed: accident severity, collision type, collision location, and number of involved vehicles.

2.3.2.3.4. Comments

A tool requiring only a stream of 30 seconds (or similar interval) observations from ubiquitous single inductive loop detectors was developed. This stream is processed to provide a continual assessment of safety, updated every interval, and based on central tendencies of flow and speed, as well as variations in flow and speed for different lanes of the freeway. The authors also provide the insights gained in relating accident and traffic flow typologies.

However, as the authors did not consider non-crash traffic conditions, how to detect pre-crash traffic conditions remains unknown. Moreover, the authors used and suggested to use 30 seconds observations as inputs for the developed model, which appears to be rather risky as 30 seconds data can have a high noise level.

2.3.2.4. Studies by Abdel-Aty et al.

2.3.2.4.1. Objectives

The main objective of the studies undertaken by Abdel-Aty et al. was to identify patterns in the freeway loop detector data that potentially precede traffic crashes. Studies of this research group were undertaken to establish the relationship between historical crashes of specific types and corresponding loop detector data for rear-end crashes (Pande and Abdel-Aty, 2005; Pande and Abdel-Aty, 2006), for sideswipe crashes (Pande and Abdel-Aty, 2006), and for both rear-end and sideswipe crashes (Abdel-Aty et al., 2004; Abdel-Aty and Pande, 2005). Moreover, the authors tested the transferability of their methodology with data collected from Dutch freeways (Abdel-Aty et al., 2008).

2.3.2.4.2. Data Used and Study Site

Most of the studies undertaken by Abdel-Aty et al. are based on a 36.25-mile road section on the Interstate-4 corridor in Orlando, Florida, United States. The section is equipped with 69 dual loop detector stations in each direction located approximately every ½ mile. These stations report speed, volume, and occupancy data every 30 seconds from the three through lanes of the corridor. Crash data for a period of five years ranging from 1999 through 2003 with the occurrences of 4189 crashes was used.

2.3.2.4.3. Methodology

The studies undertaken by the research group follow three steps:

Data preparation: defining crash and non-crash cases
Feature Selection: selecting the most important variables
Model development

Data for crashes are collected and defined according to Figure 2-3. For each crash, there is a detector station assigned to the crash and called crash station, and there are several detector stations at the upstream and downstream of the crash station. According to different studies, the number of detector stations employed varies. For example, Figure 2-3 presents the crash case definition used in (Abdel-Aty and Pande, 2005) with station F as a crash station and with five upstream stations (namely from A to E)

and one downstream station (station G). Five minute intervals were used to produce data slices. The pre-crash time interval varies from 20 to 30 minutes. For a time slice at one station, there are several data fields containing variables for three lanes. For example, in the study of (Abdel-Aty and Pande, 2005) the authors make us of the average and standard deviation of speed for each lane which produces totally 252 data fields for each crash case (7 stations x 6 slices x 3 lanes x 2 variables). The study presented in (Pande, 2005) related to rear-end crashes (chapter 5) uses 120 data fields, each field being processed differently: 5 stations (2 upstream and 2 downstream) x 4 slices (totally 20 minutes before crashes) x 3 variables x 2 ways of data aggregation over the lanes (the average and standard deviation).

Non-crash cases are generated in the same manner that crash cases are generated. Non-crash cases, in order to be used for model development, are randomly selected among non-crash cases in either the whole non-crash case population (such as in (Pande, 2005)) or in a smaller set of non-crash cases obtained by using matched-case control (such as in (Abdel-Aty, Uddin, Pande, Abdalla and Hsia, 2004; Abdel-Aty and Pande, 2005)). Finally, for each crash case, several non-crash cases were randomly selected for developing the model.

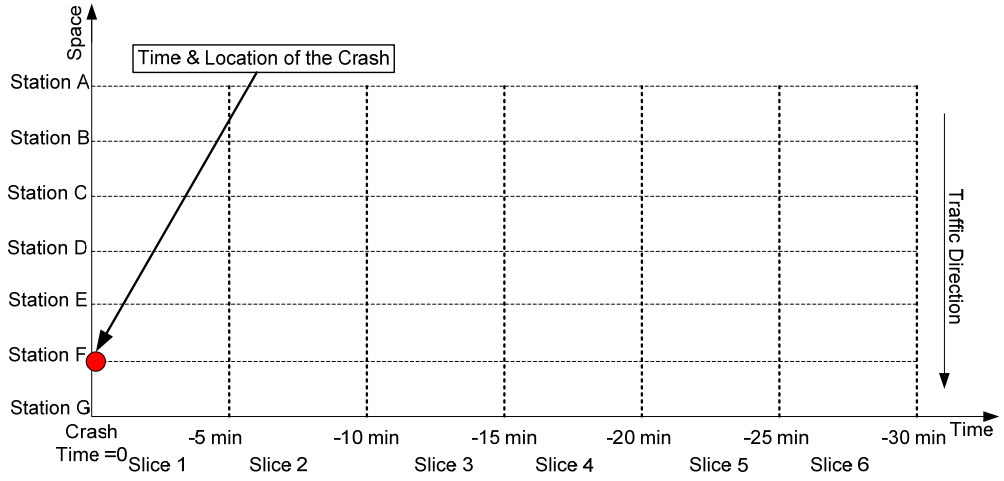

Figure 2-3: The definition of crash cases. Source: (Abdel-Aty and Pande, 2005)

With crash and non-crash cases extracted, the authors selected among a large number of variables the ones contributing the most to the differentiation between crash and non-crash cases. Depending on the studies, different techniques were used. For example, *matched case-control logistic regression* was used in (Abdel-Aty and Pande, 2005) to select three among 252 variables, whereas in (Pande and Abdel-Aty, 2006) *classification and regression trees* were used to estimate variable importance.

Finally, data including crash and non-crash cases was used as input for developing models. Different machine learning techniques were applied: Probabilistic Neural Networks (Abdel-Aty and Pande, 2005), Multi-Layer Perceptron (MPL) Network (Pande and Abdel-Aty, 2005; Pande and Abdel-Aty, 2006), or radial basis function (Pande and Abdel-Aty, 2005).

2.3.2.4.4. Comments

The authors have successfully applied a series of data mining techniques to select important variables and to develop models that differentiate crash from non-crash cases for different crash types. Crash and non-crash cases were characterized by many variables covering temporal and spatial dimensions of the pre-crash traffic evolution. Furthermore, using multiple detector stations for a crash case, the authors also developed a shockwave method for recognizing crash time more precisely.

However, the studies of this group are somehow confusing. The authors follow two different strategies in selecting non-crash cases (using matched case-control or not) and in combining detector stations (the number of detector stations used upstream) with the same study site. However, the authors did not provide a guideline about the best strategy to use. In addition, the authors re-aggregated data multiple times, which reduces the information and makes the frontier between crash and non-crash cases blurry.

2.3.2.5. Studies by Hourdakis et al.

2.3.2.5.1. Objectives

The authors (Hourdakis et al., 2006) aim at determining whether crash prone conditions can be identified and detected prior to occurrence.

2.3.2.5.2. Data Used and Study Site

The area with the highest number of crashes, a mile long section of the I-94 westbound south of downtown Minneapolis, is used as study site. Three detection and surveillance systems were deployed for capturing live crashes on video while simultaneously extracting individual vehicle speeds, headways, and classifications. 11 cameras captured and saved 12 hours of traffic every day, whereas individual vehicle measurements are collected 24/7 at 6 stations on the roadway, on a per lane basis; this represents a total of 20 detection locations. In this work, the risk identification model was developed with traffic data collected from August 2003 to January 2004. The video captures for a longer period was collected and analyzed.

2.3.2.5.3. Methodology

Thanks to the availability of the traffic surveillance video data, the authors could justify the time of crashes reported by the police. Moreover, the authors could also identify near-miss cases which represent events where no collision occurred, yet there were single vehicle crashes caused by drivers' maneuvers aiming at avoiding collisions between two or more vehicles. In total, there were 30 crashes and 122 near misses detected.

For developing the model, a number of non-crash prone conditions were identified and defined as any period of time where neither a crash nor a near miss was observed. A large number of 20-30 minute periods were randomly extracted from video surveillance data and from traffic measurements. Several factors relating to the crashes were also extracted from video surveillance data, such as the visibility conditions in terms of rain and snow, pavement condition, and sun position.

Having obtained traffic as well as video data, the authors choose three traffic flow metrics: temporal, spatial, and heuristic metrics. Temporal metrics quantifies changes in traffic characteristics over a time interval and is directly defined and derived from point measurements. The temporal metrics includes the average speed, the coefficient of variation of speed (the ratio between the standard deviation of speed and its mean over a time period), the traffic pressure (the product of density and speed variance), kinetic energy (the product of density (mass), and the square of the stream (average) speed), and the coefficient of variation of time headway. The calculation of temporal metrics is based on a moving filter over the individual vehicle speed time series. The size of the moving window is defined by the number of vehicles and varies from 15 to 120. Spatial metrics refer to metrics derived from the trajectory information of a single vehicle over a specified section of roadway and include the acceleration noise (the standard deviation of accelerations (decelerations) a vehicle performs in the space of a mile), the mean velocity gradient (a measure of the mean change in velocity per unit distance of the trip), and the quality of flow index (calculated from a vehicles average speed, absolute sum of changes in speed, and number of speed changes in a mile). As it is difficult to obtain the spatial metrics, strict assumptions were made on the speeds and headways of vehicles which are not considered to change on the one mile study site. Heuristic metrics are given by the engineering judgment and include *Max-Min-Diff* (the percentage difference in speed between the fastest and slowest vehicle in a group of *n* vehicles passing over a single detector), and *UpdownDiff* (the percentage difference in speed between the fastest vehicle at the downstream detector and slowest vehicles in the upstream one, a conservative setting). The heuristic metrics are calculated based on different sizes of vehicle groups. In addition to traffic flow metrics, three environmental factors that were also studies are: pavement condition, visibility conditions, and sun position. Globally, 1'137 parameters resulted good candidates for developing the model. By applying Backward Elimination, the authors could reduce the number of parameters to 18 most important factors.

The authors developed the model using logistic regression techniques with Maximum Likelihood Estimation, employed to determine model's coefficients. With the final model based on 18 most important factors, the authors conclude that speed variations and environment factors are two main causes of the crashes or near-miss events. The authors also tested their model in real time, for a defined number of days (not previously considered), in order to assess its true performance. From a total of 60 crashes and near-misses, the model could identify 58.33%, while from a total of 176'380 non-crash cases, the model misidentified 12'016 cases, bringing the false decision rate to 6.81%.

2.3.2.5.4. Comments

The authors possessed a large amount of data including individual traffic data, crash data, and video data. Using video data, the authors proposed the idea of justifying the exact crash time and extracting additional environmental information from video data. The idea of extracting near-miss cases was also proposed for the first time. They proposed the new idea of using a moving window of vehicles instead of moving windows of time, as commonly used in the field.

However, as the performance of the developed model is not particularly high, a lot of improvements that can be implemented. Firstly, from individual vehicle data, many more traffic related variables can be extracted, such as the percentage of heavy vehicles, the speed difference between lanes, etc. Similarly, with low spacing between detector stations, many traffic related spatial variables can be obtained instead of subjectively selected heuristic ones. Secondly, near-miss cases might have enriched the crash data set but driver's status was not justified to assure that the situations were caused by the traffic itself and not by drivers showing a strange behavior. Moreover, the model developed using logistic regression is more likely to be an exploratory study than a model applicable in real-time, specifically because of its poor performance. The authors presented a probability threshold which was subjectively determined.

Furthermore, one of inputs for the model relates to the visibility extracted from video data. The author did not explain how to capture this information in a real-time framework.

2.3.2.6. Studies by Hossain et al.

2.3.2.6.1. Objectives

The objectives of the work presented in (Hossain and Muromachi, 2010) include:

Developing models for predicting crashes in real-time.
Finding out where and how to layout the detectors to develop real-time risk identification models.
Suggest solutions when a detector is out-of-order.

2.3.2.6.2. Data Used and Study Site

Data was collected from the Shinjuku 4 Tokyo Metropolitan Expressway, in Japan. The length of the road section is of approximately 14 kilometers (for the direction bounding the downtown of Tokyo) and there are 50 detectors, 6 entries and 3 exit points. The time period for data collection was: December 2006 - November 2008.

The crash data contained 1318 cases, including information on date, time (in minutes), location, crash lane, type of crashes and vehicle involvement. Traffic data are the 5-minute station-level aggregated values of average speed and cumulative flow, for each of the 50 detectors. The traffic conditions leading to crashes were contained in 5-minute interval data, at least 9 minutes before the crash.

2.3.2.6.3. Methodology

Crash time is justified using pre-installed surveillance cameras in most parts of the expressway or using the round the clock petrol by safety vehicles.

Normal traffic conditions comparable with conditions leading to crashes were extracted at the same time of the day, and day of the week for the whole study period. Crashes occurring on weekends or during night time were simply excluded. The study taking benefit of low spacing between detector stations considers 6 combinations of traffic detectors at the upstream and at the downstream of crashes. Depending on the combination of detectors, the number of crashes considered varied from 167 to 281. Data corresponding to 30 crashes were used for the model evaluation purpose and the rest of the data was used for model building.

The authors tested the models with different combinations of one detector station upstream and one downstream of the crash location. Thereafter, for each combination of detector stations, the authors tested different combination of variables (that were derived from speed and flow at each station in order to select variables that significantly influence the differentiation between non-crash and crash cases). Variable selection was undertaken using logistic regression. Finally the differences in speed and flow were used as input variables for developing models.

Bayesian Network models were developed for each detector station combination to provide probabilities for a traffic case to become a crash case. To make a classification, a probability threshold needs to be determined.

2.3.2.6.4. Comments

The authors proposed different combinations of detector stations upstream and downstream of crash location. Whereas Abdel-Aty et al. used certain combinations of detector stations and did not compare such combinations; Hossain et al. made this comparison and recommended the best combination. The authors also suggested using the second best combination as an alternative, particularly when there is failure within the best combination (such as a detector failure).

However, there are also limitations that restrict the author's findings. First the raw datasets used are already highly aggregated in 5-minute intervals. Secondly, the number of potential variables used is limited (speed, flow, and derived values of speed and flow).

2.3.2.7. Other Disaggregate Studies

Recently, a novel trend of studies in traffic safety emerged, which utilizes disaggregate traffic flow data in which indicators of traffic risks or traffic safety are developed based on individual vehicle data, named *risk* or *safety indicators*, respectively. Below, all risk or safety indicators will be called *risk indicators*.

Time-To-Collision (TTC -(Hayward, 1971)) is one of the first indicators of this type. TTC is used to indicate the time that remains until a collision between two vehicles if the collision course and speed difference are untouched. TTC has been widely used in theoretical safety studies, and one successful application of TTC is in the context of Traffic Conflict Technique (TCT - (Archer, 2001)), which is perhaps the most developed indirect measure of traffic safety.

Derived from TTC, a series of risk indicators were invented, such as Time Exposed Time-to-collision (TET) or Time Integrated Time-to-collision (TIT) (Minderhoud and Bovy, 2001), Post-Encroachment Time (PET) (Allen et al., 1978), etc. However, most of the indicators invented were only tested in simulation models.

Several recent studies that can be listed here include the work presented in (Oh et al., 2006; Mouzon et al., 2008; Haj-Salem and Lebacque, 2009). The general objective of these studies was to create risk indicators and validate them using individual vehicle data for crash and non-crash cases. The validation results were positive as the risk indicators had different distributions in crash and non-crash cases. However, the indicators were only tested with a limited number of crashes, and more importantly, crash risks in pre-crash traffic conditions were only identified at the last minutes, which might make it impossible to prevent the crash even when the crash risk is identified.

From the traffic management point of view, there have been no models developed that were based on risk indicators such those applicable in a real-time framework. However, as the units of analyses are the individual vehicles, this direction of disaggregate studies is still promising when the vehicle-to-vehicle or vehicle-to-infrastructure communication becomes a widely spread reality.

2.4. Critical review

2.4.1. Overview

The work reviewed so far has certainly brought a significant contribution to research on traffic safety using traffic flow data. However, there several common points that can still be improved.

2.4.2. Type of Data

In disaggregate studies, datasets are a prerequisite. Different types of data are used in the previous studies. Based on the availability of such data, the objectives and methodology of studies can vary. Table 2-1 summarizes the data used in previous studies with emphasis on traffic data. There are two conditions regarding traffic data influencing the objectives and methodology of previous studies: the details of raw traffic data and the spacing between traffic detector stations.

If traffic data is collected from single loop detectors, the number of variables is limited to vehicle count and occupancy (as in (Golob and Recker, 2004; Golob, Recker and Alvarez, 2004; Golob, Recker and Pavlis, 2008) presented in section 2.3.2.3). Double loop detectors can provide more information depending on the storage configuration. Furthermore raw traffic data can be aggregate or individual vehicle data. If the raw data are aggregate, the variables are: average speed, vehicle count, and occupancy (for example, as in the studies introduced in sections 2.3.2.1, 2.3.2.2, 2.3.2.4, and 2.3.2.5). If the raw data comes from individual vehicle data, many variables such as headway, time gap, vehicle count, average speed, variations of speed, headway, and time gap, can be obtained by aggregating the data.

The spacing between traffic detector stations determines whether space-related variables can be used in developing models. In case of low spacing between detector stations, several studies make use of stations upstream and downstream of crash locations to characterize the spatial traffic evolution (as in studies reviewed in sections 2.3.2.2, 2.3.2.4, and 2.3.2.5).

As listed in Table 2-1, two research groups (Hourdakis et al. and Hossain et al.) make use of video data. The usefulness of video capturing is that crash time can be justified and environmental information such as visibility or lighting condition can be extracted. Only one group (Abdel-Aty et al.) used meteorological data.

Traffic related variables used in previous studies might be insufficiently detailed and were obtained by aggregating raw datasets at several levels (lane level or section level). Such multiple aggregation levels would reduce the frontier between crash and non-crash cases, which leads to the resulting low performance of the developed models.

Table 2-1: Types of Data used

Studies	Duration of data availability	Types of data used	Types of Raw Traffic Data	Use of multiple detector stations
Oh et al.	Feb 16 - Mar 19, 1993, 5A.M.-10A.M. & 2P.M. - 7P.M.	-Traffic -Crash	Aggregate from double loop	No
Golob et al.	Mar-Aug, 2001 (six full months)	-Traffic -Crash	Aggregate from single loop	No
Lee et al.	13 months from Jan, 1998 to Jan, 1999	-Traffic -Crash	Aggregate from double loop	Yes
Abdel-Aty et al.	1999 - 2003 (5 years)	-Traffic -Crash -Meteorological	Aggregate from double loop	yes
Hourdakis et al.	From Aug, 2003 to Jan, 2004	-Traffic -Crash -Video	Aggregate from double loop and Individual	No
Hossain et al.	Dec, 2006 to Nov, 2008	-Traffic -Crash -Video	Aggregate from double loop	Yes

2.4.3. Traffic-Related Variables

Depending on the raw data and on the availability of multiple detector stations, the variables used and the number of variables vary in previous studies. Some studies additionally used non-traffic variables, such as time of day, visibility, etc. However, only traffic-related variables are discussed in this section.

There are two views from which traffic characteristics are observed, the temporal and the spatial view. From the temporal view, traffic characteristics are quantified in variables such as average of speeds, variation of headways, etc., within a certain time interval (which is mostly a 5-minute interval in previous studies). These are the variables measuring the evolution of traffic over the time interval.

There are also variables measuring traffic evolution over space, which can be related to laterally crossing the lanes of a section or to a certain road distance. Normally, raw traffic datasets are stored based on lanes. The data can be aggregated from individual lanes to create more aggregate variables representing traffic characteristics of the whole section. In this case, traffic data from only one detector station is used and the newly-created variables are called station-based. If more than one station is used, corresponding variables characterize the traffic evolution along the road section where detector stations are used. It is worth noting that the temporal aggregation is inclusive in spatial aggregation.

The availability of variables and the raw data aggregation interval are important factors influencing the set of potential variables in previous studies. For example, Hossain et al. had limited choices in creating new variables because the aggregation interval in raw data was already on a 5-minute basis and the variables were station-based. On the contrary, Abdel-Aty et al. had more choices as their raw data was aggregated in 30 seconds from which 5-minute variables were created by re-aggregating 30-second data. Hourdakis et al. have more potential to create variables as they possessed individual data. Table 2-2 summarizes the variables and the raw data used in previous studies.

Table 2-2: Variables and aggregation intervals in raw data

Studies	Variables	Aggregation Interval
Oh et al.	Lane-based volume, occupancy, and average speed	10 seconds
Golob et al.	Lane-based volume and occupancy	30 seconds
Lee et al.	Lane-based volume, occupancy, and average speed	20 seconds
Abdel-Aty et al.	Lane-based volume, occupancy, and average speed	30 seconds
Hourdakis et al.	Individual vehicle data	-
Hossain et al.	Station-based volume and average speed	5 minutes

From the raw data characterized by the variables listed in Table 2-2, new ones were created and called *potential variables*. We define the following groups of potential variables:

- LB - Lane based. Variables belonging to this group can be obtained using traffic data of one lane and do not relate to other lanes at one detector station.
- AL - Aggregation over lanes. Variables belonging to this group are calculated based on data from more than one lane at the same detector station.
- DS - Differentiation between stations

The first two groups (LB and AL) are station-based (i.e. variables belonging to these groups are calculated based on data from one detector station). Examples of lane-based variables include: average speed, standard deviation of speed, average headway, etc. The variables representing the aggregation over lanes include the average speed or the volume over two or three lanes. Variables representing differentiation between lanes such as speed difference, flow difference, etc. also belong to group AL. Variables representing differentiation between stations involve at least two detector stations. For example, the differences between the average speeds and volumes of two stations (one upstream and one downstream of crash location) are classified in the DS group.

To obtain the potential variables, original variables in the raw data are re-aggregated (here the term *re-aggregate* is used as in most of the previous studies (except the studies by Hourdakis et al.), the raw data being already aggregated). Depending on the raw data, three re-aggregation levels can be undertaken to obtain potential variables. The first re-aggregation level (designated *level 1*) relates to the aggregation of raw data from the time interval of interest (which is 20 minutes for studies of Golob et al., 2, 5, or 8 minutes for Lee et al., vary according to Hourdakis et al., and 5 minutes for the others). Variables belonging to group LB are obtained from the re-aggregation level 1. The second re-aggregation level (*level 2*) generates variables using the variables obtained from level 1. The second level can be undertaken by aggregating data over lanes (for example, the average speed over three lanes) or over stations (for example, the speed difference on each lane between two stations). There is also a third re-aggregation level (*level 3*), such that variables obtained at this level are generated by aggregating variables obtained in level 2. Variables of the third level can be based on one station (such as the

coefficient in variation of speed over three lanes) or based on two stations (such as the difference of station-based average speeds between two stations).

The different levels of data re-aggregation used in previous studies are listed in Table 2-3. It is worth noting that different re-aggregation levels were undertaken by the authors except for Hossain et al. whose raw data was already aggregated on level 2.

The re-aggregation levels represented in Table 2-3 do not indicate the quality of the aggregating operation. Most of the aggregating operations reduce the information while others can bring more information. For example, averaging lane-based speeds to obtain the average speed at a station reduces speed information. However, subtracting the lane-based average speeds from two stations can bring additional information on speed variation between stations on one lane.

In general, the more the traffic data are re-aggregated, the more the distinction between crash and non-crash cases becomes blurry. Among previous studies, there are two study groups (Golob et al. and Hourdakis et al.) who still keep variables obtained from the first re-aggregation level. Three research groups re-aggregated data at the third level. Lee et al. and Hossain et al. created new information which includes differences of flows (by Hossain et al. only) or speeds between two stations, whereas Abdel-Aty et al. divided station-based standard deviation of speed to station-based average speed to obtain station-based coefficients in speed variation.

Table 2-3: Scope of potential variables and re-aggregation level in previous studies

Studies	Variables Scope			Re-aggregation Levels
	LB	AL	DS	
Oh et al.		Yes		2
Golob et al.	Yes			1, 2
Lee et al.		Yes	Yes	2, 3
Abdel-Aty et al.		Yes		2, 3
Hourdakis et al.	Yes		Yes	1, 2
Hossain et al.		Yes	Yes	2, 3

We define four types of variables, from VT1 to VT4, presented here in Figure 2-4 based on different traffic parameters such as speed, flow, headway, percentage of heavy vehicle, etc. Variables of type VT1 belong to the LB group, i.e. representing traffic characteristics within lanes. VT2 includes variables linking different lanes. Group VT3 includes variables linking consecutive time intervals, and group VT4 linking traffic characteristics between different stations. In previous studies, there was no variable used for characterizing traffic evolution between different time intervals.

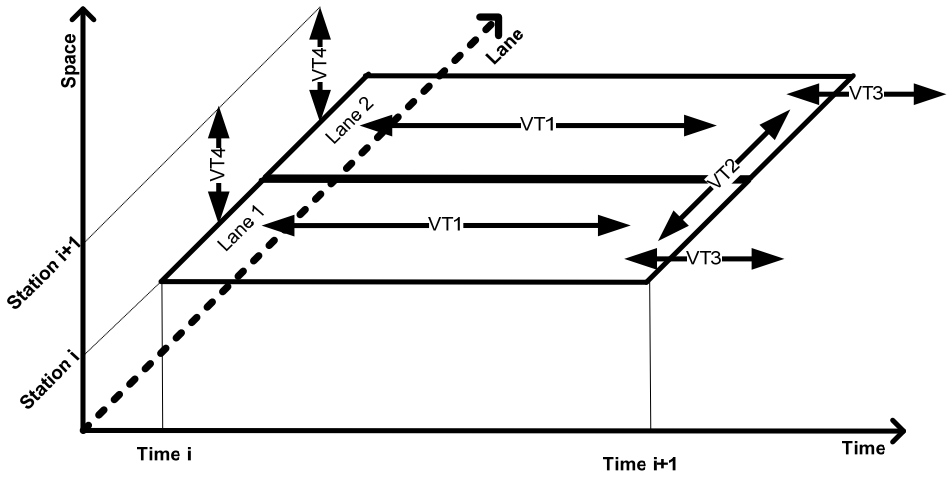

Figure 2-4: Types of variables

2.4.4. Relevancy of Selected Non-crash Data

Most of the models developed are used for assessing the risks of two crash types: rear-end and sideswipe. Rear-end and sideswipe crashes usually occur when the traffic flow is high, when the drivers cannot keep a sufficiently safe headway from the front vehicles, or when it is difficult to find a gap on the adjacent lane for overtaking. This mean that rear-end and sideswipe crashes can only occur under certain circumstances and it is less probable for crashes of these types to occur under other traffic conditions. Therefore, the non-crash cases used to compare with pre-crash cases should be selectively extracted from the set of all non-crash cases. Choosing irrelevant non-crash cases might mislead the development of a model and the subsequent interpretation of the results. As a consequence, if non-crash data are not properly selected, the applicability of the developed models can be questionable.

Techniques for selecting non-crash data are presented in Table 2-4. All data that is not related to crashes is considered as *potential non-crash data*. Several studies limit the potential non-crash data to a smaller set based on the control of factors such as crash location, time of the day, day of the week, weather conditions, etc. (called *external factors*). This means that a reduced set of all possible non-crash data having the same external factors as crash data is generated. Lee et al., Abdel-Aty et al. (in some of their studies), and Hossain et al. use this control, whereas there was no control on potential non-crash data in studies by Hourdakis et al. and in some of the studies by Abdel-Aty et al.

Thereafter, non-crash cases are selected from potential non-crash data for the comparison with crash cases to develop models. Lee et al., Abdel-Aty et al., and Hourdakis et al. selected non-crash cases at random, whereas Hossain et al. used all available non-crash cases.

Here, it can be seen that the selection of non-crash cases is arbitrary even with the control of external factors. For the selection with control of external factors, there is no guarantee that controlled non-crash cases are comparable with crash cases. Moreover, there is an uncertain assumption that there was no unrecorded incident that changed the traffic during the selected non-crash cases. For the selection without control of external factors, the comparison could be worse if non-crash cases were within post-crash periods or within periods where the traffic flow was low.

Table 2-4: Selection of non-crash cases

Studies	Non-crash Selection	Controlled?
Oh et al.	Fixed at 30' before crashes	-
Golob et al.	-	-
Lee et al.	At random	Yes
Abdel-Aty et al.	At random	Yes for some studies, no for some other studies
Hourdakis et al.	At random	No
Hossain et al.	Use of all possible non-crash cases	Yes

2.4.5. Data Imbalance

The imbalance of data sets is currently a well-known issue in data mining (Chawla et al., 2004). Data imbalance usually results in the low performance of traditional machine learning techniques (including classification and regression techniques). Among two data sets representing two classes, the data set with a much higher population is called *major class* whereas the other data set is called *minor class*. The minor class is the class of interest. In traffic safety studies using disaggregate traffic flow data, the minor class is the class of crash cases. The population of crash cases is much lower than the population of non-crash cases. In the literature, the ratio between the populations of the major class and the minor class is called *Imbalance Ratio* (*IMRO*).

Three potential solutions (two at data level and one at algorithmic level) for developing models capable of differentiating non-crash and crash cases are down-sampling, up-sampling, or use a modification of learning methods. As the population of the major class is high, the down-sampling techniques try to sample a subset of the major class such that the population of the subset is similar to the population of the minor class. On the contrary, up-sampling techniques try to clone instants of the minor class so that the final population is similar to the population of the major class. Another solution is to modify the learning method, as the modification of classification Random Forests presented in (Chen et al., 2004).

It can be seen that most of the reviewed studies (except the studies by Golob et al.) used down-sampling technique to sample non-crash data although the imbalance ratios were different. Table 2-5 presents different IMRO values used for developing models in previous studies. It is worth noting that data for validation purpose might have different IMRO values, and only IMRO values for training models are important for model's performance.

In most studies an IMRO value of 1:1 is used (i.e. Oh et al., Lee et al., and Abdel-Aty et al.). In the other ones, IMRO values were arbitrarily determined based on the availability of non-crash data.

Table 2-5: Imbalance Ratios in previous studies

Studies	IMRO
Oh et al.	1:1
Golob et al.	-
Lee et al.	1:1
Abdel-Aty et al.	*n*:1 where *n* varies
Hourdakis et al.	*n*:1 where *n* was not stated
Hossain et al.	*n*:1 where *n* =95.4; 92.1; 91.3; 92.3; 89.8; and 99.9

2.4.6. Data Mining Techniques

Various data mining techniques were used in previous studies and can be divided into two groups: techniques for feature selection and techniques for model development.

Feature selection is the process of choosing the most important features to be used in model development. Previous studies usually followed three steps in developing models: listing potential variables, selecting important variables, and developing models. Based on raw traffic data and based on the space between detector stations, the number of potential variables may vary. Several techniques were used to select the most important variables, as listed in Table 2-6.

Table 2-6: Data mining techniques used in previous studies

Studies	Feature Selection	Model Development
Oh et al.	*t*-statistics	Probabilistic Neural Networks
Golob et al.	Principal Component Analysis	-
Lee et al.	Arbitrary	Categorization
Abdel-Aty et al.	Logistic regression; classification and regression trees; Random Forests	MPL neural networks, probabilistic neural network
Hourdakis et al.	Backward Elimination	Logistic regression
Hossain et al.	Logistic regression	Bayesian Networks

Two advanced feature selection methods were used by Abdel-Aty et al.: Classification and Regression Trees, and Random Forests. The remaining work, and the techniques used for model development were

rather traditional. In this regard, traditional machine learning techniques would support imbalanced data sets badly, (such techniques provide low performance for minor classes). As crashes are rare events on motorways, compared to non-crash events, the aforementioned techniques would perform poorly if non-crash data were more properly sampled. Another issue that traditional techniques suffer from is that potential variables are highly mutual correlated. The correlativity in data would mislead the listed techniques and result in the wrong identification of important factors and would ultimately lead to wrong results.

2.4.7. Performance of the Approaches

The performance of existing approaches cannot be verified under the conditions of the current research, because the data required as input in approaches is not available. Therefore, the performance discussed in this section relates to each of the approaches and is summarized in Table 2-7.

In Table 2-7, the missed alarm rate is the percentage of pre-crash cases that are incorrectly identified as non-crash, and the false alarm rate is the percentage of non-crash cases incorrectly identified as pre-crash. In general, a regression approach would require three data sets including training, calibration, and validation. The training data set is used for establishing the relationship between pre-crash and non-crash cases. The calibration data set is used for defining regression thresholds aiming to classify new traffic cases into pre-crash or non-crash. The validation data set can be considered as the set of new traffic cases which are tested with the developed model together with the defined thresholds. In a classification approach, training and validation data sets are required.

It is important to note that the missed alarm rate and false alarm rate presented in Table 2-7 are only applied for the validation data set. As there is only one data set used the first two studies (Oh et al, 2001 and Lee et al, 2003), missed alarm and false alarm rates are not available.

Table 2-7: Performance of existing approaches

Studies	Missed Alarm	False Alarm	Note
Oh et al, 2001	-	-	One data set for Training, calibration, and validation
Lee et al, 2003	-	-	One data set for Training, calibration, and validation
Hourdakis et al, 2006	41.67	6.81	One data set for calibration, and validation
Abdel-Aty et al, 2005-	26.10	30.00	Two data sets for training and validation
Pande et al, 2005-2007	26.00	34.00	Two data sets for training and validation
Hossain et al, 2010	36.67	20.00	One data set for calibration, and validation

In the other studies presented in Table 2-7, the best reported performance of the corresponding approaches is presented. The main objective of the above-mentioned approaches is to detect risky traffic conditions. Therefore, obtaining missed alarm rates as low as possible is better than having low false alarm rates.

The best missed alarm rate in Table 2-7 is 26.00% (Pande et al, 2005-2007), with a maximum of only 74% of pre-crash cases can be identified. The cost for that low missed alarm rate is 34.00% of the non-crash incorrectly identified. Hourdakis et al, 2006 offer an approach having a low false alarm rate of 6.81%, yet on the other side, the missed alarm rate is really high (41.67%).

2.4.8. Applicability for Crash Prevention in Real-time

The applicability of the developed models in term of crash prevention is based on the following factors:

Is the input data easy to get in real-time?
When the crash risk is identified, is it possible to activate preventive measures to clear or diminish the risk?

Only when inputs for the models are obtainable *online* and the crash risks are identified a certain time before crash occurrence, the models for crash prevention in real-time are applicable. Here, the performance of the models is assumed to be high. The model developed by Hourdakis et al. requires input from video data that cannot be automatically extracted. Therefore, it is not applicable online. The models developed by Oh et al. and Lee et al. are not applicable they identify the crash risk right before the crash occurs; therefore the crash is not preventable. The model developed by Golob et al. is not applicable as it does not differentiate between crash and non-crash case.

Table 2-8: Applicability of develop models in crash prevention

Studies	Inputs	Ready to Prevent	Applicable
Oh et al.	Yes	No	No
Golob et al.	Yes	No	No
Lee et al.	Yes	No	No
Abdel-Aty et al.	Yes	Yes	Yes
Hourdakis et al.	No	No	No
Hossain et al.	Yes	Yes	Yes

2.5. Conclusions from Literature Review

Driving is a complex task. In most of the cases, it is driver's behaviors that make traffic unsafe. Among unsafe driving behaviors, lapses, errors, and violations are potential driving situations that can lead to collisions. Lapses and errors are driving misbehaviors that depend on individual drivers and cannot be

influenced when the driver is on the road. Violations are intentional actions prone to a certain level of collision risk. Fortunately enough, violations are spontaneous and can therefore be influenced by smart-design of the road environment.

Motorway traffic safety studies using traffic flow data aim to identify traffic conditions that can lead to collisions. The ultimate objective of such studies is to provide a prediction on traffic crash risks such that certain smart designs can be activated, helping to clear crash related risks or to diminish collision consequences. However, before attaining that objective, it is necessary to differentiate risky traffic conditions. In the literature, there are studies that attempt to address this challenge. In spite of the significant contributions by existing studies, the following issues can be undertaken to improve the state of the art:

i. Non-crash data should be systematically used in model development where relevant non-crash data is selected to be compared with pre-crash data. There is a need for a methodology to sample non-crash relevant data with pre-crash data.
ii. There is a need for a new methodology in order to develop risk identification models with better performance. Existing methodologies offers models with high missed alarm and false alarm rates. The new methodology should improve the model performance by reducing these two alarm rates.
iii. Existing approaches neglect the data imbalance problem, which might be a serious shortcoming that limits the performance of developed models. The new methodology needs to address this problem.
iv. Existing approaches are limited to crash risk assessment (i.e. verify whether the last traffic case is risky or not) and do not provide any prediction on future crash risks. Thereby, it is unclear how to implement countermeasures aiming to reduce the crash risk. Analysis on crash likelihoods would increase the confidence of countermeasure activation decision.

In the next chapters, the above risk related issues will be tackled.

Chapter 3 Methodology

The methodology of the current research is here introduced in order to fill the gaps in the state of the art of Chapter 2. The choices of techniques at different methodological phases are discussed in order to improve the performance of the models developed using the proposed methodology.

3.1. Overview

As a result of the critical review analysis presented in Chapter 2, the methodology to be established needs to confront issues that were either not properly addressed or could be improved:

- Non-crash data sampling: selection of non-crash traffic data that are comparable with and relevant to pre-crash traffic data,
- Performance improvement: imbalance between non-crash and pre-crash traffic data and performance improvement of risk identification models, and
- Real-time applicability

Thus, the methodology must include the main steps illustrated in Figure 3-1.

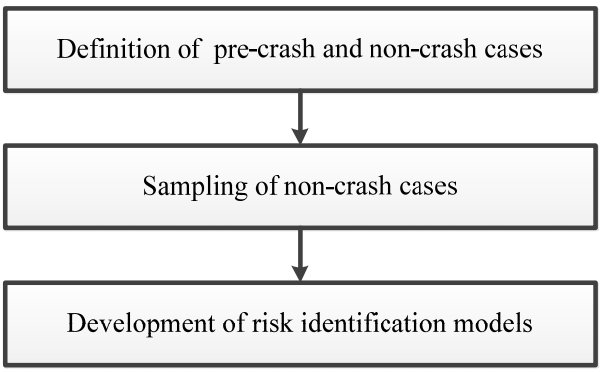

Figure 3-1: Mains methodological steps

Once the methodology is applied to the input data, pre-crash and non-crash cases will be defined. Thereafter, non-crash cases will be sampled such that only relevant non-crash cases are used and compared with pre-crash cases. Finally, risk identification models will be developed based on pre-crash cases and relevant non-crash cases. The final result given by applying the methodology is a set of models capable of identifying real-time traffic crash risks on motorways. The three main methodological steps are respectively discussed in sections 3.2, 3.3, and 3.4.

3.2. Traffic Situations - TS

3.2.1. Definitions

A *Traffic Situation (TS)* is an ensemble of instant information indicating the traffic state and external factors that might have influences on the traffic state from the traffic safety point of view.

As a TS is characterized by instant information, the time duration employed for the characterization is called *aggregation time interval*. The definition of TS is applied for one traffic direction only, because the traffic state on each traffic direction may vary.

3.2.2. Traffic State

According to Figure 2-4, there are four types of traffic-related variables that can be used to characterize the temporal traffic state:

- Type 1: traffic state on each lane (i.e. average speed, volume, occupancy during aggregation time interval).
- Type 2: traffic variation over lanes (i.e. over-lane speed difference).
- Type 3: traffic variation over aggregation time intervals, and
- Type 4: traffic variation over road sections

Among these four type variables, the first three ones are measurable using one traffic detector station, whereas the last type variable is only measurable when low spacing detector stations are available. Moreover, aside from variables of type 1 that characterize the traffic on each lane, all variables of other types should bring additional information about temporal and spatial traffic variations.

3.2.3. External Factors

External factors are all of the non-traffic factors influencing the traffic state; here can be many. This section inexhaustibly discusses these factors, especially the ones that can influence the traffic safety.

3.2.3.1. Instantaneity

According to the statistics in section 4.5.1, the time of the day and the day of the week can be temporal factors having influences on traffic occurrences.

3.2.3.2. Weather Conditions

Weather effects on traffic and traffic safety are reported in many studies, such as Satterthwaite (1976), Edwards, (1999), and more recently Andrey, (2010). The effects can come from many meteorological factors, such as wind, precipitation, or temperature.

3.2.3.3. Pavement Conditions

The assessment of the effect of pavement conditions on traffic safety is discussed by Mayora et al, (2009). Pavement conditions mentioned in this section is mainly related to the contact between pavement surface and vehicle tires via skid resistance, and influenced by temporal external conditions such as rain. When the pavement becomes wet, snowy, or icy, the skid resistance is reduced, causing longer braking. Pavement surface damages are not taken into account in the present study.

3.2.3.4. Visibility

There is a wide range of alternatives to characterize the visibility depending on the factors influencing it. The lighting conditions related to night, day, facing direction with the sun, etc. can affect the visibility. The weather can also influence visibility: heavy rain and fog reduce it dramatically.

3.2.3.5. Other External Factors

Many other external factors can influence motorway traffic safety, rock avalanches, the occurrence of the first accident, and the collapse of tunnels (in case the motorways through tunnels), just to name a few. However, these are rare events in motorway traffic crashes and are difficult to quantify. Therefore, such external factors are not considered as part of traffic situations.

3.2.4. Types of TS

Two TS types are distinguishable in this research based on historical crash occurrences: *non-crash TS* and *pre-crash TS*. The definitions are presented below.

For each crash, there is a duration called *crash period,* where the traffic conditions relating to the crash before and after occurrence are illustrated in Figure 3-2. The crash period is divided into three parts: the traffic conditions after the crash called *post-crash period*, the traffic conditions right before the crash called *pre-crash period* and the traffic conditions at the start of the crash period called *pre-crash buffer period*.

The time period outside of all the crash periods is called *non-crash period*. Here, there might still be traffic conditions where traffic risks were high but did not end up in crashes. However, there is no reason to justify these traffic conditions and hence, they are neglected.

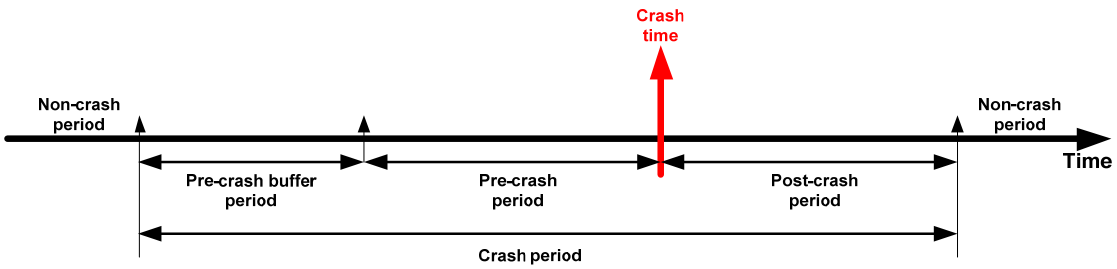

Figure 3-2: Separation of non-crash and crash-related traffic conditions

It is worth noting that for a new TS, which can be a TS in real-time or in validation data set, its status as non-crash or pre-crash is unknown and needs to be identified applying the methodology presented in the present chapter.

3.2.4.1. Pre-crash TS (PTS)

PTS are TS within pre-crash periods. Traffic conditions right before crashes are known to be truly risky as they obviously are followed by crashes. PTS are used to characterize such traffic conditions.

3.2.4.2. Non-crash TS (NTS)

A non-crash TS (NTS) is a TS appearing in non-crash periods.

3.2.4.3. Unused Traffic Intervals

Post-crash periods are not used as the traffic conditions are abnormal. Pre-crash buffer periods are also unused as they might contain both NTS and PTS.

3.3. Data Sampling

3.3.1. Motivation

As reported in sections 2.4.4 and 2.4.5, there can be two issues related to the selection of non-crash data in the comparison with pre-crash data developing risk identification models. Two issues are:

1. In some studies, non-crash data are arbitrarily selected. The potential consequence is that the developed model might be misled.
2. In other studies, the imbalance between non-crash and pre-crash data, because crashes are rare events that reduce the performance of developed models.

To prove the data imbalance issue, data collected at study sites presented in section 4.4 (Chapter 4) are used to generate PTS and NTS, following the steps presented in section 5.2. The imbalance ratio between PTS and NTS is 1:1'612.27 (i.e. in average 5 pre-crash minutes can be observed every 5.6 days). By ignoring the data imbalance issue, all NTS and PTS are used to develop a risk identification model using Random Forest (Breiman, 2001). The results are summarized in Table 3-1. The developed model performs well with training data sets. However, its performance with validation data of PTS is unacceptably low.

Table 3-1: Low performance foe to data imbalance issue

Training (%)		Validation (%)	
NTS	PTS	NTS	PTS
100.00	100.00	100.00	12.50

Concerning the arbitrary selection of non-crash data (in comparison with pre-crash data), a test is implemented with all the non-crash data sampling methodologies applied in the previous studies. Results show that existing methodologies do not perform well with validation data sets (for both NTS and PTS). The results are summarized in Table 8-2.

3.3.1.1. Objectives

Here, data sampling is the process of matching relevant NTS with PTS such that matched NTS are comparable with PTS. As most of rear-end and sideswipe crashes occurred during high flow conditions (see Figure 4-14), one of the expected results from this process is that NTS occurring during free flow conditions are eliminated and unused for comparison with PTS. Moreover, the right NTS are selected for comparison with PTS.

3.3.2. Sampling Steps

3.3.2.1. Overview

The main idea of NTS sampling process is to cluster NTS into a certain number of NTS groups. Each group is represented by the group center. Thereafter, PTS are classified into NTS groups. A PTS is classified into an NTS group if the PTS is closer to the center of that group than to the center of any other group.

Here, NTS and PTS are characterized by different variables depending on the input data such as traffic volume, average speed, occupancy, etc. There can be two issues regarding these variables:

- Issue 1: The ranges of the variables are different. For instant, the occupancy can range from 0.0 to 100.0 (in percentage), and the average speed from 0 to 150 (in km/h). In the current research, the roles of all variables are equally considered. Therefore, in order to avoid the greater influence of variables having greater value ranges, it is essential to normalize the values of the variables.
- Issue 2: The number of variables can be rather large (depending on input data). As traffic situations are represented under the form of vectors, the number of vector dimensions is high when more variables are needed. The high number of dimensions together with the high number of traffic situations would make up a huge data matrix, which would cause overloaded computer memory and would require huge amount of calculation time. Thereby, to facilitate the clustering process, the number of vector dimensions is reduced before clustering.

As illustrated in Figure 3-3, the NTS sampling is undertaken in three steps (represented by solid border rectangle boxes):

i) normalizing and reducing dimensions of NTS,
ii) clustering NTS input vectors into groups, and
iii) classifying PTS input vectors into N groups of NTS.

We name X_{NTS} a matrix whose rows are NTS and whose columns are variables representing NTS. The number of X_{NTS} rows is R and the number of columns is D.

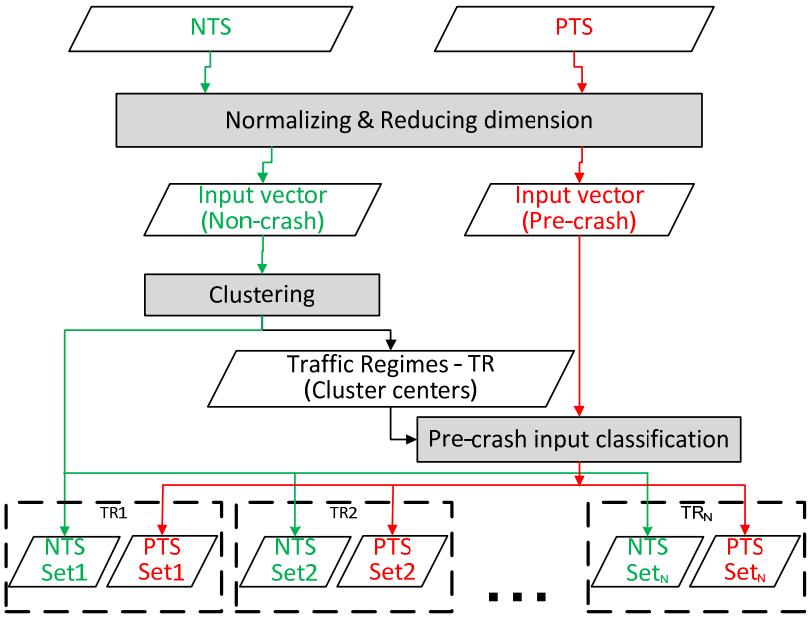

Figure 3-3: NTS sampling steps

Further details on steps i), ii), and iii) are discussed in sections 3.3.2.2, 3.3.2.3, and 3.3.2.4, respectively.

3.3.2.2. Normalization and Dimension Reduction

Dimension reduction in statistics can be divided into feature selection and feature extraction. With a given input data set, the variables used for characterizing traffic situations are selected. The number of variables can be rather high, and therefore feature extraction is applied to map traffic situation data to a lower dimensional space such that the variance of the resulted data is maximized.

Many feature extraction methods, both linear and non-linear, can be good candidates for reducing the dimensions of traffic situations in the current research. With a specific traffic data set, the following choices can be made:

i) Choice of normalization mothod
ii) Choice of feature extraction method
iii) Choice of D' - number of dimensions after dimension reduction ($D'<D$)

If *Norm* and *FEx* are normalization and feature extraction methods, respectively, then:

- $\bar{X}_{NTS} = Norm(X_{NTS})$
- $X'_{NTS} = FEx(\bar{X}_{NTS})$.

Results of the dimension reduction step include the functions *Norm* and *FEx* and a data matrix – X'_{NTS} whose number of rows is equal to the number of NTS – R and the number of columns D' is smaller than the number of variables used for characterizing NTS - D.

3.3.2.3. NTS Clustering

X'_{NTS} is used as input of the clustering method. There exist also many clustering methods potentially applicable in the current research. When the methodology is applied to a particular set of traffic situations, the following choices need to be made regarding to NTS clustering process:

i) Choice of clustering method
ii) Choice of N - number of clusters

Results of NTS clustering process include:

A set C of N cluster centers: $C=\{C_1, C_2, \ldots, C_N\}$. C_i with $0< i \leq N$ being vectors with D' elements.
A column vector I having R elements: $I=<I_1, I_2, \ldots, I_R>$ with I_j $(0 < j \leq R)$ being integers and $0 < I_j \leq N$. The value of the j-th element of the vector I represents the index of the cluster from which the j-th NTS is clustered into.

3.3.2.4. PTS Classification

X_{PTS} is the matrix of PTS. $\bar{D}$ and $\bar{R}$ are respectively the number of dimensions and the number of rows of X_{PTS} $(\bar{D} = D)$. With the function *FEx* and a fixed C obtained from previous NTS sampling steps, PTS are classified as below:

- Calculate $\bar{X}_{PTS} = Norm(X_{PTS})$
- Calculate $X'_{PTS} = FEx(X_{PTS})$
- For a k-th row of X'_{PTS} $(0 < k \leq \bar{R})$, the k-th PTS is classified into a i-th cluster if the distance between the k-th row of X'_{PTS} and C_i is smaller than the distance between the k-th row of X'_{PTS} and C_l with any $l \neq I$ *and* $0 < l \leq N$.
- Find values of $\bar{I}_k$ $(0 < k \leq \bar{R})$ that are indices of clusters which the k-th PTS are classified into $(0<\bar{I}_k \leq N)$.

3.3.3. Results

Finally, at the end of the NTS sampling process, N clusters are obtained. Under each cluster i, there are:

A set of NTS which are the j-th NTS clustered into the i-th cluster, i.e. all j such that $Ij = i$.
A set of PTS which are the k-th PTS classified into the j-th cluster, i.e. all k such that $\bar{I}_k = j$.

As a result of the clustering process, traffic situations, including both NTS and PTS under one cluster, are more similar when compared to traffic situations under other clusters. As traffic situations are characterized by traffic state and by external factors, clusters are called *Traffic Regimes (TR)*.

It is important to note that normalization and dimension reduction are only applied to match NTS with PTS. Once Traffic Regimes are determined, NTS and PTS represented by different variables are reused for further analyses.

3.4. Model Development

3.4.1. Motivation

One of the main motivations for developing a new methodology is to improve the performance of risk identification models in comparison to the performance of models developed using methodologies in the literature. Section 3.3.1 also introduces the fact that data sampling can contribute to improve the performance of the developed model by selecting the relevant NTS to be compared with PTS. However, data sampling is not sufficient as the model performance also depends on the machine learning method used to develop the model.

The modeling technique used should be able to overcome the following challenges:

- Certain variables characterizing traffic situations can be categorical. An example of this variable type is the time of the day and the day of the week.
- Although data sampling is applied, there is still an imbalance between NTS and PTS under each traffic regime.

In the next sub-sections, the steps to select the most suitable machine learning technique are discussed.

3.4.2. Supervised Learning

Subsequent to the data sampling process, there are a set of NTS and a set of PTS under each traffic regime (TR). A TR-based Risk Identification Model (RIM) is developed in order to differentiate NTS and PTS under that TR.

As TR-based RIM are trained using available NTS and PTS, the learning technique is a supervised one. In supervised learning, data used for model development are labeled with predefined classes. For the current research, traffic situations (TS) in training data include NTS and PTS classes. Here, there are two options of supervised techniques for the TR-based RIM development:

Classification approach to classify TS population into two classes: NTS and PTS.
Regression approach to estimate the probability for a TS to be PTS (or NTS).

The regression approach can also be converted to a classification approach by defining a probability threshold to separate probabilities into NTS and PTS zones.

3.4.3. Data Imbalance Issue

Wang and Wu (2006) identified that one of the challenges of supervised learning approaches to develop TR-based RIM is the imbalance between the populations of two classes (i.e. the population of NTS, called

major class, is much higher than the population of PTS, called *minor class*). The Imbalance RatiO - *IMRO*, which relates to the population of the major class divided by the population of the minor class, is usually used for indicating the imbalance, as introduced in section 2.4.5. In data mining, learning imbalanced data sets is gaining increasing interest (Chawla et al., 2004) as traditional learning techniques only perform well with data sets of equal or approximately equal sizes.

One of the widely recognized effects of the high IMRO value on learning technique performance is that models developed with imbalanced data sets seem to classify all data into the major class, whereas it is the minor class that is of interest. Even though the overall performance of the developed models is high, the percentage of observations of the minor class that are correctly classified is rather low.

X is a matrix of input data with $X_1, X_2 \ldots X_M$ as column variables and rows $x_1, x_2 \ldots x_P$ as observations, and with Y as a column vector of outputs such that y_i is the known output of the observation x_i. For the classification problem, y_i is a class label, whereas for the regression problem, y_i is a numerical value. As such, the set of training examples is $\{(x_1, y_1), (x_2, y_2) \ldots (x_P, y_P)\}$. A supervised learning algorithm is a function g which approximates an original function $f: X \rightarrow Y$ where $y_i = f(x_i) + \varepsilon$ with ε representing data noise in measuring y_i and x_i. The output of g with x_i as an input is an estimated value of $y'_i = g(x_i)$. Suppose that the function $d(y_i, y'_i)$ measures the discrepancy between the real output y_i and the estimated output y'_i. The best function g should minimize the training error, as presented in Equation 1.

Equation 1: the training error

$$\mathrm{Error}_{\mathrm{training}} = \sum_{\mathrm{i} \in \mathrm{TrainingSet}} \mathrm{d}\left(\mathrm{y_i}, \mathrm{y'_i}\right)$$

Where the *TrainingSet* is the set of indices i used for training. If there is some parameter to be tuned in function g, another index set called *CalibrationSet* is used. If the developed model is used for predicting new values in the future, the index set called *ValidationSet* is used. The CalibrationSet and ValidationSet index sets are discussed and used later on.

When using regression approaches, the sum of squared errors can be used for estimating $Error_{Training}$ as presented in Equation 2.

Equation 2: Training error using sum of squared errors

$$\mathrm{Error}_{\mathrm{training}} = \sum_{\mathrm{i} \in \mathrm{TrainingSet}} \left(\mathrm{y_i} - \mathrm{y'_i}\right)^2$$

As the TrainingSet can be divided into the NTS and PTS sets called TrainingSetNTS and TrainingSetPTS, respectively, Equation 2 can be rewritten as presented in Equation 3.

Equation 3: Sum of training errors combined from NTS and PTS classes

$$\mathrm{Error}_{\mathrm{training}} = \sum_{\mathrm{i} \in \mathrm{TrainingSetNTS}} \left(\mathrm{y_i} - \mathrm{y'_i}\right)^2 + \sum_{\mathrm{i} \in \mathrm{TrainingSetPTS}} \left(\mathrm{y_i} - \mathrm{y'_i}\right)^2$$

The proportion of PTS is low compared to that of NTS; the training error due to PTS class is also lower than that caused by NTS class. As consequence, the sum of training error of both PTS and NTS mostly reflects the training error due to NTS. If least squares method is used to train a regression learner, the obtained learner (i.e. the developed model) will neglect the presence of PTS and result in low PTS detection rate.

Therefore, the performance of a model should be estimated based separately on the error due to PTS class - $\mathrm{ErrorPTS}_{\mathrm{training}}$ and the error due to NTS class - $\mathrm{ErrorNTS}_{\mathrm{training}}$ as presented in Equation 4.

Equation 4: Separate training errors due to the NTS and PTS classes

a) $\mathrm{ErrorPTS}_{\mathrm{training}} = \sum_{i \in \mathrm{TrainingSetPTS}} \left(y_i - y'_i\right)^2$

b) $\mathrm{ErrorNTS}_{\mathrm{training}} = \sum_{i \in \mathrm{TrainingSetNTS}} \left(y_i - y'_i\right)^2$

3.4.4. Prediction Power Issue

One of the objectives in the current research is to estimate the traffic crash risk in real-time. Therefore, the prediction power of supervised learning methods plays a vital role in the methodology. Together with the performance for imbalanced data sets, the prediction power is a criterion for choosing a learning method.

Equation 2 presents the training error with training data. To estimate the prediction error, one should perform the expectation of both sides of Equation 2, which gives Equation 5*(a)*. The transformations from Equation 5*(a)* to Equation 5*(l)* indicate that

Equation 5: Bias – Variance Decomposition

(a) $\mathrm{E}(\mathrm{Error}_{\mathrm{training}}) = \mathrm{E}\left\{\sum_i \left(y_i - y'_i\right)^2\right\}$

(b) $\mathrm{E}(\mathrm{Error}_{\mathrm{training}}) = \left\{\sum_i \mathrm{E}\left(y_i - y'_i\right)^2\right\}$

(c) $\mathrm{E}(\mathrm{Error}_{\mathrm{training}}) = \left\{\sum_i \mathrm{E}\left(y_i - f_i + f_i - y'_i\right)^2\right\}$

(d) $\mathrm{E}\left(y_i - y'_i\right)^2 = E(y_i - f_i)^2 + E\left(f_i - y'_i\right)^2 + 2\mathrm{E}(y_i - f_i)\left(f_i - y'_i\right)$

(e) $\mathrm{E}\left(y_i - y'_i\right)^2 = E(\varepsilon^2) + E\left(f_i - y'_i\right)^2 + 2(\mathrm{E}\{y_i f_i\} - \mathrm{E}\{y_i y'_i\} - \mathrm{E}\{f_i f_i\} + \mathrm{E}\{f_i y'_i\}$

(f) $\mathrm{E}\left(y_i - y'_i\right)^2 = E(\varepsilon^2) + E\left(f_i - y'_i\right)^2$

(g) $\mathrm{E}\left(y_i - y'_i\right)^2 = E(\varepsilon^2) + E\left(f_i - \mathrm{E}\{y'_i\} + \mathrm{E}\{y'_i\} - y'_i\right)^2$

(h) $\mathrm{E}\left(y_i - y'_i\right)^2 = E(\varepsilon^2) + E\left\{\left(f_i - \mathrm{E}\{y'_i\}\right)^2\right\} + E\left\{\left(\mathrm{E}\{y'_i\} - y'_i\right)^2\right\} + 2E\{(f_i - \mathrm{E}\{y'_i\})(\mathrm{E}\{y'_i\} - y'_i)\}$

(i) $bias = E\left\{\left(f_i - \mathrm{E}\{y'_i\}\right)^2\right\}$

(j) $variance = E\left\{\left(\mathrm{E}\{y'_i\} - y'_i\right)^2\right\}$

(k) $\mathrm{E}\left(y_i - y'_i\right)^2 = E(\varepsilon^2) + bias + variance + 2\left(\mathrm{E}\{f_i \mathrm{E}(y'_i)\} - \mathrm{E}\left\{\mathrm{E}\{y'_i{}^2\}\right\} - \mathrm{E}\{f_i y'_i\} + \mathrm{E}\{y'_i \mathrm{E}\{y'_i\}\}\right)$

(l) $\mathrm{E}\left(y_i - y'_i\right)^2 = E(\varepsilon^2) + bias + variance$

For the transition from Equation 5*(e)* to Equation 5*(f)*, we have:

$E\{y_i f_i\}=f_i^2$ as f_i is deterministic and $E\{y_i\}=f_i$.
$E\{f_i f_i\}=f_i^2$ as f_i is deterministic
$E\{y_i y'_i\}=E\{ y_i(f_i+\varepsilon)\}=E\{ y_i f_i \}+E\{ y_i \varepsilon\} = E\{ y_i f_i \}$ as is ε noise.
Therefore, $E\{y_i f_i\}- E\{y_i y'_i\}+ E\{f_i f_i\}- E\{ y_i f_i \}=0$.

Similarly, for the transition from Equation 5(k) to Equation 5*(l),* we have:

$E\{f_i E\{y_i\}\}=f_i E\{y_i\}$
$E\{E\{y'_i\}\}^2= E\{y'_i\}^2$
$E\{f_i y_i\}=f_i E\{y_i\}$
$E\{y'_i E\{y'_i\}\}= E\{y'_i\}^2$
Therefore, $E\{f_i E\{y_i\}\} - E\{E\{y'_i\}\}^2 - E\{f_i y_i\} + E\{y'_i E\{y'_i\}\}=0$

Equation 5*(i)* defines as *bias* the difference between outputs of the original function *f* and the expected outputs of the approximated function *g*. The expected outputs of *g* for new inputs should be similar to the outputs of *g* for inputs from the training data set. Therefore, the bias represents the discrepancy between the functions *f* and *g* and can be estimated using training data. A model that can fit all training data is a zero bias model.

Equation 5*(j)* defines as *variance* the difference between the expected outputs $E\{y'_i\}$ of function *g* and real outputs y'_i of function *g* for the new inputs. The variance is represented by the difference between the performances of the model for training data and for validation data.

According to Equation 5*(j),* the prediction error of a model depends on three components:

- The noise ε that can be available in the data
- The bias of the model
- The variance of the model

To reduce the prediction error, it is necessary to reduce the components. However, with a given data set and no further specification of function *f*, the noise cannot be reduced. Therefore, reducing the prediction error depends on reducing the bias and the variance.

However, there is a trade-off between bias and variance (Geurts, 2010) that prevents optimizing both bias and variance at the same time. A reduction of the bias leads to an increase in variance, and a reduction of variance leads to the increase in bias. An example of obtaining low variance and high bias is to approximate a complex original function *f* by a constant function *g*. *g* gives the same constant value whether the inputs come from the training data set or from new data. Therefore, the variance is obviously zero. However, the constant function *g* is very far from the complex function *f*, creating a high bias between two functions. An example of low bias and high variance is represented by fitting training data generated by a sine function *f* by a polynomial function *g*. One can increase the order of *g* until all training data are fitted by *g*, i.e. there is no bias between the two functions. However, when there are new inputs, *g's* predictions are rather poor as *g* is linear while *f* is non-linear (a sine function). The result is that the bias between *g* and *f* is low (zero), whereas the variance is high.

Although there is no approach that can simultaneously reduce both bias and variance to zero, it is possible to find approaches providing acceptable bias as well as variance.

3.4.5. Summary

The machine learning approach plays an important role in model development and contributes greatly to the developed model. Therefore, appropriate approaches have to be selected when the methodology is applied to a particular data set. For an approach to be selected, it should be able to overcome the following issues:

- Certain variables can be categorical.
- Imbalance ratios can be high. The selected approach needs to resist imbalance.
- Prediction power should be high, i.e. the performance of developed model is high for both NTS and PTS validation data sets.

3.5. Summary

This chapter presents the methodology and the choice of techniques used for developing risk identification models. The traffic is quantified into Traffic Situations (TS) characterized by variables that can be categorical or numerical, and there can be traffic factors or external factors. Thereafter, TS are clustered into Traffic Regimes (TR) where TR-based Risk Identification Models are developed. Figure 3-4 summarizes the methodology consisting of the following three main steps:

i) *Data Preparation*: Input data are integrated to define NTS and PTS.

ii) *Data sampling* : NTS are sampled in order to group PTS with relevant NTS into clusters called traffic regimes

iii) *Risk Identification Models (RIM) development*: RIM are developed under each traffic regime

It is important to note that there are design choices at each step of the methodology. For a concrete study site, the design choices need to be made such that the final results are optimized for the available data from the study site. Thereby, the proposed methodology can be executed by following one of two ways below:

1) Setup the initial design choices and apply the methodology with available data from study site. From the initial results obtained, design choices at each step are optimized aiming to maximize the performance of developed risk identification models.
2) Design choices are made at each step of the methodology.

In the current research, our approach is to fix the initial design choices to obtain the first results. Thereafter, the developed models will be optimized in order to improve the performance of risk identification models. This approach is appropriate as the number of design choices is large and one cannot test all the choices at every step.

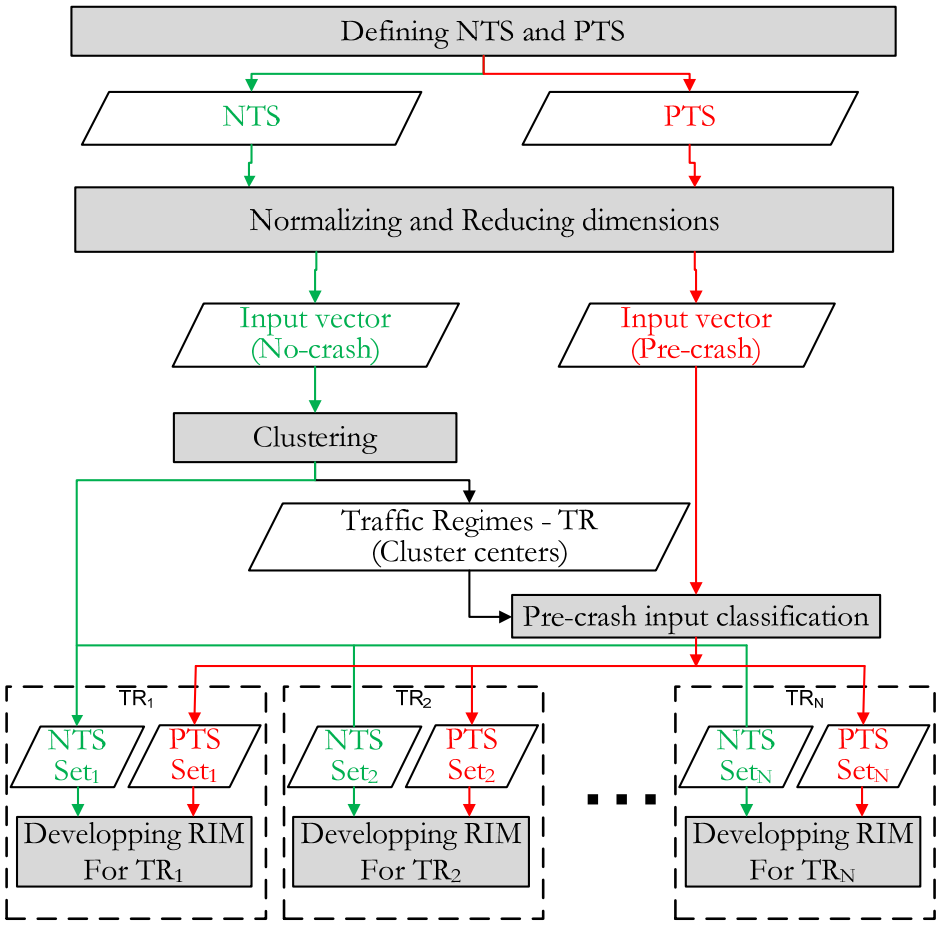

Figure 3-4: Summary of methodology

Chapter 4 Data and Study Sites

The methodology that is developed and presented in Chapter 3 will be applied under Swiss conditions. This chapter presents the research conditions in Switzerland. The discussion on data, data format, and study sites as well as the issues relating the data. Preliminary crash analyses are also presented.

4.1. Overview

4.1.1. Swiss Motorway Network

In Switzerland, motorways are called *autobahnen* in German, *autoroutes* in French or *autostrade* in Italian. Two of the most important motorways are the A1, connecting St. Margrethen in eastern canton of St. Gallen to Geneva in western part of the country, and the A2, connecting Basel in the northern part to Chiasso in canton of Ticino in the south. The general speed limit on Swiss motorways is 120 km/h.

The Federal Roads Office (FEDRO) is the Swiss authority responsible for the country's road infrastructure and private road transport. As of the January 1st, 2008, its range of duties increased significantly. With the entry into effect of the redistribution of financial responsibility and the accompanying division of duties between the federal government and the cantons, it assumed the functions of developer and operator of the motorway network. It belongs to the Federal Department of the Environment, Transport, Energy and Communications (DETEC), and focuses on securing sustainable and safe mobility on the country's roads.

According to FEDRO, (2009), a total of 1'765.6 kilometers of motorway are currently in operation. The network is planned to comprise 1'892.5 kilometers. The remaining 126.9 kilometers are expected to be completed within the next 15 years. Until the end of 2008 a total of 1'765.6 kilometers of motorways were in operation including:

- 7-lane sections 1.2 km
- 6-lane sections 80.7 km
- 4-lane sections 1'300.8 km
- 3-lane sections 1.9 km
- 2-lane sections 269.5 km
- Mixed sections 111.5 km

This corresponds to 93.3 percent of the planned network presented in Figure 4-1.

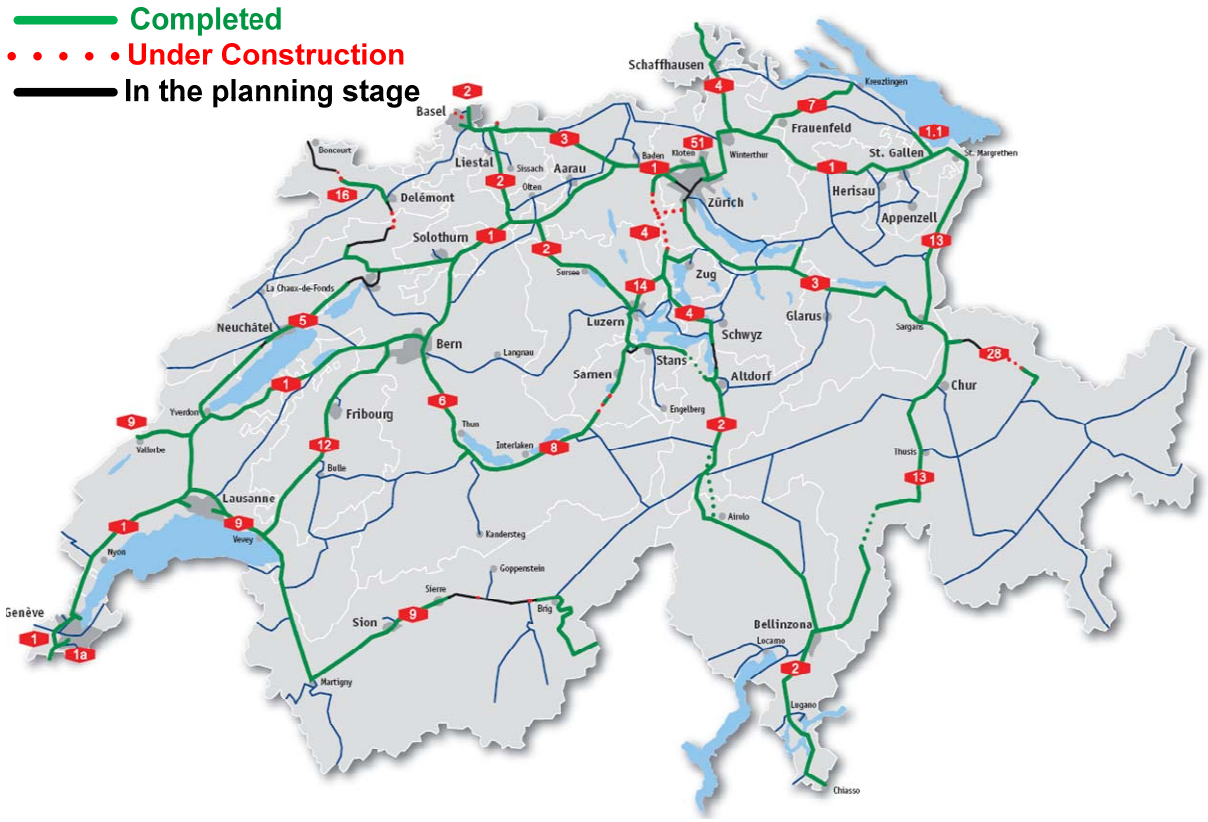

Figure 4-1: Swiss motorway networks. Source: FEDRO

4.1.2. Data Sensors

Until Jan, 2010, about 300 traffic detector stations are installed on Swiss motorways and provide online downloadable data. The map of the stations is presented in Figure 4-2. According to FEDRO, (2009), there are criteria to choose the location on the road for installing traffic detectors. The criteria serve for two main purposes: traffic statistics and better traffic management. The traffic data collected are used by engineering companies or traffic operators.

Most of traffic detectors on Swiss motorways are double inductive loop detectors. Each loop detector records the time of passage and the duration of presence of each vehicle on the detector. Two loop detectors installed at a distance of 4 meters make up a double loop detector on one road lane collecting individual vehicle information when vehicles pass by the detectors.

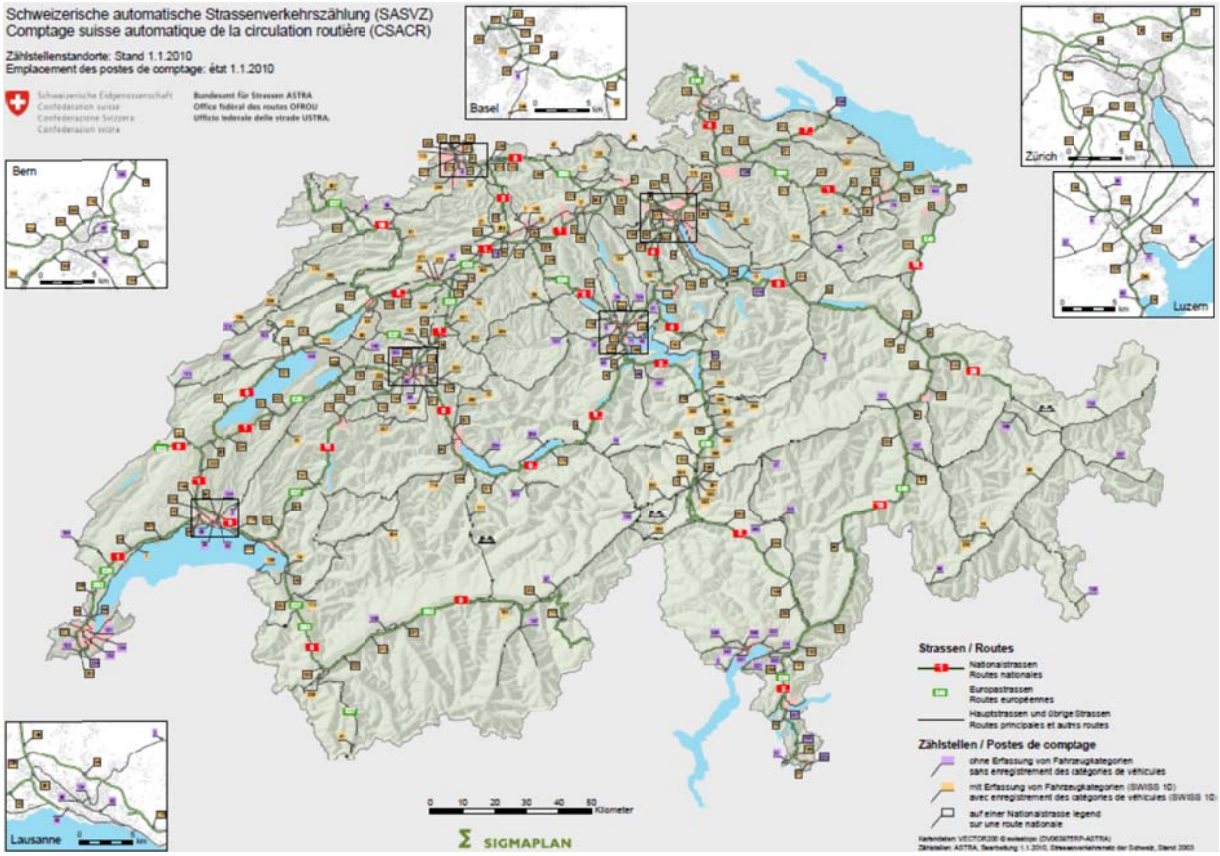

Figure 4-2: Swiss automatic road traffic counts

Together with traffic data detectors, there are also meteorological sensors for measuring weather information. There are two meteorological station systems exist in parallel in Switzerland: MeteoSwiss system and Boschung system.

The Boschung stations are installed along Swiss roads, including motorways for collecting road weather information aiming to improve road maintenance and safety. According to Boschung, (2010), the information collected depends on particular stations. Unfortunately, due to technical problems, only data from Boschung stations on motorways in Vaud canton are accessible and used in this study.

The other meteorological system in Switzerland comprises MeteoSwiss stations providing weather information for general purposes including daily weather forecast. MeteoSwiss stations are installed to collect information for the whole area around the station.

4.1.3. Guidelines for Study Site Selection

This study is classified as a disaggregate traffic safety study using traffic flow data, i.e. pre-crash traffic conditions are compared with non-crash traffic conditions to find the relationship between traffic flow data with crash occurrences. However, Swiss motorway network is large with various characteristics from

sections to sections. Therefore, a road section to be selected as a study site needs to satisfy the following conditions:

i. There must be at least one traffic detector station installed in that section. Besides, the data collected from the station should be accessible. This condition limits the search space of study sites into the list of more than 300 detector stations presented in Figure 4-2.
ii. From the position of the selected traffic detector stations, there should be as many crashes recorded as possible. Together with the first condition, this criterion is decisive for the current study. The numerousness of crashes results in the high number of pre-crash cases usable in the present study.
iii. There should be meteorological stations close to the traffic detector station.

There are also other rules applicable to select a study site. However, the guidelines would limit dramatically the search space of potential study sites.

4.2. Data Specification

4.2.1. Introduction

Three types of data used in the present study include traffic data, meteorological data, and crash data. The traffic data is provided by FEDRO. The meteorological data is provided by Swiss Federal Office of Meteorology and Climatology (MeteoSwiss). Alternative meteorological data is provided by Boschung – a private company who manages road weather stations. Crash data is provided by Swiss Federal Statistics Office (FSO).

4.2.2. Traffic Data

Traffic data in Switzerland is collected from double inductive loop traffic detectors installed under the road pavement and is stored under individual vehicle format, i.e. the information of individual vehicles is recorded when the vehicles pass by the detectors. The traffic data format is presented in Figure 4-3. It is worth noting that the headway of a vehicle is the time distance between front bumper of that vehicle to the front bumper of the vehicle preceding the current vehicle whereas; the time gap of a vehicle is the time distance between front bumper of that vehicle to the rear bumper of the preceding vehicle. Therefore, time gap of a vehicle is always smaller than headway.

In Figure 4-3, the classes of vehicles are extracted according to the instruction presented in (FEDRO, 2009). The lane of passage is the lane index ranging from 1 to the total number of lanes for both directions. Depending on the location, there can be more lanes on one direction than on the other direction. For each detector station, FEDRO provides a guideline about the direction for each lane index. In most of 4-lane sections, the traffic on lanes with indices of 1 and 2, or 3 and 4 is on the same direction. The traffic direction in Figure 4-3 is always 1.

Motorway traffic detectors are installed aiming to monitor the traffic and to generate statistics. Network coverage is analyzed to cover the best these two objectives. The amount of traffic, the number of metering station on the section, the risk of congestion are all factors taken into consideration. As such, as illustrated in Figure 4-2, traffic detectors are installed with higher density around big cities such as Geneva, Lausanne, Bern, Basel, and Zurich. Detectors are also installed near intersection between two motorways.

Index	Date	HHMM	Sec	ms	Reserved	Lane	Dir	Headway	Time Gap	Speed	Length	Veh. Class
023198	150303	0001	21	30	000000	1	1	43.6	43.5	120	467	2
023199	150303	0001	38	42	000000	1	1	17.1	16.9	125	989	3
023200	150303	0001	47	12	000000	4	1	46.9	46.8	113	428	2
023201	150303	0001	50	58	000000	4	1	3.4	3.3	119	423	2

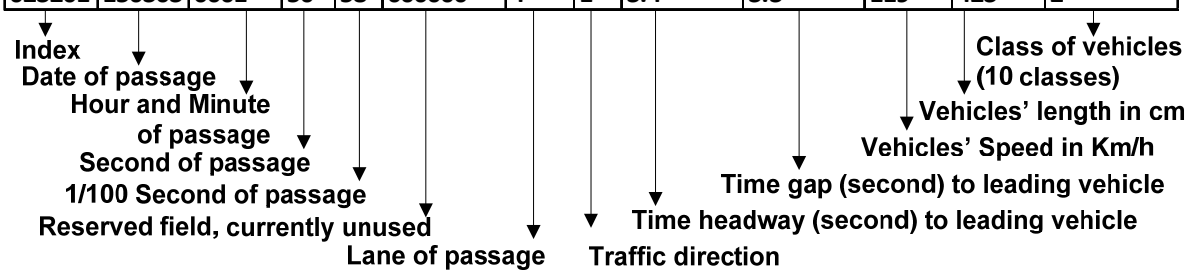

Figure 4-3: Traffic data format

4.2.3. Meteorological Data

4.2.3.1. Boschung Stations

Depending on locations, Boschung stations provide meteorological data with the following data fields:

i. Station code
ii. Station name
iii. Date time
iv. Air Temperature
v. Soil Temperature
vi. Relative of Humidity
vii. Dew Point
viii. Type of precipitation (Three types: rain or snow or no precipitation)
ix. Quantity of precipitation (Five levels: nothing, weak, normal, strong, very strong)
x. Number of spraying (this is based on the deicing program with salt solution)

Each data line represents a query for 5-minute intervals. While data fields iv, v, vi, vii, and x contain quantitative values, data fields viii and ix are categorical.

4.2.3.2. MeteoSwiss Stations

MeteoSwiss data from all MeteoSwiss stations are managed and stored on a central server. Data extraction is undertaken via a Java-based software Clipmap. The software can provide meteorological data at many aggregation levels. The most detailed data represent information for 10-minute intervals. Some example meteorological parameters are instant temperature at two meters above ground (*tre200s0*), instant relative humidity at two meter above the ground (*ure200s0*), the total precipitation (in mm) of the last 10 minutes (*rre150z0*).

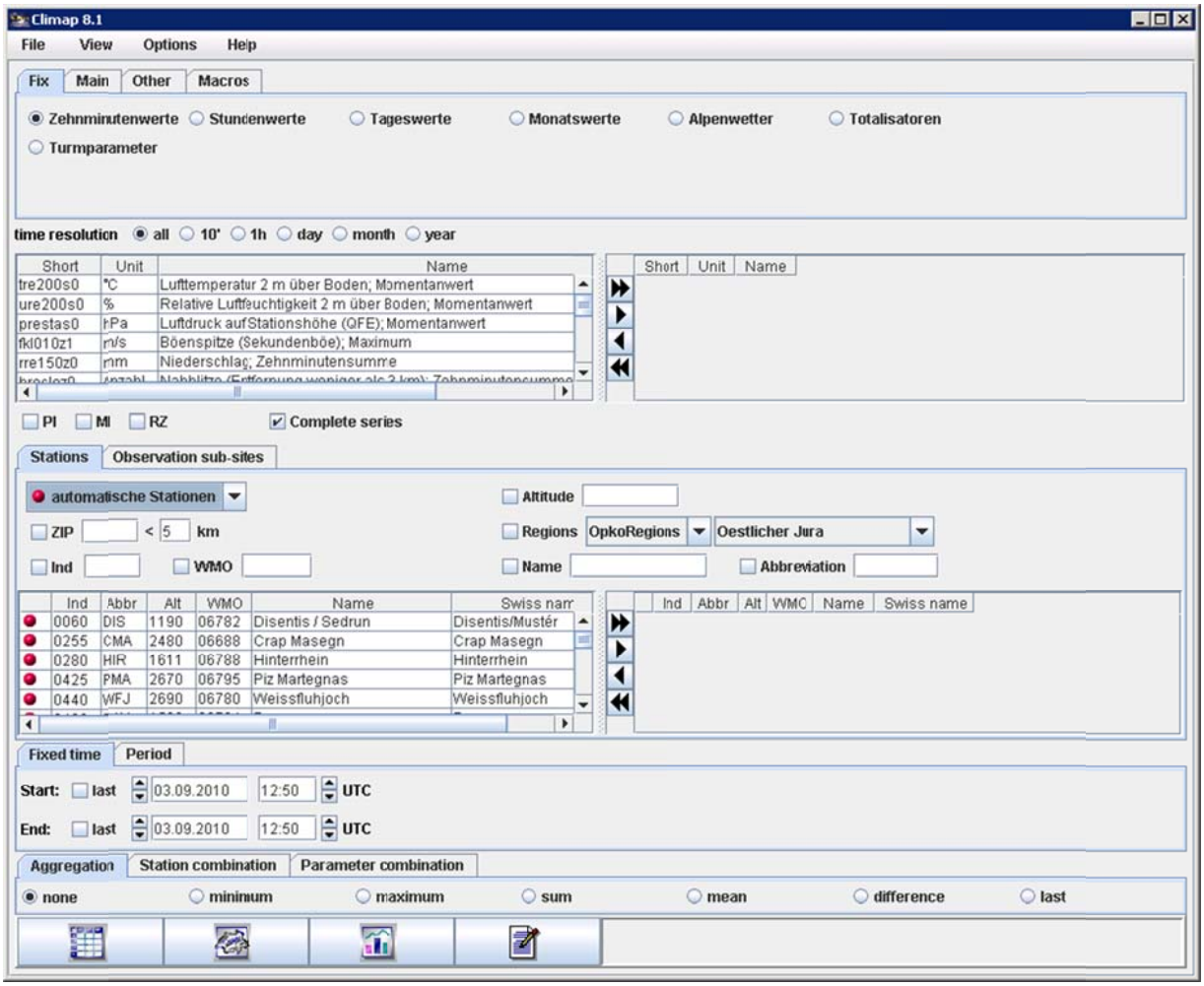

Figure 4-4: Clipmap interface (in English and German)

4.2.3.3. Summary

MeteoSuisse and Boschung stations are alternatives that provide meteorological data. Whereas Boschung data is specific to road weather conditions, MeteoSuisse data is more general for localities. The common characteristics of these two types of stations include the lack of sight distance, the presence of fog or the information on ambient conditions.

However, data from these stations can still be useful for the current study with information such as precipitation, temperature, etc.

4.2.4. Crash Database

In Switzerland, all traffic crashes recorded by the police are informed to FSO who stores the data for statistics purposes. According to FSO, (2010), the concept of crash depends on the mode of transport. Road traffic crashes since 1992 are all crashes occurring on public roads. Yet before that until 1991, only

crashes causing damage estimated at more than 500 Swiss francs (until 1975 the limit was set at 200 francs) were taken into account. Since 2002, a crash is considered as road traffic crash if it occurs on public roads and causes injuries. Anyone who has suffered injuries, regardless of gravity, falls into the category of injured.

According to FSO, (2010), since 1992, the road traffic police collect anonymous individual data on traffic crashes, vehicles and people involved with features registered including:

Circumstances of the crash (date, time, type of crash, type of road, location, conditions, etc.)
Type of vehicles involved, information on drivers (goal and driving license) and the people involved (position in the vehicle security system, consequences of the crash, sex, age)

The submission of crash data to the FSO needs to be transmitted using informatics or statistical survey forms which is listed in (FSO, 2005) or presented in Figure B-1 and Figure B-2. The statistical survey forms are considered as the statement of crashes or a guideline for the census of the crashes. The FSO also provides the guideline on how to fill the survey form (FSO, 2005).

The survey form can be used for any road traffic crash type and for any road type. Ten crash categories are used to distinguish different crash types as listed in Table 4-1. More details about crash categories are illustrated in Figure B-3.

Motorways are safe by design such that the probability for some accident types such as accidents of categories C, G, and H to occur is minimized whereas; some other crash types such as crashes of categories B, D, E, and F are more common. Figure 4-5 presents the percentages of four most common crash types plus the percentage of all other crashes for six years from 2002 to 2007 with a total of 46'641 crashes on Swiss motorways.

Table 4-1: Ten crash types for all road types

Category	Name
A	Crashes related to pedestrian
B	Skidding or losing control
C	Crossing crashes
D	Crashes while overtaking
E	Rear-end crashes
F	Crashes while changing lanes (for pre-selection) or bypassing
G	Crashes while changing directions
H	Crashes while turning (without changing the direction)
I	Crashes caused by an animal
K	Other crashes

Note that crashes under category B are considered as single vehicle crashes. According to FSO, (2005), a crash is classified into category B when the driver tries to avoid a collision or deviates from the carriageway by his own fault. There should have been no prior collision with another road user; otherwise it is another type of accident. The crash due to evasive action, for example triggered by an overtaking maneuver of a vehicle on other direction, is classified into this category. When a driver overtake or being overtaken loses control of his vehicle, the correct type is however an accident when overtaking.

The category D of crashes includes only crashes related to overtaking. The collision occurs between the overtaking vehicle and a vehicle overtaken, a vehicle traveling in the opposite direction or a vehicle traveling behind the first vehicle and already conducting an overtaking.

Rear-end crashes under category E indicates crashes by a vehicle hitting another vehicle, whether moving or momentarily stopped, which borrows the same way. If the first vehicle hits a parked or permanently stopped vehicle, it is an accident by slipping or loss of control (of category B), or other type of crash.

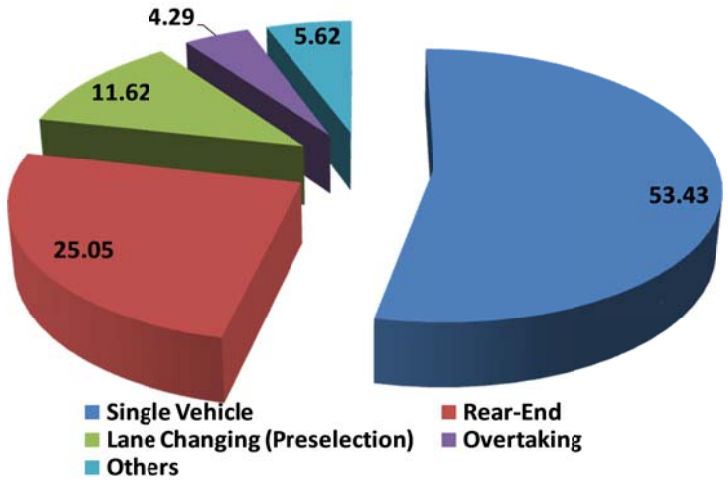

Figure 4-5: Percentages of most common crash types

Crashes under category F can be divided into two cases: bypassing and pre-selecting. The bypassing collision between two vehicles traveling in the same direction applies when none of the two vehicles intend to change lanes. A pre-selecting collision occurs when a vehicle collides before, during, or after a lane change, with another vehicle traveling in the same direction on a different lane. The lane change is planned or executed in order to pre-select before a junction, an on/off-ramp of motorways, but not to overtake or attempt to overtake.

The details of crashes are presented in the statistics survey form (FSO, 2005). There are three parts to be filled in for each crash: general about the crash; the object causing the crash; and the persons involved in the crash. The general information about a crash includes time, location, type, and external conditions (such as weather, pavement status, lighting conditions, etc.). Most of the general information is important for the present study. In the second part of a crash record, the information about the cause of the crash is important as it allows identifying whether the crash was caused by the driver's status, the vehicle's technical problems, or external factors. The third part of the crash record is not of interest of the current study.

4.3. Crash Related Issues

4.3.1. Overview

Among a total of 46'641 crashes on Swiss motorways during six years from 2002 to 2007, there are a number of crashes used and matched with traffic data at study site. To obtain better results in later stages of the current research, every individual crash recorded at the selected study site should guarantee that:

The impact of the crashes on traffic such as abnormal changes in speed or flow is observable in traffic after crash occurrences. This is essential as crash time and location are not known with precision and will be estimated based on traffic data
The crash is traffic-induced, i.e. not caused by unusual human state or technical error, etc. The specification of considered crashes is discussed in detailed in section 4.3.2.

Following sub-sections discuss about these issues.

4.3.2. Traffic-Induced Crashes

According to FSO, (2005), faults and influences of crashes should be obtained as much as possible at the location of crashes. By observation and notice, the person who establishes the crash report can mention up to three possible faults and influences to allow identifying all the factors having played a role in the development of the crash. There are four main faults and influences for road traffic crashes:

Human influence such as the state of drivers, visibility, driving skill, attention, etc.
Road and environment such as visibility, signalization, meteorological state, animals, etc.
State of vehicles
Traffic and the violation of traffic rules.

In the current research, not all crashes are considered. Each crash is verified whether it will be used. For example, crashes caused by bad state of drivers (such as drunk, influenced by drugs, unqualified for driving (i.e. no driving license), etc.) are not considered. Crashes with the appearance of animals on the road are excluded. Crashes due to technical problems of vehicles are also not included. Almost all crashes due to traffic and the violation of traffic rules are included. There are nine sub-categories of causes due to traffic and violation of traffic rules. The sub-categories are:

- About speed such as speed inappropriate for traffic conditions or speed over the speed limit, etc.
- About left/right movement and pre-selection such as not looking when changing lanes, quitting a platoon while travelling closely behind another vehicle, etc.
- About overtaking situation such as before or on a curve, etc.
- About overtaking decision during the traffic such as when vehicles are too close, premature return on the right lane, overtaking on the right lane, etc.
- About priority such as entering the motorway.
- About other movements in traffic such as following front vehicle too closely.
- About the movement of bicycles and motorcycles (this is not applicable for motorways).
- About pedestrians (this is not applicable for motorways).

In each sub-category, there are still certain traffic violations that are not applicable for motorways such as imprudent reverse gear, not stop in front of pedestrian crossing line, etc. Crashes caused by such violation are not present in the motorway traffic crash database. There are also crashes caused by drivers when the drivers' states were good such as overtaking on the right lane, exceeding speed limit. It is those drivers who are personally responsible for the crashes and those crashes are unconsidered.

Finally, crashes considered in the present study belong exclusively to categories D, E, and F in Table 4-1, i.e. crashes while overtaking; rear-end crashes; and crashes while changing lanes for pre-selection or bypassing, respectively. Although crashes while overtaking and crashes while changing lanes for pre-selection or bypassing are distinguishable by the aim of the movement, the development at the beginning of the crashes is more interesting for the current study. Therefore, these two crash types are called *sideswipe crashes* in this dissertation. According to Figure 4-5, the rear-end and sideswipe crashes contribute more than 41% out of all crashes on Swiss motorways during six years from 2002 to 2007.

4.3.3. Crash Observation on Traffic Data

Here, one of the main objectives is to link traffic flow conditions to crash occurrences. Therefore, it is critical that crash locations should by close to traffic detectors' location in order that traffic evolution preceding crashes can be observed in traffic data via speed or flow changes. For this reason, crashes to be considered are limited within a buffer of one kilometer from traffic detectors (for each direction).

4.3.4. Types of Crash

When a crash occurs, the police come to the site and follow a guideline (FSO, (2005)) to determine the type of the crash. The guideline includes a series of questions whose answers are *yes* or *no* such that the crash type is determined when the answer *yes* is given. The questions need to be asked according to a given order. The two questions in the list are used in the example below:

Question 1: Does the crash occur between a pedestrian and a vehicle such that the driver of the vehicle did not lose control, did not tend to avoid and did not divert his trajectory?

If the answer is *yes*, that is a pedestrian related crash. There are several exceptions:

If there are only pedestrian involved in the crash, the crash is not traffic crash.
If that is a crash with a vehicle such that the driver lost control or tended to avoid or diverted his trajectory, that crash is classified into crash category due to slippery or loss of control.
If an animal hold by the pedestrian is wounded and the pedestrian is fine, the crash is animal-related.

Question 2: is there a frontal collision between two vehicles traveling in inversed directions?

If the answer is *yes*, the crash is crossing crash.

As the road is motorways, popular types of crashes include rear-end, overtaking, lane changing for pre-selection crashes (see Figure 4-5) which are of interest of the current research and single-vehicle crashes which are the most numerous. By asking and answering the series of questions, the type of crashes on motorways can be precisely determined.

4.3.5. Crash Time and Location

In Switzerland, when a crash occurs on the motorway, the police of the canton where the crash occurs are responsible for the crash report. In the survey form (FSO, 2005), there are fields to precise the crash location such as the name of the road, the direction, the kilometrical position, the geographical position, etc. However, the crash reports returned to the FSO are not completely filled in. For the study site CH023 for example, the police from Bern canton did not precise the crash locations in term of lane where the crash occurred although the GPS coordinates of the crash are available. The deviation in geographical coordinates can be as much as 3 meters and therefore, it is impossible to identify the lane where the crash occurred. In many crash cases, the traffic direction where the crashes occurred is not identifiable. It is worth noting that the geographical position of crashes is still important and accurate enough for limiting crashes in the buffer of one kilometer from traffic detectors.

As a similar issue, the crash time is not known with precision. According to FSO, (2005), if it is not possible to determine the precise time when the crash starts to occur, the person responsible for filling the survey form needs to indicate the most plausible time based on his observation.

4.4. Study Site

According to guidelines in section 4.1.3, a study site is selected and used in this dissertation. The procedure of study site selection includes the following steps:

i. From the list of more than 300 traffic detector locations providing traffic data as specified in section 4.2.2, count for each detector location the number of crashes satisfying conditions discussed in section 4.3.2 within the buffer of 1km from the location, called *considered crashes*. Re-arrange the list in descending order of the number of considered crashes.
ii. Remove from the ordered list all traffic detector locations that there is no meteorological station within a buffer of 15km from the traffic detector locations. The remaining detector locations make up the reduced list of locations.
iii. Select the first location from the reduced list of locations.

As road crashes are monitored by cantonal police (there are 26 cantons in Switzerland), crash positions can be undetailed in the crash records for several cantons such as Vaud or Ticino, etc. Therefore, study site cannot be selected in those cantons.

The selected study site is at 27^{th} position of the considered list. The study site is named according to the name of traffic detector stations: CH023. The location of site CH023 is illustrated in Figure 4-6. Site CH023 lies on motorway A1 connecting two cities Bern and Zurich. The site CH023 is not the location where crash occurrences are the most numerous yet is selected because traffic, meteorological and crash data are altogether available, which is crucial in development of risk identification models by applying the methodology proposed in Chapter 3.

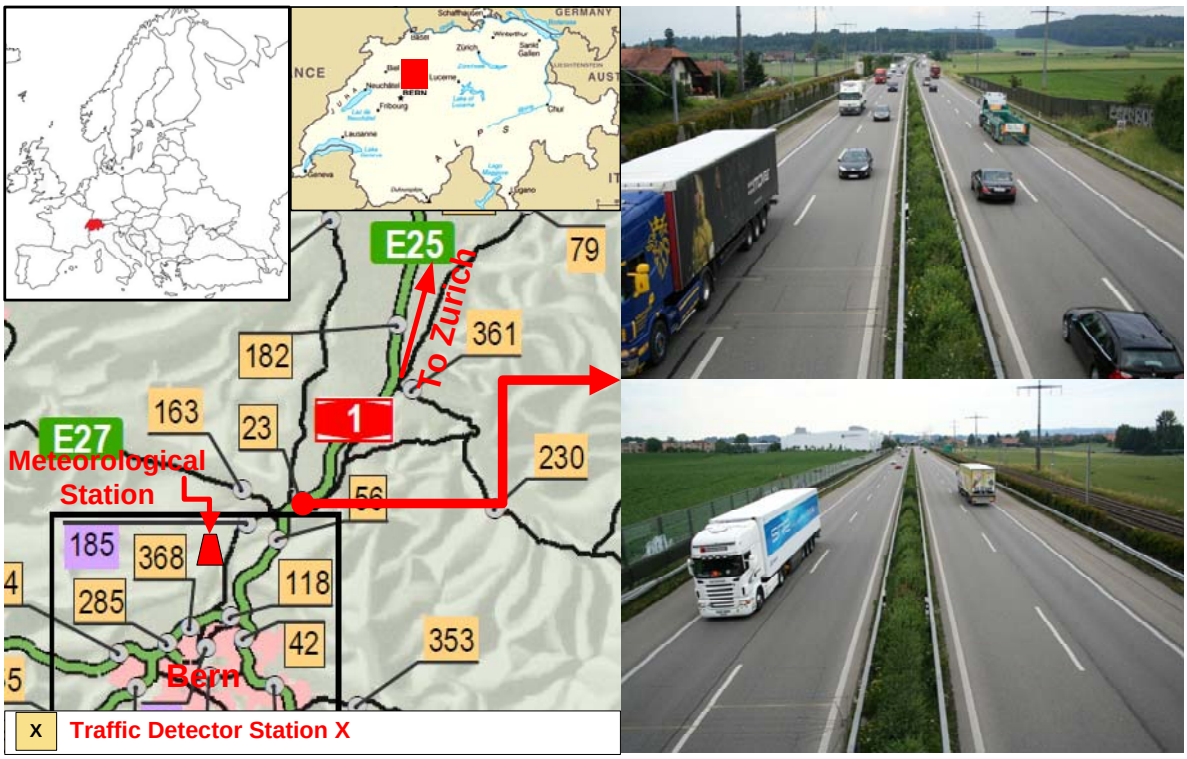

Figure 4-6: Study site

As the accessibility of Boschung data is only limited in Vaud canton, MeteoSwiss data are used to indicate meteorological state at the study site. There is a MeteoSuisse station locates at about 10km from the study site.

The road section at site CH023 includes two traffic dictions: Bern - Zurich and Zurich – Bern. On each of directions, the road pavement is divided into two lanes. The lane on the left is called *normal lane* whereas; the lane on the right is called *overtaking lane*. Within 1km from the location of traffic detectors, road sections are straight and the inclination of the section is almost zero.

4.5. Preliminary Crash Analysis

This section provides different statistics on crashes based solely on crash data. The statistics in this section can be referenced in the next chapters.

4.5.1. Crash Distribution by Time of the Day and Day of the Week

Figure 4-7 presents the crash distribution. Here the crash time in the crash records is used. In Figure 4-7, the last row represents the distribution of crashes by time of day averaged by all the days of the week. The

last column represents the crash distribution by day averaged by all time of the day. The number of crashes is the highest on Friday and then on Saturday. Crashes occur more often in late afternoon on Friday (from 16:00 to 19:00) than in any other time. Although the crash frequency is high on Saturday, crash occurrences did not concentrate on a particular hour of the day. Crashes occur more often in early Sunday mornings than in early morning of any other day of the week.

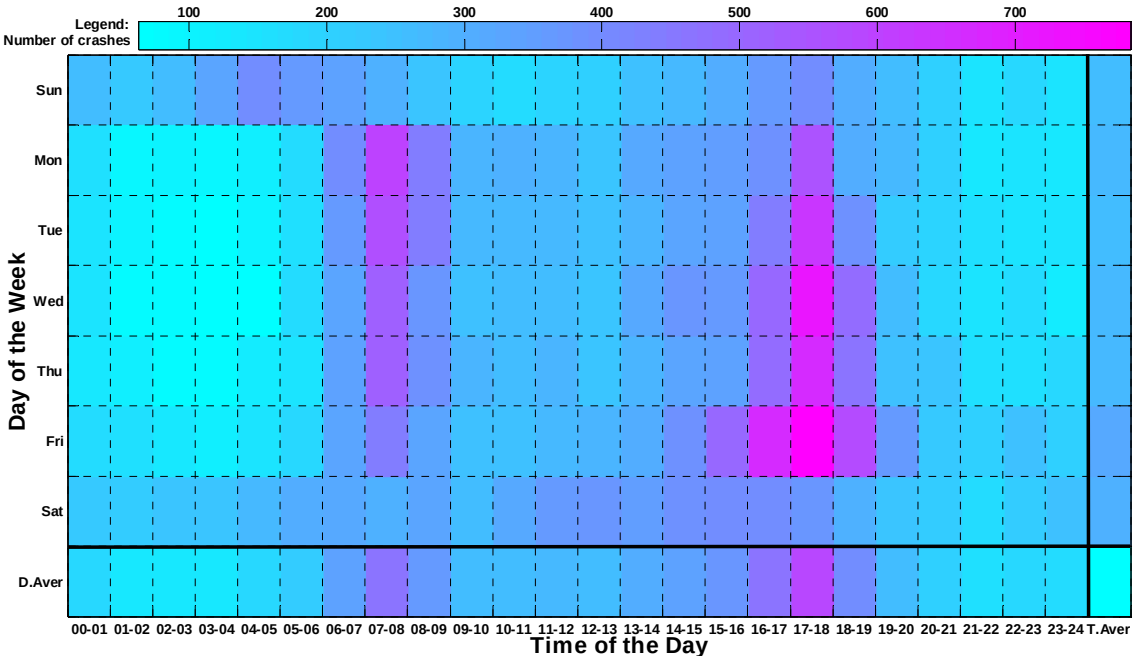

Figure 4-7: Crash distribution by day of the week and time of the day

The statistics in Figure 4-7 indicate that time of the day and day of the week contribute to crash occurrences.

4.5.2. Crash Distributions by Weather Conditions and Pavement Conditions

The crash distribution by weather conditions is presented in Figure 4-8. The crash distribution by pavement conditions is presented in Figure 4-9. The two distributions show that even when there is no precipitation in the air, the pavement can still be non-dry, i.e. there can be other sources for the non-dry state of the pavement.

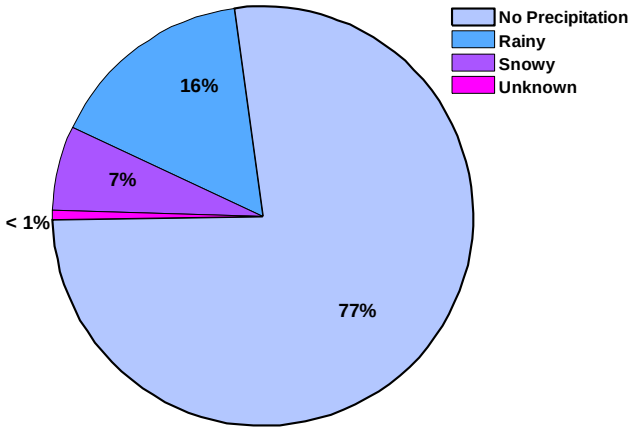

Figure 4-8: Crash distribution by weather conditions

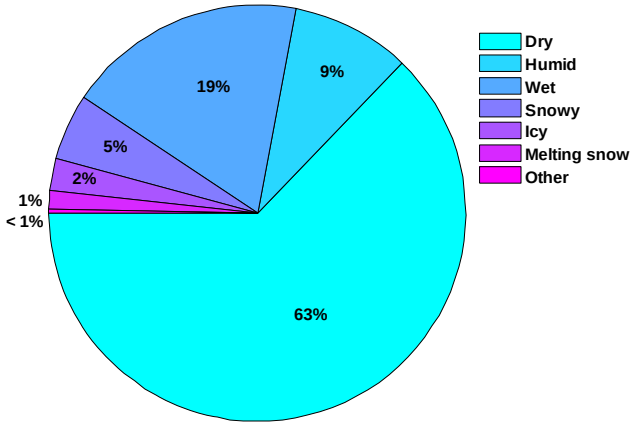

Figure 4-9: Crash distribution by pavement conditions

4.5.3. Crash Distributions by Crash Severity

There are many criteria to evaluate the severity of crashes. Figure 4-10, Figure 4-11, and Figure 4-12 three different views on the crash severity.

In Figure 4-10, the number of objects involved in a crash can be as high as 17. As the crashes are on motorways, the objects are vehicles. Due to the high speed practiced on the motorways, it is not abnormal to observe many vehicles involved in a crash.

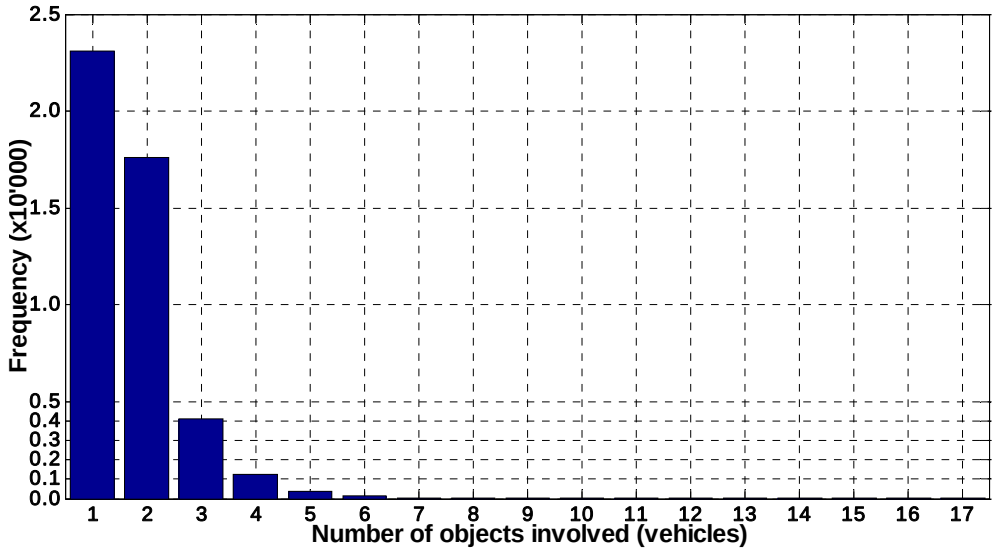

Figure 4-10: Crash distribution by the number of objects involved

In Figure 4-11, the crash severity is viewed according to the number of persons involved in the crash. Depending on the capacity of vehicles, there can be more than one person occupying a vehicle. The maximum number of persons involved in a crash is 86.

Figure 4-12 presents the crash distribution by the number of injuries. The number of crashes without any injury is high, contributing up to about 75% of all crashes. However, there are also crashes with the number of injuries increasing up to 27 or 57, which is high.

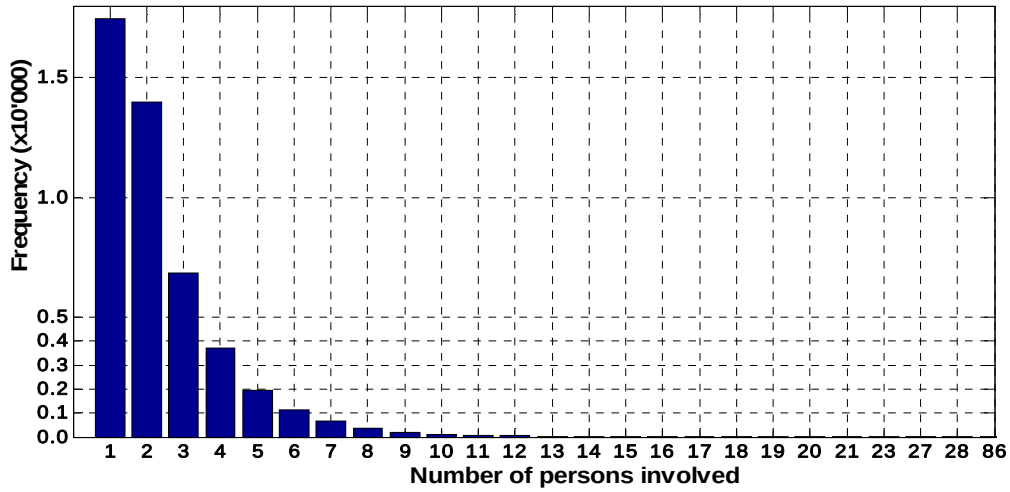

Figure 4-11: Crash distribution by the number of persons involved

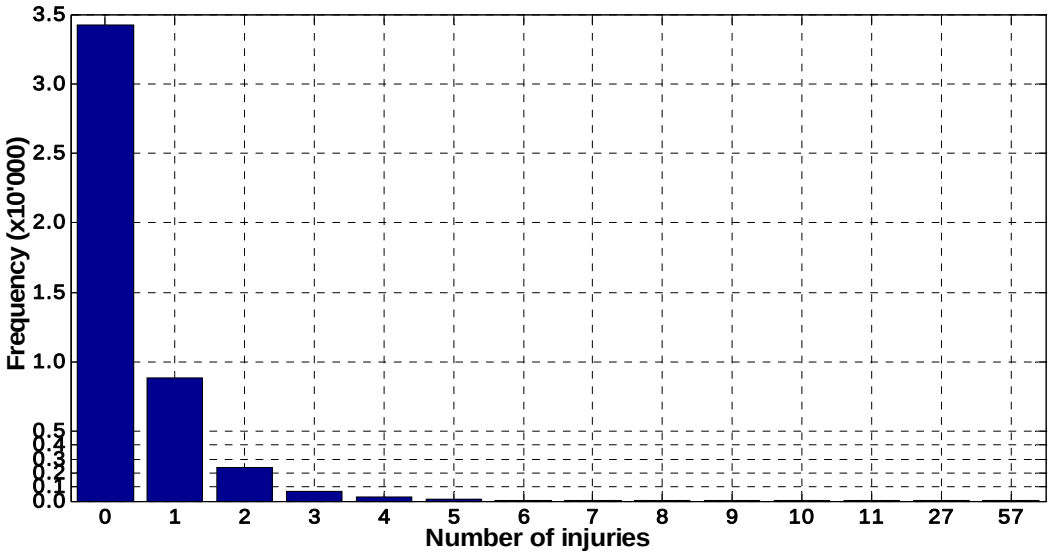

Figure 4-12: Crash distribution by the number of injuries

4.5.4. Crash Statistics at Study Site CH023

The statistics presented in Figure 4-13 and Figure 4-14 are based on all original crash data at study site CH023, i.e. the crash time was not corrected and crashes other than traffic-induced crashes are still counted. On a road section of 2 km (i.e. one km each direction from the location of traffic detectors) and during 7 years (i.e. from 2002 to 2007), there are 85'841'219 vehicles passing by traffic detectors. On the same road section and during the same period, 170 crashes are reported by the police. As result, the crash rate is for the road section during the period is 99.02 million of vehicle kilometers traveled.

Figure 4-13 suggests that most of crashes (about 67.6% of all crashes) at study site CH023 are rear-end whereas the proportion of single vehicle crashes is about 26.5%.

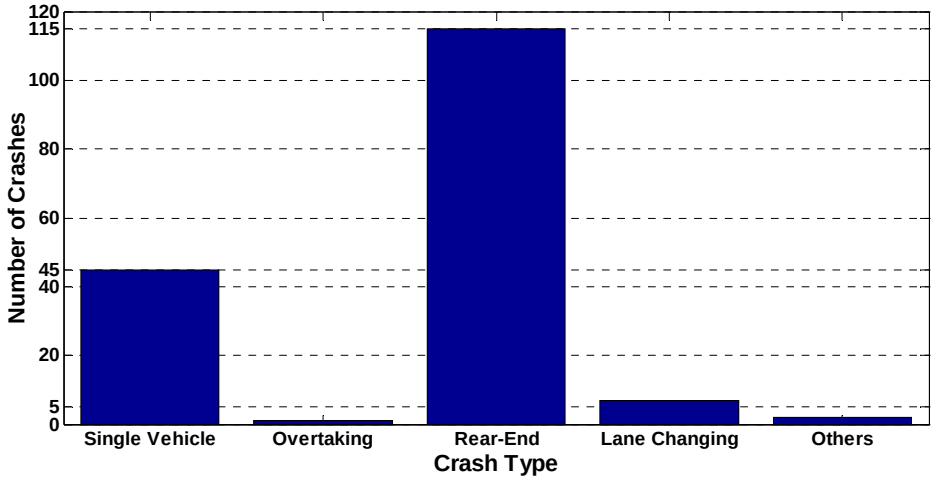

Figure 4-13: Crash distribution by crash types (Site CH023)

Figure 4-14 suggests that most of the crashes recorded at CH023 (about 51.2% of all crashes) occurred from 16:00 to 18:00 (i.e. period of two hours in the late afternoon). Besides, two crash peaks in Figure 4-14 also correspond to morning and evening peaks when the traffic flow is high.

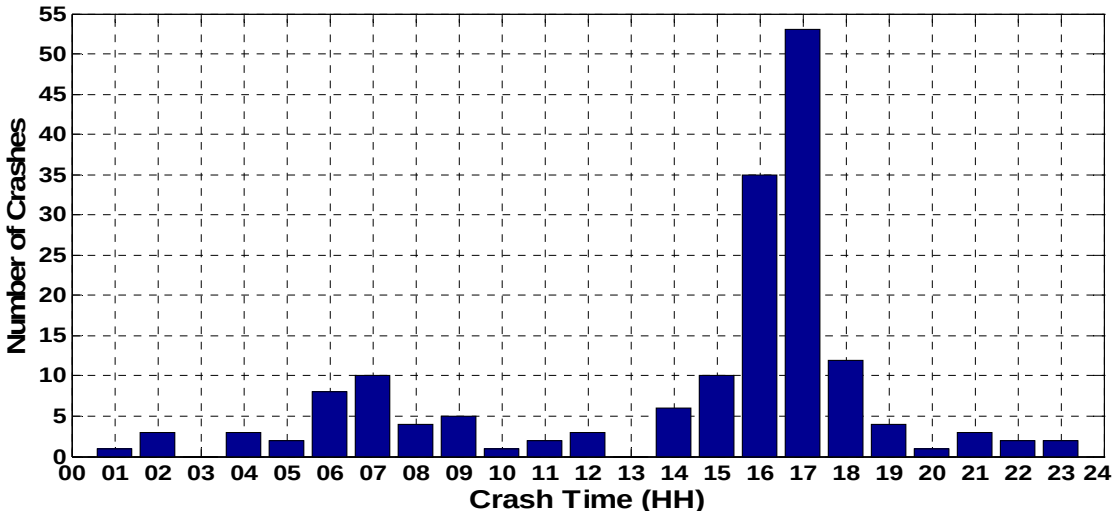

Figure 4-14: Crash distribution by crash time (Site CH023)

Speed/flow relationship of traffic conditions on the right lane before crashes is presented in Figure 4-15. Here, most of crashes occurred under high flow conditions. Several crashes occurring under low flow – high speed conditions are single-vehicle crashes. Crashes occurring under low flow – high speed conditions are rear-end crashes under congestions.

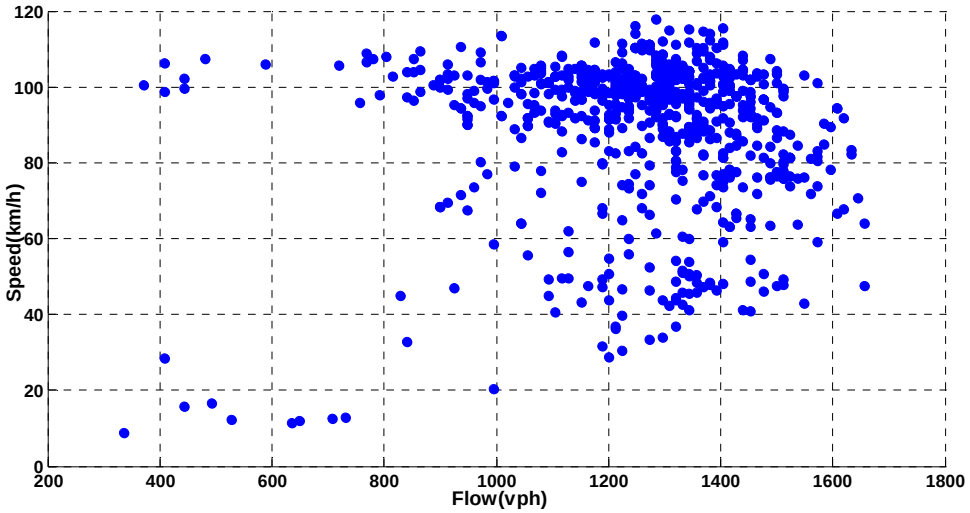

Figure 4-15: Speed/Flow diagram for traffic conditions preceding crashes on the right lane. Dots representing 5-minutes aggregated speed and flow

The speed profile based on the time of the day is presented in Figure 4-16. The clear speed drop can be observed from 11:00 on weekend and from 16:00 on weekdays which starts the afternoon peaks. This means that the congestions did occur although without high frequency. There is almost no congestion during the morning peak at the study site.

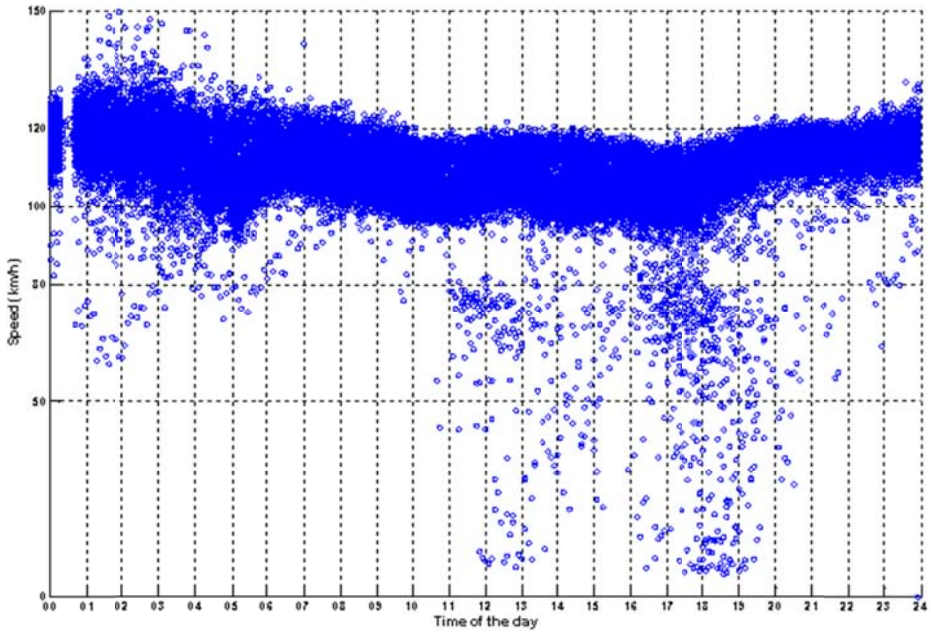

Figure 4-16: Speed profile according to time of the day (weekdays and weekend included)

4.6. Summary

Compared to existing studies in the literature (see section 2.3.2), except the study by Hourdakis et al., the available traffic data in Switzerland are more detailed with individual vehicle data. Due to the objectives of traffic detector installation, the spacing between detector stations in Switzerland is high. Therefore, variables of group VT4 (i.e. spatial variation of traffic, see section 2.4.3) are not considered in the current research.

Study site CH023 is selected for the current research according to several criteria. Road sections at the study site are straight with no inclination. At the study site, traffic, meteorological and historical crash data are altogether available.

There exist issues regarding to accuracy of crash records such as faults and influences causing crashes, the process to determine the types of crashes as well as crash time and crash location. There are guidelines to determine faults and influences causing crash and the types of crashes. Although the recorded faults, influences and types of crashes are not used to condemn drivers, they should be consistent with juridical crash records that are used by insurance companies and local authorities. Therefore, the exactitude of faults, influences and types of crashes are guaranteed. However, crash time and location are not precisely input in crash records. Crash time is not precise yet is acceptable as crash time in crash records is used as the reference to find the suitable crash time using traffic data. The crash time is

corrected to guarantee that pre-crash traffic data is collected before crash occurrences. This is important as the one of objectives of the current research is to identify the crash risk at some time before the crash occurs so that preventive measures are deployed and take effect on traffic.

Chapter 5 Traffic Situations

From this chapter, the application of the methodology presented in Chapter 3 to data selected from study site presented in Chapter 4 is presented according to steps presented in Figure 3-1. The present chapter addresses the first step: definition of traffic situations as well as two types of traffic situations that are appropriate for the selected study site.

5.1. Introduction

The application of proposed methodology to data at selected study site requires making methodological choices that agree with the available data. The represent chapter attempts to make the following decision regarding to the definition of traffic situations (see section 3.2) and the distinction between two types of traffic situations (i.e. NTS and PTS, see section 3.2.4):

- Choice of variables that are used for characterizing traffic situations (with the guideline presented in section 3.2).
- Determination of crash time for all crashes knowing the issues discussed in section 4.3.
- Determination of the durations of crash period, pre-crash period, post-crash period, pre-crash buffer period based on the available data at study site (see section 3.2.4).

5.2. TS Specification

5.2.1. Traffic Characteristics

According to section 2.4.3, there are four types of traffic related variables, namely VT1, VT2, VT3, and VT4. Depending to the equipment available at the study site, the number of variables used varies. For the study site selected in section 4.4, traffic data from one traffic detector station is used for characterizing traffic state. Therefore, traffic-related variables of type 4 – VT4 (see section 2.4.3) that require the availability of multiple detector stations are unavailable for consideration in the present research. The remaining variable types VT1, VT2, and VT3 include variables representing traffic characteristics within lanes, variables relating different lanes and variables linking consecutive time intervals, respectively.

As traffic data employed in the present research make use of individual traffic data, the number of potential variables characterizing the traffic state on each lane during an aggregation interval (i.e. variables belonging to VT1) can be high. Popular variables include the average speed, the flow, the occupancy, the percentage of heavy vehicles, the average time headway and time gap, the variation of speed, headway and time gap, etc. Several risk indicators such as Time-To-Collision – TTC and Mean Absolute Relative Speeds - MARS, can also be calculated using individual traffic data. To generate variables belonging to VT2, any variable of type 1 for one lane can be combined with that variable of other lanes. Similarly, to generate variables belonging to VT3, any variable of type 1 or 2 for one aggregation period can be combined with that variable for the previous or the next time aggregation intervals. Therefore, the choice of variables belonging to VT1 is decisive.

In our previous study (Mouzon et al., 2008), sensitivity analysis of risk indicators before crashes and during conditions where there is no crash was undertaken. Obtained results indicate that risk indicators

such as TTC and PBTR do not provide good indication of risky traffic conditions. It means that before few crashes, TTC (represented by inversed value) or PBTR values are relatively high. However, compared to overall TTC or PBTR distribution, those values are not high. Figure 5-1 extracted from (Mouzon et al., 2008), illustrates a typical example of the judgment. Other irrelevant TTS values was set to zero and is not shown in Figure 5-1 to make it easier to read. A crash occurred at about 21:00. TTC values before and after the crash occurrence remain similar making it not a good indicator for crash risk identification.

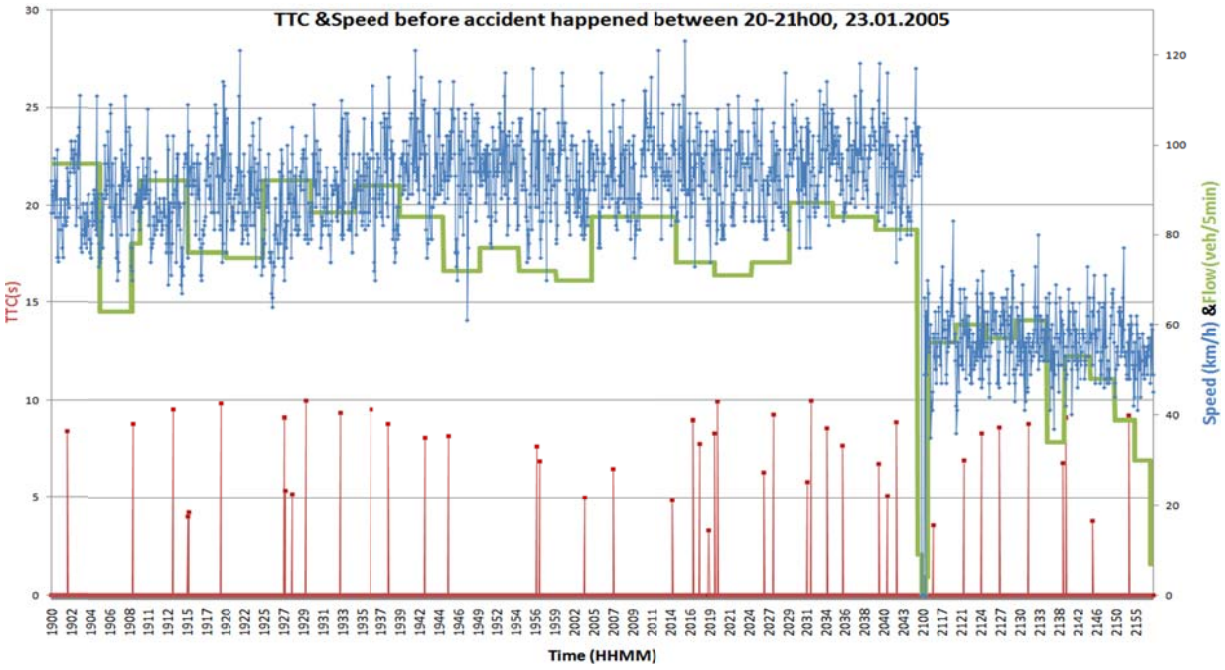

Figure 5-1: TTC distribution before a crash

Therefore, the utilization of risk indicators within the initial design choices is not elected. Instead, the decision of traffic variables used for characterizing traffic situations is listed below:

- Fundamental variables belonging to VT1 are used: flow, average speed, average headway, occupancy, percentage of heavy vehicles. These are variables aggregated directly from individual traffic data presented in Figure 4-3. Data fields in raw data such as time gap and vehicle length are correlated with considered variables (for instant, headway and time gap, vehicle length with class of vehicles – that produces percentage of heavy vehicles, etc.) and therefore, are not used.
- Two variables that are used in previous studies are headway and speed variations are also used in the current study.
- Variables characterizing traffic evolution between different time intervals – VT3 are not used in previous studies. Here, two variables for each lane representing temporal evolution of flow and average speed are used as initial choices.
- Variables characterizing traffic difference between adjacent lanes – VT2 are not considered in previous studies. Here one variable representing the difference of speed between two lanes is also employed as initial choices.

To summarize, most of traffic variables used are related to speed or flow – the fundamental traffic information. These variables are considered as initial choices and at this stage, there is no guarantee that the performance of the final risk identification models will be high. The foreseen benefice of using fundamental information related to speed and flow is that it will facilitate the interpretation of obtained results. Extended information such as the use of risk indicators can be considered to improve the performance of developed models.

5.2.2. Weather Information

As MeteoSuisse station is not next to traffic detectors, the utility of meteorological information is limited. Besides, information provided by MeteoSuisse station is local and might not be correlated with meteorological information at the location of traffic detectors. Therefore, the only usable meteorological data field is the precipitation. However, the intensity of precipitation at the location of MeteoSuisse station is different from the intensity at the location of traffic detectors due to the distance of about 10km. Therefore, precipitation is discretized into two values: 0 representing precipitation equal to zero and 1 representing positive precipitation.

Finally, only one variable representing meteorological information is used and can have one of two values 0 and 1.

5.2.3. Other Information

The moment of traffic situations can be used as variables. The moment of a vehicle passing by traffic detectors is recorded in traffic data (see section 4.2.2) and characterized by positions in two cycles: time of the day and day of the week. Time of the day can be any moment during the period of 24 hours and day of the week can be any day from Monday to Sunday. Therefore, these two temporal variables are categorical.

As no other data sources are located within study site, no other external variable can be considered in TS characteristics.

5.2.4. Aggregation Time Interval

The duration of 5 minutes is chosen as the duration of aggregation time interval. The motivation for this choice includes:

- Duration of 5-minute is chosen in previous studies. Using this duration makes it easier to compare new findings in the present research with previous studies. For instant, comparing methodologies for sampling non-crash data in Chapter 6 or comparing the performance of developed models in Chapter 8 are based on 5-minute aggregation.
- 5-minute aggregation interval is the initial choice to develop the initial models and can be optimized based on the performance of the models once the models are developed.

5.2.5. Summary

Three considered non-traffic variables are time of the day, day of the week, and discretized precipitation.

The number of traffic related variables depends on road configuration at the study site. As the selected study site is two-lanes-per-direction, 19 traffic related variables are used. Totally, 22 variables are used to characterize TS and introduced in Table 5-1.

Among 22 variables, *X1* and *X2* are categorical and represented as number. For example, *X2* values range from 1 to 7 representing from Monday to Sunday, respectively. *X1* also represents the index of the time period in a day. As TS is based on 5-minute aggregation time intervals, there are 288 different *X1* values ranging from 0 to 287 representing the intervals from 00:00-00:05 to 23:55-24:00, respectively for a day. If 1-minute interval is chosen, *X1* value range is from 1 to 1'440.

Variable *X22* contains precipitation information. *X22* values are either 1 or 0 representing whether there is precipitation during the aggregation interval or not. When the precipitation is positive, that amount is discretized to 1.

In case there are more than two lanes per direction at the study site, the following adjustments are foreseen in comparison with TS definition illustrated in Table 5-1:

- Addition of status for extra lanes. There are seven variables specifying the status of one lane.
- Addition of speed difference between two adjacent lanes. If there are L lanes, (L-1) speed differences are used in TS's definition.
- Addition of traffic evolution for each additional lane. Two variables (representing average speed and flow) are needed for each additional lane.

Therefore, the number of variables for three lanes is 32, for 4 lanes is 42, etc.

Table 5-1: Variables characterizing TS in case of two lanes

Variable	Alias	Meaning	Numeric/ Categorical	Unit	Value range	Specification
X1	*TDay*	Time of the day	Categorical	-	0 - 277	Instantaneity
X2	*WDay*	Day of the week	Categorical	-	1, 2, 3, 4, 5, 6, and 7	
X3	*LFlow*	Right lane's flow	Numeric	Vehicle per hour (vph)	0 - 4000	Right lane's status
X4	*LASpd*	Right lane's average speed	Numeric	Kilometers per hour (km/h)	0.0 - 200.0	
X5	*LAHw*	Right lane's average headway	Numeric	Second	0.0 - 100.0	
X6	*LOcc*	Right lane's occupancy	Numeric	Percentage (%)	0.0 - 100.0	
X7	*LVHw*	Right lane's standard deviation of headway	Numeric	Kilometers per hour (km/h)	0.0 - 100.0	
X8	*LVSpd*	Right lane's standard deviation of speed	Numeric	Second	0.0 - 50.0	
X9	*L% HV*	Right lane's percentage of heavy vehicles	Numeric	Percentage (%)	0.0 - 100.0	
X10	*HFlow*	Left lane's flow	Numeric	Vehicle per hour (vph)	0 - 4000	Left lane's status
X11	*HASpd*	Left lane's average speed	Numeric	Kilometers per hour (km/h)	0.0 - 200.0	
X12	*HAHw*	Left lane's average headway	Numeric	Second	0.0 - 100.0	
X13	*HOcc*	Left lane's occupancy	Numeric	Percentage (%)	0.0 - 100.0	
X14	*HVHw*	Left lane's standard deviation of headway	Numeric	Kilometers per hour (km/h)	0.0 - 100.0	
X15	*HVSpd*	Left lane's standard deviation of speed	Numeric	Second	0.0 - 50.0	
X16	*H%HV*	Left lane's percentage of heavy vehicles	Numeric	Percentage (%)	0.0 - 100.0	
X17	*Spd#*	Speed difference between two lanes	Numeric	Kilometers per hour (km/h)	-200.0 - 200.0	Discrepancy between lanes
X18	*LFCg*	Right lane's flow difference compared to previous TS	Numeric	Vehicle per hour (vph)	-4000 - 4000	Traffic evolution
X19	*LSCg*	Right lane's speed difference compared to previous TS	Numeric	Kilometers per hour (km/h)	-200.0 - 200.0	
X20	*HFCg*	Left lane's flow difference compared to previous TS	Numeric	Vehicle per hour (vph)	-4000 - 4000	
X21	*HSCg*	Left lane's speed difference compared to previous TS	Numeric	Kilometers per hour (km/h)	-200.0 - 200.0	
X22	*Prec*	Precipitation	Numeric	-	0 and 1	Weather conditions

5.3. Crash Time

5.3.1. Overview

As stated in section 4.3.5, crash time and location are not known with precision. While crash time is estimated by the police, crash location is under the form of GPS coordinates without indicating the traffic direction. One of the main objectives of the current research is to identify the crash risk before it turns into crashes. This objective can be achieved by examining historical crashes. The crash time of historical crashes needs to be corrected such that the corrected crash time is earlier than the real crash time.

Here, crash time and location are used as references to identify the suitable crash time and location using traffic flow data. As the traffic direction where the crash occurred is not known, traffic data for both directions are used to determine time and location of crashes. The precision about the lane where crashes occurred is not required as traffic situations are characterized by variables on both lanes.

5.3.2. Shockwave Propagation

The crash time is determined by shockwave theory. Let C denote the capacity after the incident. An incident causes a backward propagating shockwave only if the flow q (uninterrupted) is larger than C. If not, the incident is hard to observe from the upstream detector. The speed of the shock w can be determined from the fundamental diagram as illustrated in Figure 5-2. We thus get: $t_p = -d/w$, where t_p and d are the crash time and the distance from crash location to detector, respectively. Density can be estimated by: $k=q/v$ where q is flow or capacity and v is the average speed.

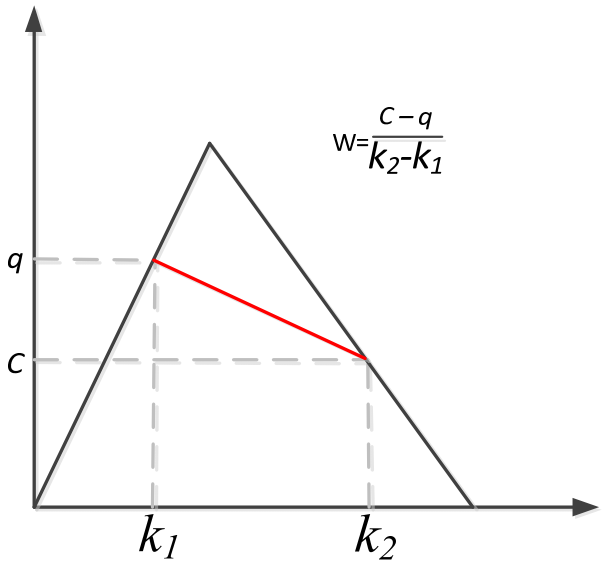

Figure 5-2: Calculation of the shockwave speed in case of incident, for q > *C*

The shockwave speed is substantially different than the speed observed at the detector. In case of a very severe incident, for which C would be almost zero, i.e. road section is closed, the speed of the shock would be large and the propagation time would be relatively small.

5.3.3. Crash Time Estimation

The calculation of the crash time is straightforward if q and C are known. However, q and C are unknown and need to be searched in the period before and after the recorded crash time. For each crash, traffic data on each direction for thirty minutes before and thirty minutes after its occurrence (according to recorded crash time) are used to detect crash time. Two flows with the largest absolute flow difference are considered as q and C. It is worth noting that rear-end and sideswipe crashes in the current study only occur under high flow conditions. Therefore, a crash would always provoke a capacity lower than the uninterrupted flow before the crash.

The following procedure is used to estimate the crash time for each traffic direction:

1) Order traffic data according to the time of passage of vehicles such that the last vehicle is the most recent vehicle.
2) For i=1 to T_{veh} - the total number of vehicles, calculate average speed v_i, density k_i, and flow of the direction q_i during the last 5 minutes from the time of passage of i-th vehicle (with i-th vehicle included).
3) *Max_q_diff*=0
4) *Time1*=1
5) *Time2*=1
6) For i=1 to T_{veh}
    For j=i+1 to T_{veh}
        If (crash is downstream of detector) and (q_i - q_j> *Max_q_diff*) then
            Max_q_diff = q_i - q_j
            Time1=*i*
            Time2=*j*
        End
    End
End
7) Among two directions, choose the direction having greater *Max_q_diff* value – that is the direction where the crash occurred.
8) *Max_q_diff* is the flow drop when the crash occurs. q and C are flows at time *Time1* and *Time2*, respectively. k_1 and k_2 are also densities at time *Time1* and *Time2*, respectively.

Then the crash time is $t_p = -\frac{d}{w} = -\frac{d(k_2-k_1)}{C-q}$ with d calculated using GPS coordinates of the crash and traffic detectors.

5.3.4. Shifted Crash Time

As the duration of traffic situations is 5 minutes, the crash time is shifted 2.5 minutes earlier and then is shifted again to the end of the last traffic situation. The shifted crash time is from 2.5 to 7.5 minutes earlier than the estimated crash time.

Figure 5-3 illustrates how the shifted crash time can be obtained. In the upper case, the shifted time is earlier than the estimated time for more than 2.5 minutes but less than 5 minutes. In the lower case, shifted time is 7.5 minutes earlier than estimated time.

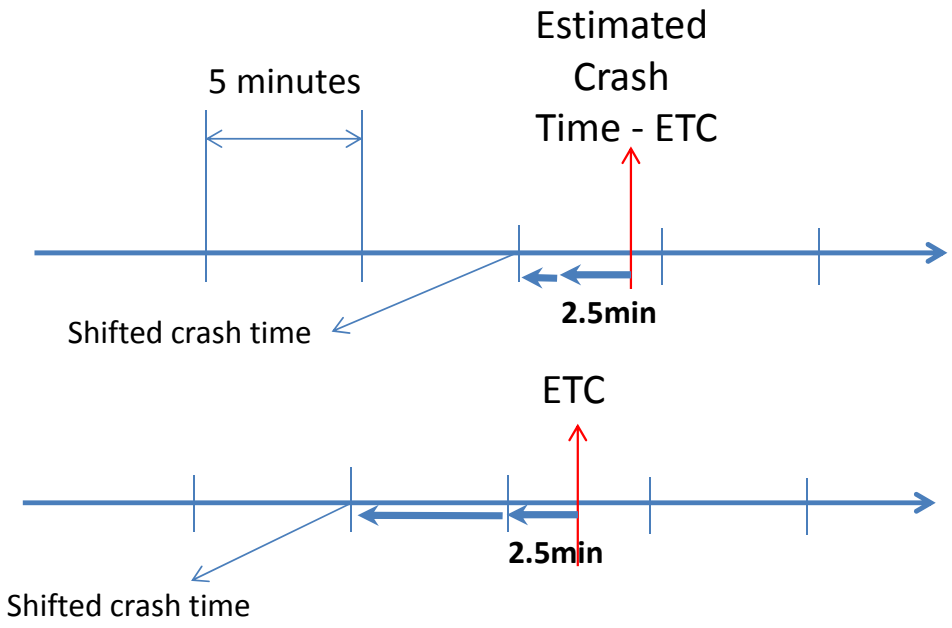

Figure 5-3: Illustration of shifting crash time

It is worth noting that shifting crash time does not violate the objectives of the current research as crash risks should be identified earlier before the crash occurrence so that actions can be taken to avoid the risk.

5.4. NTS & PTS

According to section 3.2.4, once the crash time is estimated, NTS and PTS can be defined.

Here, the duration of pre-crash period is decided to 30 minutes. A shorter pre-crash period is undesirable as crash risk might not be identified sufficiently early so that preventive measures can be activated and take effect. As TS are characterized by 5-minute intervals, 6 PTS are needed for each crash.

Crash risk might also appear at more than 30 minutes before crashes. Then the pre-crash buffer period can be a mixture of non-crash and pre-crash conditions. To reduce the ambiguity of this period, pre-crash buffer period is taken out of consideration.

Similarly, post-crash period is not of interest in the current research and is not considered. The duration of the post-crash period is achieved based on the impact of crashes on traffic. According to our analysis, the longest impact of a crash on traffic can last up to 210 minutes after the crash occurrence. Therefore, 210-minute interval is set for post-crash zone to assure that there is no part of post-crash zone contained in non-crash zone.

With almost six years data (from Jan 22, 2002 to Dec 31, 2007), the population of NTS is 1'160'831. The PTS population for 120 crashes is 720. It is obvious that the sizes of two data sets (NTS and PTS) are imbalanced with the imbalance ratio IMRO=1'612.27.

5.5. Data Preparation

Data preparation includes pre-processing steps to obtain NTS and PTS sets from raw individual vehicle data.

5.5.1.1. Raw Data Cleansing

Raw data are stored in text files and therefore, it's necessary to control whether data lines (with class of vehicles excluded) can be converted into numerical data. Raw data can be corrupted under many forms such as:

- Reset text of detectors. Traffic detectors usually reset counters after midnight when the traffic is low. Special text appears in data files indicating the reset. Reset duration usually lasts for about 15 minutes.
- Failure of detectors. The frequency of failure is higher in 2002, 2003, and 2004 at study site CH023. Special characters appear when the failure happens.
- Errors within data lines. Data lines should be in the format presented in Figure 4-3. Otherwise, the data lines should be removed.

When all the corrupted text in raw data is removed, raw data are aggregated based on lanes.

5.5.1.2. Aggregated Data Cleansing

As 5-minute intervals are used in the current study for aggregating data, any data interval containing insufficiently raw data for the interval is removed. This usually occurs when there is a detector reset and the reset starts within an aggregation interval. The aggregation interval starts at minute 0 or minute 5 (for example, at 12:00 or 17:45 or 21:15). If the reset starts at 12:33 and finishes at 12:47, raw data from 12:33 to 12:47 is not available. Moreover, the intervals from 12:30-12:35 and from 12:45-12:50 are also removed as there is not sufficiently raw data for those intervals. The same data cleansing is also applied to the intervals when there is failure of detectors.

In case of errors within data lines, corresponding lines are simply removed.

5.5.1.3. Missing Values

When data for two lanes of the same direction are aggregated, the aggregated data for one lane is matched with aggregated data for the other lane to generate traffic situations. There are periods when there is no vehicle on one lane whereas there are vehicles on the other lane. In this case, data for the no-vehicle lane are generated to match with the lane where there are vehicles. 7 lane-based parameters are generated for no-vehicle lane (see Table 5-1):

- Volume: 0 vehicles.

- Occupancy: 0%.
- Average Speed: 0 km/h.
- Average headway, standard deviation of headway, standard deviation of speed, percentage of heavy vehicle are also set to zero, similarly to average speed.

The zero value of volume and occupancy reflects the reality as there is no vehicle during the aggregation interval. However, the zero value for speed is a dummy value because if there is no vehicle, there is no speed. Similarly, dummy zero values are assigned to Average headway, standard deviation of headway, standard deviation of speed, percentage of heavy vehicle. These dummy values should not influence the final results as the corresponding traffic situations will be clustered into a special group.

5.6. Summary

This chapter discusses the application of the first step of the methodology to the data collected form study site. 22 variables are used to characterize Traffic situations that are aggregated for 5-minute intervals. Before NTS and PTS are specified, crash time correction is discussed with crash time estimation using shockwave theory. To agree with aggregation time interval, crash time is shifted earlier to match with the end of the last traffic situation.

Other important design choices are made for pre-crash period such that before shifted crash time, there are 6 PTS characterizing the traffic evolution before the crash. There are data unused which are data within pre-crash buffer period and post-crash period. The exclusion of these periods should not influence the overall performance of developed models.

Finally, the working data set of the current research with the selected study includes 1’160’831 NTS and 720 PTS (for 120 crashes). The imbalance ratio of the classes is IMRO=1’612.27.

Chapter 6 Data Sampling & Traffic Regimes

This chapter presents design choices with regard to data sampling process and provides in depth analysis of the traffic conditions obtained from clustering process called *traffic regimes*. Traffic situations are transformed before being sampled. After the data sampling process, traffic situations under original form are used in subsequent chapters.

It is worth noting that the terminology "traffic regime" might be used elsewhere to indicate totally different meanings. Traffic Regimes in the current research should not be linked to any other study.

6.1. Introduction

According to preliminary crash analyses presented in section 4.5.4, most of the crashes occurring at the selected study site are rear-end and sideswipe crashes (120 out of totally 170 crashes). Figure 4-15 in particular show that these crashes happen mostly under high flow or congested conditions. This means that there must be certain dominating traffic conditions for the considered collisions to appear and there is low chance for these types of crashes under other traffic conditions.

Section 3.3.1 presents an example of model development using all available pre-crash and non-crash data without sampling non-crash data with bad results obtained. It means that there is an imbalance between pre-crash and non-crash data that lower the performance of machine learning methods.

Previous studies in the literature also mention partly to the need of sampling non-crash data as summarized in section 2.4.4. However, the data sampling approaches are rather arbitrary and there is non-crash data unused in the sampling process.

Here, the data sampling methodology proposed in section 3.3 is applied to NTS and PTS data sets introduced in Chapter 5. Section 6.2 will discuss about the design choices of non-crash data sampling methodology. Section 6.3 will provide analysis on results of data sampling process which are called *traffic regimes*. Section 6.4 discusses the link between NTS and PTS with traffic regimes.

6.2. Design Choices

6.2.1. Overview

According to section 3.3.2, three main sampling steps are required:

- Normalization and dimension reduction
- NTS clustering
- PTS classification

The following sub-sections examine these steps with making design choices appropriate to available data at the selected study site.

6.2.2. Normalization and dimension reduction

6.2.2.1. Motivation

The motivation for normalization and dimension is presented in section 3.3.2.1. With the data from selected study site, the facts below motivate the reduction of dimensions.

According to section 5.6, the NTS matrix contains 1'160'831 observation characterized by 22 variables making it a large data matrix. Reducing number of dimensions is important as the clustering process can be time and computer memory consuming. The expected model should work in real-time and hence, decreasing the processing time can make the model more pro-active.
There exists high correlation between several pairs of variables. Figure 6-1 presents the correlation coefficients between couples of 22 variables (the variables are listed in Table 5-1). Here, the high coefficients indicate high correlations represented by darker color. The cells in top-left to bottom-right diagonal represent the correlation of each variable to itself, which is always 1. There are pairs of highly correlated variables such as *X10 (HFlow)* and *X13 (HOcc)*, *X11 (HASpd)* and *X17(Spd#)*, and *X5(LAHw)* and *X7(LVHw)*. Two correlated variables represent similar information and therefore bring spare information to the clustering process.

Variable X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	1.00	0.00	0.45	0.06	0.50	0.24	0.54	0.21	0.25	0.27	0.27	0.37	0.23	0.06	0.31	0.07	0.29	0.05	0.02	0.04	0.00	0.00
2	0.00	1.00	0.09	0.14	0.11	0.15	0.09	0.21	0.42	0.04	0.11	0.05	0.05	0.09	0.03	0.07	0.06	0.00	0.00	0.00	0.00	0.01
3	0.45	0.09	1.00	0.36	0.73	0.80	0.78	0.30	0.05	0.85	0.25	0.62	0.77	0.42	0.42	0.06	0.36	0.13	0.01	0.03	0.00	0.04
4	0.06	0.14	0.36	1.00	0.22	0.65	0.24	0.03	0.37	0.49	0.11	0.25	0.64	0.24	0.20	0.08	0.21	0.03	0.24	0.09	0.01	0.08
5	0.50	0.11	0.73	0.22	1.00	0.59	0.89	0.22	0.01	0.53	0.49	0.40	0.47	0.05	0.52	0.02	0.55	0.06	0.00	0.01	0.04	0.03
6	0.24	0.15	0.80	0.65	0.59	1.00	0.62	0.07	0.43	0.80	0.10	0.53	0.86	0.40	0.38	0.02	0.31	0.09	0.09	0.08	0.01	0.03
7	0.54	0.09	0.78	0.24	0.89	0.62	1.00	0.26	0.03	0.58	0.45	0.48	0.51	0.11	0.52	0.03	0.52	0.05	0.01	0.01	0.02	0.03
8	0.21	0.21	0.30	0.03	0.22	0.07	0.26	1.00	0.33	0.22	0.06	0.15	0.22	0.06	0.02	0.06	0.07	0.02	0.01	0.00	0.02	0.04
9	0.25	0.42	0.05	0.37	0.01	0.43	0.03	0.33	1.00	0.23	0.06	0.12	0.23	0.16	0.08	0.15	0.06	0.01	0.16	0.11	0.01	0.02
10	0.27	0.04	0.85	0.49	0.53	0.80	0.58	0.22	0.23	1.00	0.17	0.53	0.92	0.47	0.39	0.03	0.32	0.06	0.04	0.15	0.01	0.03
11	0.27	0.11	0.25	0.11	0.49	0.10	0.45	0.06	0.06	0.17	1.00	0.29	0.08	0.32	0.52	0.05	0.95	0.01	0.02	0.02	0.45	0.04
12	0.37	0.05	0.62	0.25	0.40	0.53	0.48	0.15	0.12	0.53	0.29	1.00	0.48	0.38	0.20	0.10	0.21	0.00	0.01	0.02	0.39	0.02
13	0.23	0.05	0.77	0.64	0.47	0.86	0.51	0.22	0.23	0.92	0.08	0.48	1.00	0.42	0.37	0.01	0.28	0.04	0.09	0.14	0.02	0.02
14	0.06	0.09	0.42	0.24	0.05	0.40	0.11	0.06	0.16	0.47	0.32	0.38	0.42	1.00	0.20	0.06	0.24	0.00	0.01	0.03	0.14	0.02
15	0.31	0.03	0.42	0.20	0.52	0.38	0.52	0.02	0.08	0.39	0.52	0.20	0.37	0.20	1.00	0.08	0.57	0.03	0.06	0.04	0.11	0.02
16	0.07	0.07	0.06	0.08	0.02	0.02	0.03	0.06	0.15	0.03	0.05	0.10	0.01	0.06	0.08	1.00	0.07	0.01	0.03	0.01	0.05	0.01
17	0.29	0.06	0.36	0.21	0.55	0.31	0.52	0.07	0.06	0.32	0.95	0.21	0.28	0.24	0.57	0.07	1.00	0.02	0.06	0.01	0.44	0.02
18	0.05	0.00	0.13	0.03	0.06	0.09	0.05	0.02	0.01	0.06	0.01	0.00	0.04	0.00	0.03	0.01	0.02	1.00	0.05	0.33	0.03	0.00
19	0.02	0.00	0.01	0.24	0.00	0.09	0.01	0.01	0.16	0.04	0.02	0.01	0.09	0.01	0.06	0.03	0.06	0.05	1.00	0.31	0.03	0.00
20	0.04	0.00	0.03	0.09	0.01	0.08	0.01	0.00	0.11	0.15	0.02	0.02	0.14	0.03	0.04	0.01	0.01	0.33	0.31	1.00	0.03	0.00
21	0.00	0.00	0.00	0.01	0.04	0.01	0.02	0.02	0.01	0.01	0.45	0.39	0.02	0.14	0.11	0.05	0.44	0.03	0.03	0.03	1.00	0.01
22	0.00	0.01	0.04	0.08	0.03	0.03	0.03	0.04	0.02	0.03	0.04	0.02	0.02	0.02	0.02	0.01	0.02	0.00	0.00	0.00	0.01	1.00

Variable X

Figure 6-1: Correlation coefficients between pairs of variables

6.2.2.2. Normalization

There are various normalizations in statistics such as standard score, Student's t-statistic, Studentized residual, Standardized moment, and Coefficient of variation, etc. Here, standard score method is used because all traffic data are used and therefore, the population parameters are known.

6.2.2.3. Dimension Reduction Techniques

As discussed in section 3.3.2.2, a method for dimension reduction in the current research is a feature extraction method. As such, there are many candidates that can be used to reduce the number of dimensions of NTS matrix. Several examples of feature extraction techniques include principal component analysis, multifactor dimensionality reduction, partial least squares, or self-organizing maps, etc.

The selection of technique for dimension reduction can greatly influence the performance of models developed in the next chapters. However, this choice is challenging at the current step as the models have not been developed so far. Therefore, a technique is selected for the initial design choices based on the characteristics of available data which is NTS matrix in the current research.

As results, Principal Component Analysis – PCA is chosen as the technique for dimension reduction because:

- PCA is a mathematical procedure transforming linearly a number of possibly correlated variables into a smaller number of uncorrelated variables called *Principal Components* (*PC*) according to Jolliffe, (2002). The order of a PC indicates its importance in representing the variability in the data. For example, the first PC is the most important and accounts for as much of the variability in the data as possible. The second PC is the second most important and accounts for as much of the remaining variability as possible and so on.
- PCA is used to transform normalized data into new data space. As the number of the most important PC is smaller than the number of original variables, data dimension reduction is obtainable. In addition to that, the PCs which are less important are eliminated such that only the most relevant information is kept for the clustering process. Finally, with only two or three first principle components, transformed data can be visualized in a manner that most of the variations of the original data are represented.

It is worth noting that the choice of PCA as dimension reduction does not guarantee the optimal performance of the developed models. Therefore, if the final performance is low, the choice of techniques for dimension reduction is subject to an optimization process.

6.2.2.4. NTS Transformation

As discussed in section 5.2.3, the first two variables TDay and WDay characterizing traffic situations are categorical, which are not applicable for the standard PCA. These variables are excluded from PCA transformation process and will be used again in development of risk identification models (presented in the next chapters). Thereby, the remaining NTS matrix composes of 20 data fields from *X3* to *X22*. As results, 20 eigenpairs are obtained. Figure 6-2 presents the obtained eigenvalues sorted descending (the first eigenvalue is the greatest) under the form of pareto chart.

According to Figure 6-2, the first eigenvalue represents up to 30% of variances in the original data, the second eigenvalue up 15%, i.e. only the first two eigenvalues can represent up to 45% of variances in the original data. Variables contributing the most to the first PC include *X5*, *X7*, *X3*, *X10*, *X6*, and *X13*. The difference between the sign of *X5* and *X7* (positive) and the sign of *X3*, *X10*, *X6*, and *X13* (negative) for the first PC indicates that the first PC is highly influenced by the volume of traffic on the road section. It is reasonable that when *X3* (right lane's flow) is high, *X10* (left lane's flow) is also high. As *X6* (right lane's occupancy) and *X13* (left lane's occupancy) are highly correlated with *X3* and *X10*, respectively (according to Figure 6-1), if *X3* (or *X10*) is high, *X6* (or *X13*) is also high. On the contrary, if *X3* is high, *X5* (right lane's average headway) and *X7* (right lane's standard deviation of headway) are low as *X5* and *X7* are also highly correlated to *X3* but inversely. When the traffic volume becomes low, the average headway and the standard deviation of headway will become high.

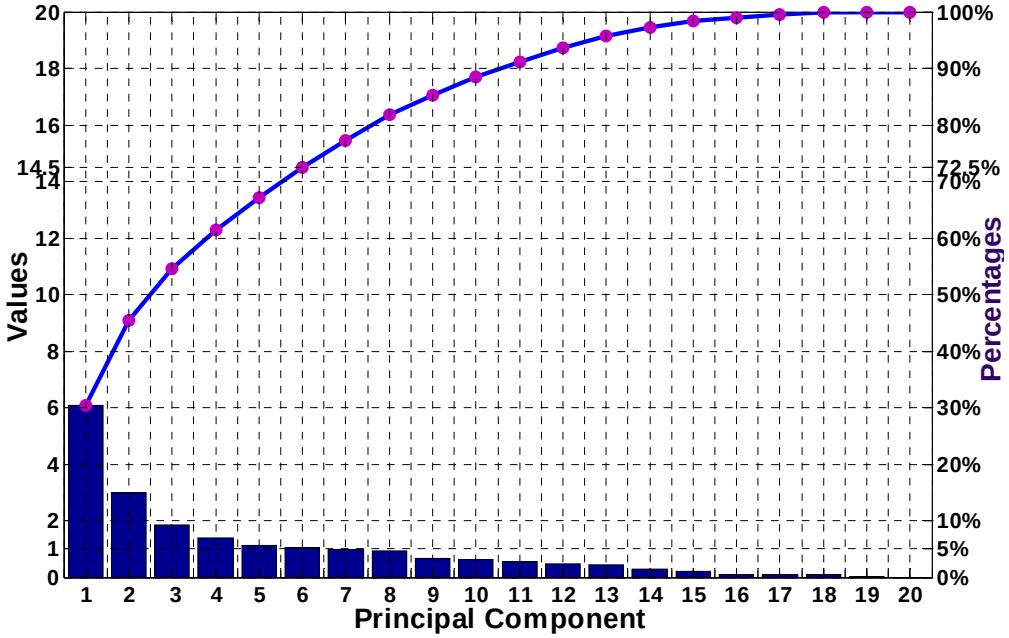

Figure 6-2: Principal Components as outputs of PCA transformation

For the second PC, the contribution of left lane's variables such as *X11* (average speed), *X21* (speed change), *X14* (standard deviation of headway), *X12* (occupancy), and *X15* (standard deviation of speed) is high. *X17* highly correlated with *X11* is also an important variable contributing to the second PC.

According to Figure 6-2, the third PC representing 9.2% of the total variance is influenced mostly by speed-related variables one the right lane (*X4*, *X8*, and *X19*). Besides, the percentage of heavy vehicles – *X9* on the right lane is the variable contributing the most to the third PC.

The forth PC is influenced by the traffic volume evolution represented by *X18* and *X20* (flow differences compared to the previous TS on the right lane and on the fast lane, respectively). The low influence of the speed changes on the two lanes indicates that the volume does increase but the section is still not under congestion. The fifth PC is more likely to represent the speed variation on the left lane (*X8*), on the left lane (*X15*), and speed evolution on the left lane (*X21*). The sixth PC is highly influenced by the precipitation information. Each of the fourth, fifth, and sixth PC represents smaller than 7% of the total variance.

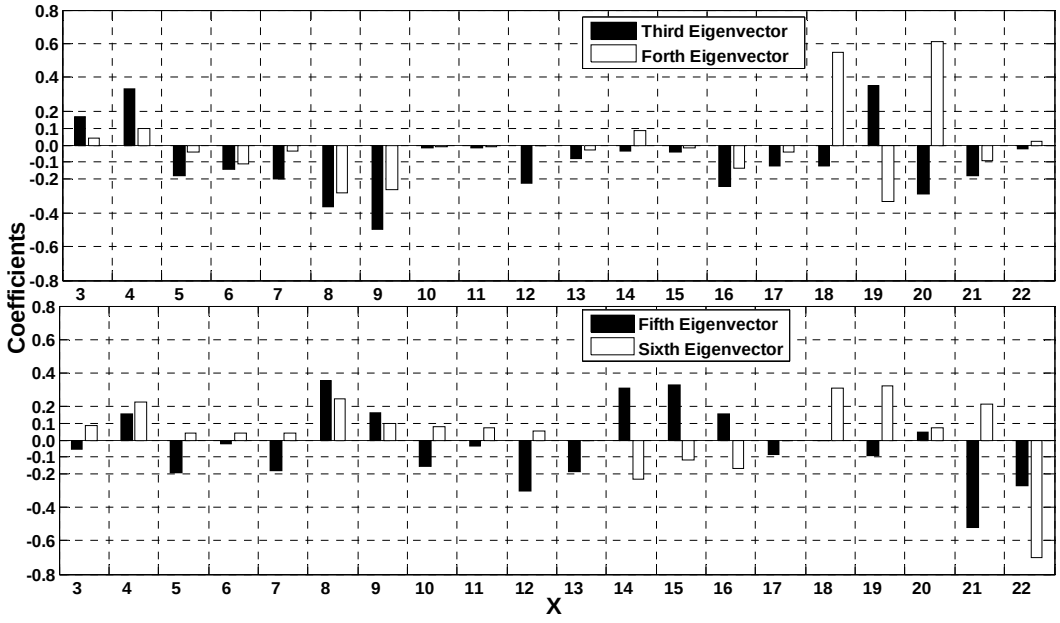

Figure 6-3: Coefficients of four eigenvectors ranked from the third to the sixth

Figure 6-4 presents the coefficients for 20 original variables in the space of the first two PC. As the first two eigenvalues represent up to 45% of variances in the original data, the data variances are visualized under the space of the first two PC better than under the space of any other two PC. Vectors representing variables *X3* (*LFlow*), *X10* (*HFlow*), *X6* (*LOcc*), and *X13* (*HOcc*) are the longest vectors in the first two PC. Therefore, they have higher influence on the data distribution in the space of the first two PC. In the other way, data points in the space of the first two PC are distributed according to high influence of flow and occupancy.

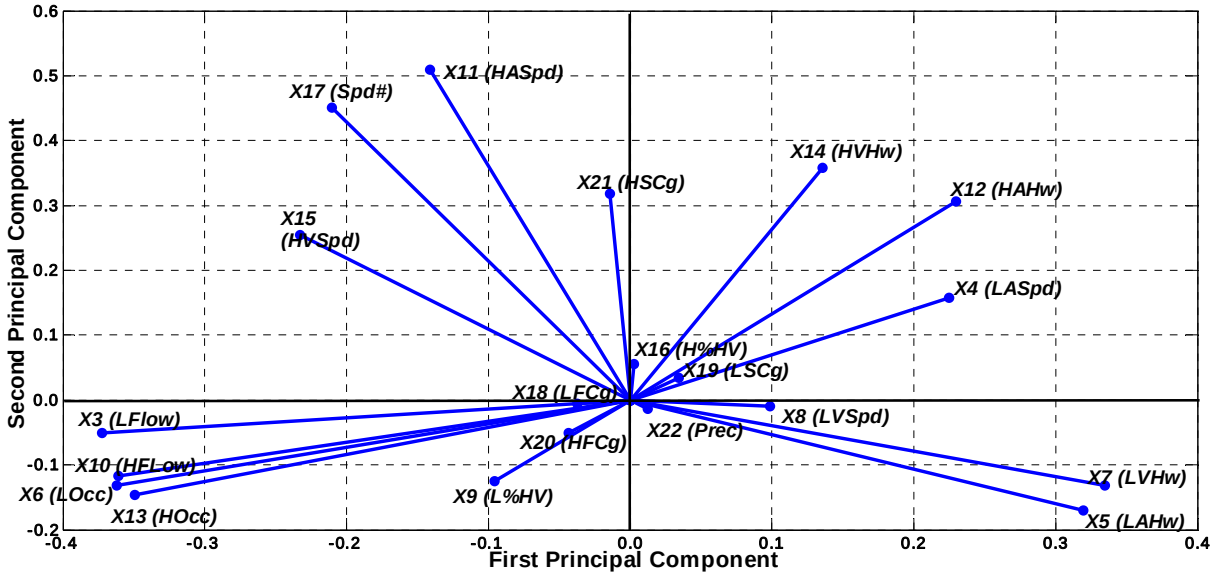

Figure 6-4: Coefficients of the first two eigenvectors

Figure 6-5 presents the transformed data in the space of the first two PC representing two most important data dimensions of the transformed data. Lines in Figure 6-4 represent coefficients of the first two eigenvectors and are the same to the lines illustrated in Figure 6-4. Note that the lengths of coefficient vectors are scaled up ten times.

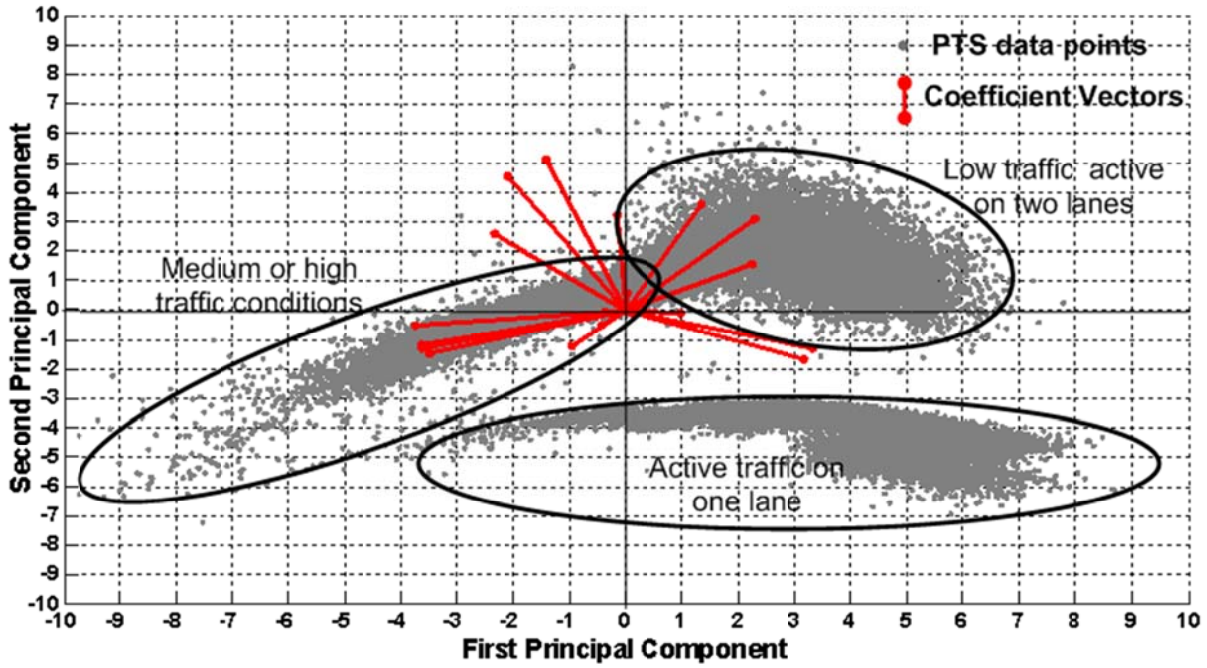

Figure 6-5: Transformed TS in the space of the first two PC

Under this space of the first two PC, data points can be intuitively partitioned into three groups (cross-checked with data in matrix *X*):

i) Active traffic on one lane. This occurs when the traffic volume is low and vehicles travel mostly on the left lane. There are also high-flow and low-speed data points belonging to this group and lying in the third quarter corner (where the first and the second PC are negative), i.e. the traffic is congested on one lane and the other lane is closed for some reason.
ii) Low traffic active on two lanes. There are vehicles on both of the lanes but the volume is still low.
iii) Medium or high traffic conditions. The traffic volume is higher compared to the traffic volume of group ii). Data points belonging to this groups and close to left edge of Figure 6-5 represent congested conditions with very high flow and very low speed.

The data can also be divided into smaller groups if other PC are taken into account.

6.2.2.5. PTS Transformation

PTS are not combined with NTS before NTS transformation using PCA. However, outputs of NTS transformation are used to transform PTS. Outputs of NTS transformation include:

- Normalization function *Norm*: $\bar{X}_{PTS} = Norm(X_{PTS})$ that is characterized by two parameters: mean and standard deviation vectors of NTS matrix.
- Transformation function *FEx:* $X'_{PTS} = FEx(X_{PTS})$ that is characterized by six eigenpairs (eigenvalue and eigenvector) corresponding the first six PC.

Figure 6-6 presents the positions of PTS data points in the space of the first two PC. It can be seen that most of the PTS data points focus in the area where there is group iii) of NTS data points presented in section 6.2.2.4. It is expected that the clustering process will be able to match PTS with those NTS as they (i.e. both PTS and NTS) represent similar traffic conditions that makes them (PTS and NTS) comparable.

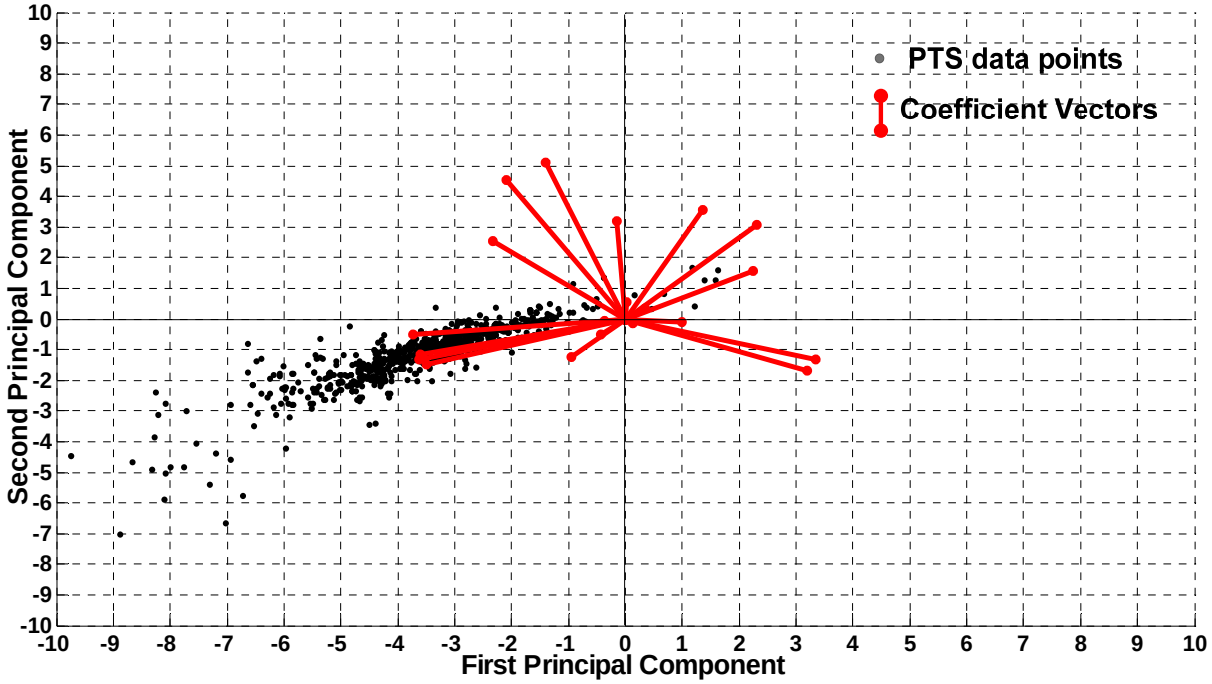

Figure 6-6: PTS in space of the first two PC

6.2.2.6. Choice of Dimensions

As NTS matrix is normalized with zero means before being transformed using PCA, the sum of obtained eigenvalues is the number of dimensions, i.e. 20 and the average eigenvalue is 1.

Many approaches can be applied to determine the value of *R* representing the number of PC to be retained (according to Hayton et al., (2004)). The main idea of the approaches is to retain PC until additional PC account for trivial variance according to some PC retention criteria.

One of the most popular PC retention criteria is to retain all PC having eigenvalues greater than the average eigenvalue (which is 1.0 in our case). The rationale behind the criterion is that *a factor should account for the amount of variation in at least one variable* (Kaiser, 1960), i.e. a factor to be retained should have the eigenvalue not smaller than the average eigenvalue.

Applying this criterion, all the PC having eigenvalue not smaller than 1.0 are retained. As results, the first six PC satisfy this criterion. As illustrated in Figure 6-2, the first six PC can preserve up to 72.5% of the variance in the original data. The other 27.5% of variance in the original data is excluded from clustering process.

It is important to note that six PC representing 72.5% of variance in the original data are only applied to the clustering process. Later on when traffic risk identification models are developed, the excluded 27.5% of variance will be re-considered in order to discover the variation between NTS and PTS.

6.2.3. NTS clustering

6.2.3.1. Clustering Method

The objective of using a clustering algorithm in the current study is to make the natural classification of TS data such that groups of NTS can be obtained before PTS are classified into the groups. Clustering is the core part of data sampling process aiming to produce a base for matching NTS with PTS. Many clustering techniques can help to achieve this objective such as Self-Organizing Maps, K-means, etc.

K-mean (Jain, 2010) is selected as clustering technique in the current research. This technique is still widely used until today thanks to its simplicity, efficiency, and ease of implementation. The other motivation for choosing K-means as clustering algorithm in the present study includes:

i) Hierarchical clustering techniques are not desirable as there is no underlying structure in NTS data set and the objective of clustering NTS is to find the natural classification of NTS.
ii) K-means method requires less memory than other clustering techniques including hierarchical clustering techniques and Self-Organizing Maps.
iii) The number of clusters is flexible to be chosen and can be any positive integer number. For several clustering algorithms such as Self-Organizing Maps (Kohonen, 1982), the number of clusters needs to be an integer divisible by an integer greater than 1 to form a rectangular cluster map.
iv) Clustering using K-means can more easily find the global near-optimal solutions if the inputs are transformed by PCA. This is because PCA automatically projects to the subspace where the global solution of K-means clustering lie (Ding and He, 2004).

Apart from to the choice of clustering technique, Euclidean metric is the distance metric used in K-means calculations.

6.2.3.2. Number of Clusters

Let C the set of cluster centers: $C=[C_1, C_2, ..., C_K]$ where C_j ($j=1 \div K$) is the cluster center of the j-th cluster, D_{NTS} is the set of index sets: $D_{NTS}=[D_{NTS1}, D_{NTS2}, ..., D_{NTSK}]$ where D_{NTSj} ($j=1 \div K$) is the set of indices of all data points x_i belonging to of the j-th cluster (x_i is NTS data transformed using PCA). The algorithm K-means tries to iteratively minimize the clustering error presented in Equation 6.

Equation 6: Squared Errors of K-means algorithm

$$Err = \sum_{j=1 \div K} \sum_{(i=1 \div P) \& (i \in D_j)} (x_i - C_j)^2$$

Err value reflects the homogeneity/variance of data within each cluster. If each data point x_i represents a cluster, *Err* will be zero, i.e. there is no variance under each cluster. *Err* becomes greatest when all data points are grouped into one cluster, i.e. the variance is maximum and the homogeneity is minimum. When the number of clusters increases, the variance decreases and the homogeneity increases. The objective of this sub-section is to determine the number of cluster the most appropriate for the NTS data.

Figure 6-7 presents four different indicators on clustering errors for different numbers of clusters (from 1 to 20). The current study limits the number of clusters to be considered to 20 as each cluster represents a traffic regime and for operational point of view, more than 20 traffic regimes are not necessary.

The curve *E1* represents the absolute clustering errors (i.e. *Err* in Equation 6) corresponding to numbers of clusters. If the whole data set is considered as a cluster, the clustering error is the highest at around 14.10^6. When there are two clusters, the clustering error reduces dramatically to below 10.10^6 compared the clustering error when one cluster is used. The strong reduction of clustering error slows up until there are six clusters. Thereafter, the clustering error keeps reducing at much lower rate.

The curve *E2* provides another view on the reduction of clustering error: *E2* value at each number of clusters *K* is the difference between the clustering error when there are *K* clusters and the clustering error when there are (*K*-1) clusters. As such, there is no clustering error difference for K=1. Therefore, E2 value for *K*=1 is zero. In fact if *E1(K)* and *E2(K)* are the absolute error and the error difference, respectively, when there are *K* clusters, *E2(K)* is calculated as: $E2(K) = E1(K-1) - E1(K)$

E2 value is much reduced until K=7. After that with K=8, 9, the reduction of E2 is stable at low speed before coming to another drop when K changes from 9 to 10. Therefore, the candidates for K value according to E2 values are 7, 8, 9, and 10.

The curve *P1* represents the percentages of error reduction $P1_K$ when changing from *K*-1 clusters to *K* clusters compared to the clustering error with 1 cluster. $P1_K$ is calculated according to Equation 7. $P1_K$ represents the additional error reduction that can be gained when one more cluster is used compared to the initial clustering error Err_1.

Equation 7: Percentages of error reduction compared to Err_1

$$P1_K = \frac{Err_{K-1} - Err_K}{Err_1}, where\ Err_l\ is\ clustering\ error\ with\ l\ clusters$$

The curve *P2* includes $P2_K$ that is similar to $P1_K$ except that the error difference is compared with the error for (*K*-1) clusters. $P2_K$ is calculated according to Equation 8. $P2_K$ represents the relative error reduction when deciding to increase the number of clusters from *K*-1 to *K*.

Equation 8: Percentages of error reduction compared to the last error

$$P2_K = \frac{Err_{K-1} - Err_K}{Err_{K-1}}, where\ Err_l\ is\ clustering\ error\ with\ l\ clusters$$

For two curves *P1* and *P2*, *K*=7, 8, 9 and 10 are also the good candidates. Therefore, other data insights need to be undertaken to choose one of the four values. The criteria to one of four numbers *K* include:

i) Lower number of clusters is more preferred. This is important from traffic operational point of view.
ii) As one of the main objectives of this study is to differentiate PTS and the relevant NTS, the additional cluster should facilitate the differentiation.

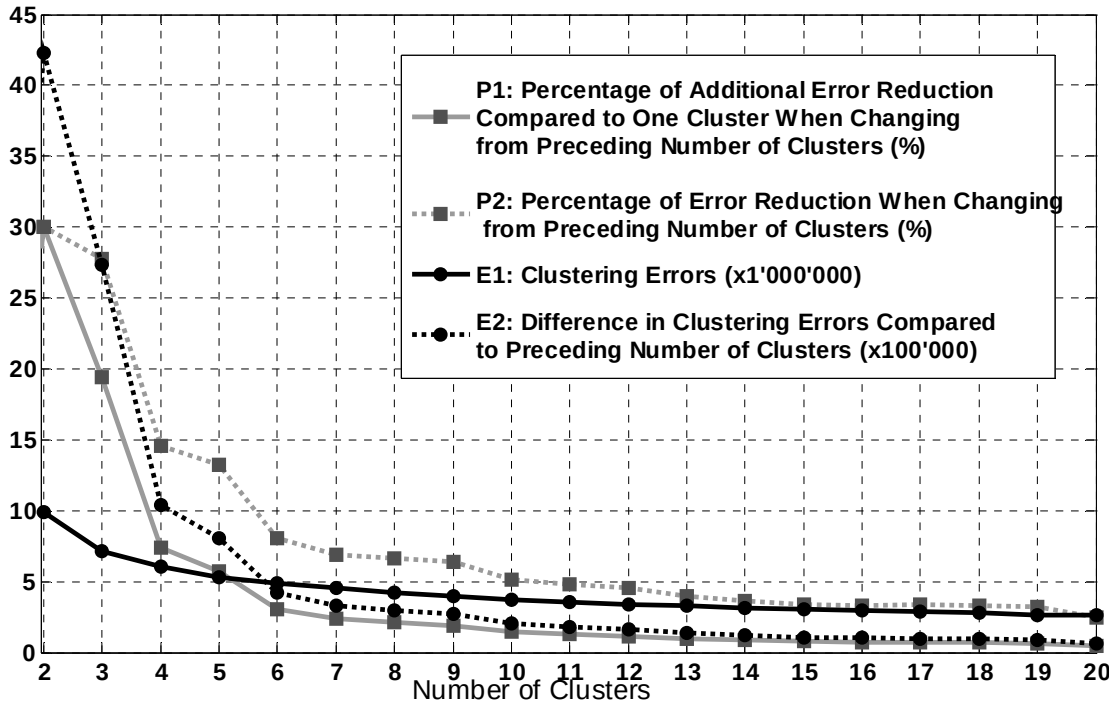

Figure 6-7: Clustering errors by the number of clusters

Figure 6-8 illustrates cluster centers in the space of the first two PC corresponding to different numbers of clusters *K* with the symbol *CK.l* indicating the the *l*-th cluster center when there are *K* clusters. Although the illustration is in two-dimension spaces, the analysis is undertaken in the space of six PC. As the first two PC represent more than 62% out of the total variance represented by the first six PC (i.e. 45% versus 72.5%), the illustration in Figure 6-8 almost matches with the analysis under the space of six PC.

The first two cluster centers are almost identical with any *K* value. According to the criterion i), 7 clusters are the most preferred, then 8 clusters, 9 clusters, and finally 10 clusters. When changing from 7 clusters to 8 clusters, a new cluster is added: three clusters C7.5, C7.6 and C7.7 are redistributed into four clusters C8.5, C8.6, C8.7, and C8.8. It's clearly illustrated through Figure 6-8 that clusters C8.5, C8.6, C8.7, and C8.8 lie in an area where there PTS data points. Therefore, the change from 7 to 8 clusters creates an additional cluster that can facilitate differentiating NTS and PTS and *K*=7 is rejected.

Consider the change from 8 to 9 clusters. An additional cluster is generated by dividing cluster C8.3 into two clusters C9.3 and C9.4. According to Figure 6-6, there is no PTS data point corresponding to cluster C8.3 area. Therefore, the additional cluster does not contribute to NTS and PTS differentiation and the number of clusters equal to 9 (K=9) is eliminated. Consequently, K=10 is also rejected. Finally, K=8 is selected as the number of clusters in the current study.

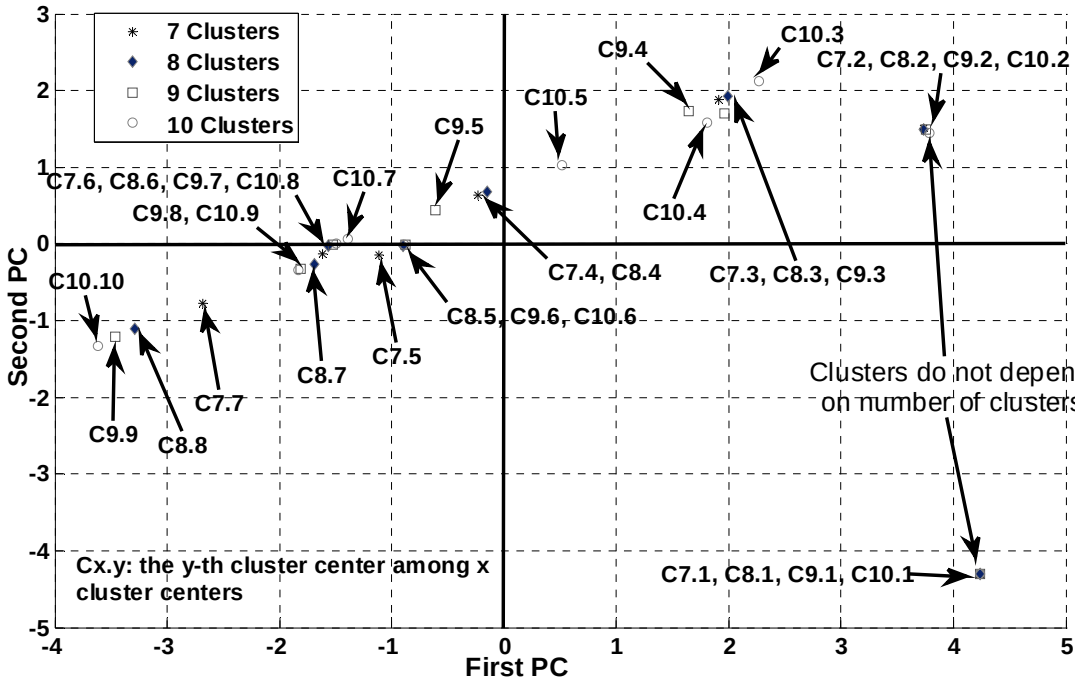

Figure 6-8: Cluster centers in cases of 7, 8, 9, and 10 clusters in the space of the first two PC

6.2.3.3. PTS classification

PTS in matrix X_{PTS} are normalized and transformed to X''_{PTS}. Thereafter, rows of X''_{PTS} are compared with cluster centers C_1, C_2... C_8. Similar to new *TS*, a *PTS* is classified into a cluster C_k ($1 \le k \le 8$) if the distance between the *PTS* data point and the center of cluster C_k is the lowest compared to the distances of the *PTS* data point to the centers of other clusters.

Let D_{PTS} is the set of index sets: D_{PTS} =[D_{PTS1}, D_{PTS2}, ..., D_{PTSK}] where D_{PTSj} ($j=1 \div K$) is the set of indices of all transformed PTS data points belonging to of the j-th cluster.

6.2.3.4. NTS & PTS Distribution

We call clusters as *Traffic Regimes – TR*, and named from *A* to *H* corresponding to cluster centers C_1, C_2...C_8, respectively. Under j-th TR, there are D_{PTSj} and D_{NTSj} – the sets of indices of PTS and NTS belonging to j-th TR. It is possible that the set D_{PTSj} is empty because there is no PTS classified into j-th TR. It is worth noting that the terminology "Traffic Regime" might be used elsewhere with different meanings and has no link with the Traffic Regimes used in the current research.

We define a parameter called *Risk Chance* representing the ratio between PTS population and NTS population under each TR. Risk Chance under a TR represents the priori probability for a new TS to become PTS when the TS is classified into that TR. Risk Chance value of a TR is the inverse value of Imbalance Ratio under that TR.

Figure 6-9 presents the distribution of NTS and PTS and the Risk Chance under each TR. There is no PTS available under regimes A and F. Therefore, Risk Chance value for these regimes is zero. Under regimes D and E, PTS populations are low leading to low Risk Chance. To obtain significantly statistical results, PTS population should be greater than 20. Therefore, regimes A, D, E, and F are declared *risk-free*, i.e. when a new *TS* is classified into one of regimes A, D, E, and F, it will be automatically identified as *NTS*. Under regimes B, C, G, and H, TR-based Risk Identification Models (RIM) are developed to differentiate NTS and PTS aiming to classify new TS into one of two classes NTS or PTS.

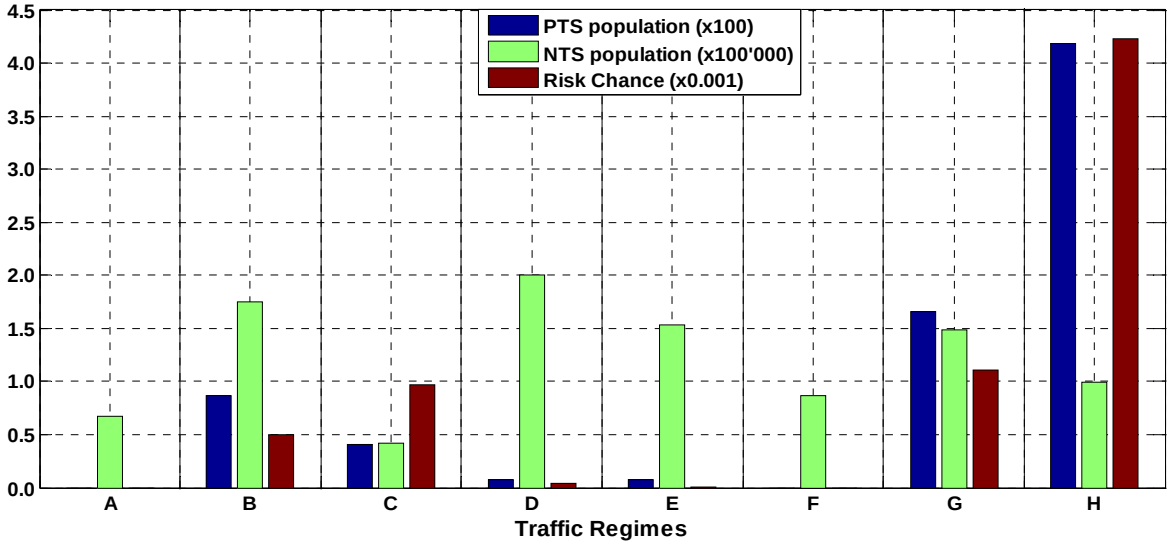

Figure 6-9: Distribution of NTS and PTS and Risk Chance under each Traffic Regime

6.3. Traffic Regime Analyses

6.3.1. Preliminary

According to Figure 4-14, PTS occur during day time, especially during peak hours (6:00-8:00AM and 4:00-6:00PM, weekday) whereas; there is low number of PTS occurring during the night (from 8:00PM to 6:00AM) when the traffic demand is low. This distribution of PTS is reasonable as the traffic must be sufficiently dense for traffic – induced crashes (i.e. rear-end and sideswipe crashes in the present study) to occur.

In Figure 6-10, the locations of traffic regimes represented by cluster centers are characterized by several variables pointing in the same direction. For example, regime F is characterized by high values of *X5*, *X7*, *X8*, and *X22* (i.e. lying in the fourth quarter of the first two PC space) and low values of *X17* (speed difference becomes negative), *X11* (left lane's speed becomes zero), and *X15* (speed variation becomes

zero). It means that regime F represent traffic conditions where there are vehicles only on one lane due to low traffic or due to some other reasons.

Two regimes A and E lie on the first quarter of the first two PC space and are most underlined by high values of *X4*, *X12*, and *X14* and low values of *X3*, *X10*, *X6*, and *X13*. It means that regimes A and E represent traffic conditions where there are vehicles on both lanes yet the flow is low. More principal components are needed to explain the difference between regimes A and E.

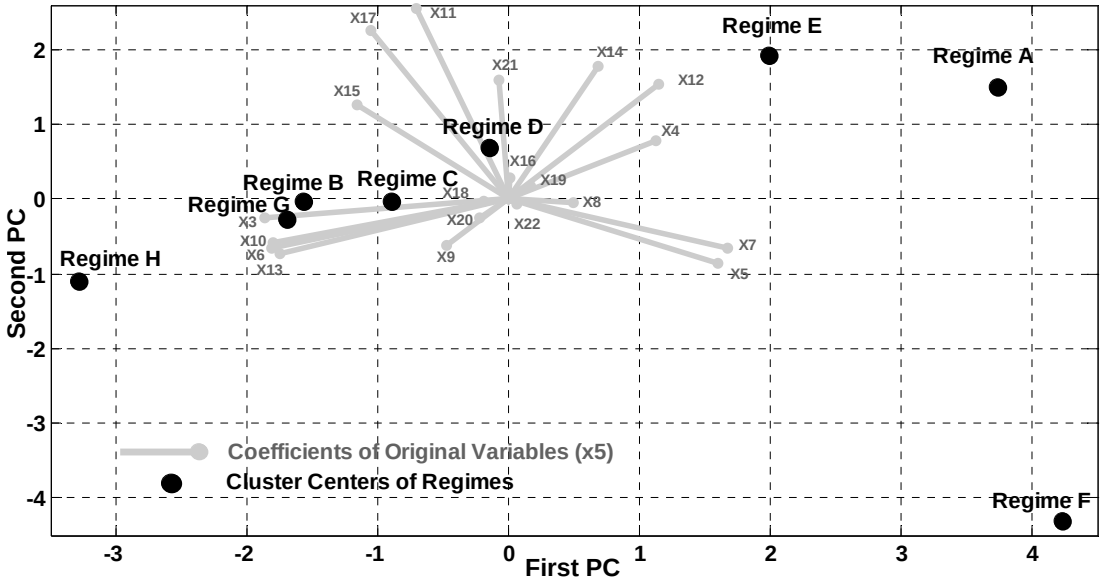

Figure 6-10: Location of cluster centers the first two PC space

Regime D is the only regime lying in the second quarter of the first two PC space and is strongly supported by *X15*, *X17*, *X11*, and *X21*. Regimes B, C, G, and H lie on the third quarter of the first two PC space and are indicated by high *X3*, *X6*, *X10*, and *X13*. B, C, G, and H are the traffic regimes where the population of PTS is high and the Risk Chance is also high.

Combining Figure 6-5 with Figure 6-10, traffic data under each TR are divisible into three main groups:

- group of traffic active on one lane: regime F
- group of low traffic active on two lanes: regimes A and E
- group of medium or high traffic on two lanes: regimes B, C, D, G, and H

Figure 6-11 and Figure 6-12 present traffic regimes in the fundamental speed/flow diagram for the right and left lanes, respectively. It can be seen that data points from one regime overlap with data points from some other regimes. However, the overlapping is apparent as there are other variables used for separating different traffic regimes, not only speed and flow on two lanes.

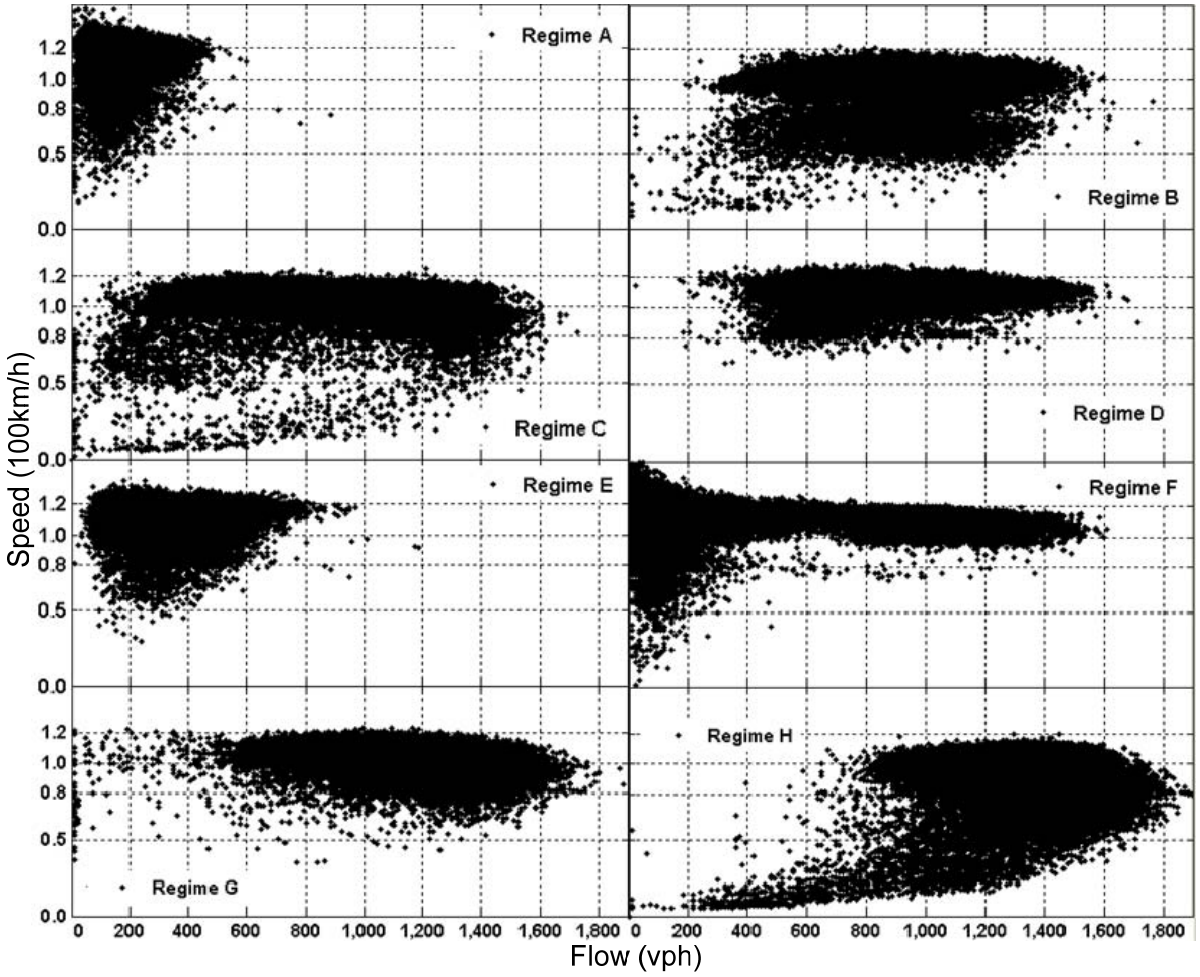

Figure 6-11: Traffic regimes under speed-flow diagram for the right lane

It can be easily recognized that the traffic under regime F is only active on one lane, mostly on the right lane and some Traffic Situations on the left lane. According to Figure 6-11, traffic flow on the right lane can be high up to 1400vph. This means that regime F represent traffic conditions where the left lane is closed due to some unknown reason.

Two regimes A and E also represent low flow conditions. The difference between regimes A and E is that the traffic flow is lower under regime A than under regime E, especially on the left lane.

For other regimes (i.e. regimes B, C, D, G, and H), speed and flow on two lanes cannot fully explain the differences.

To obtain more insights for each regime, more variables need to be used. In the subsequent sections, the detailed characterization of each regime will be presented.

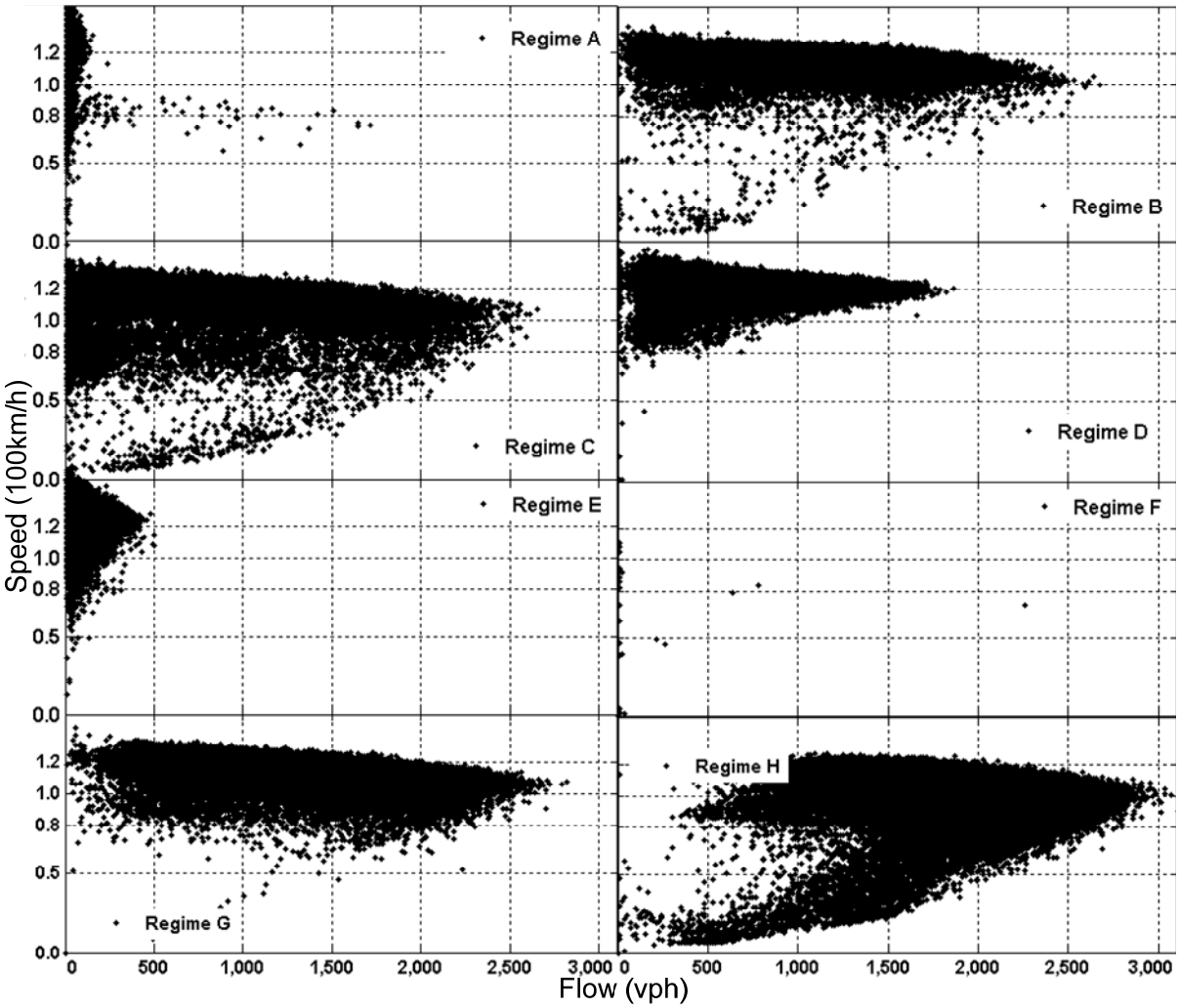

Figure 6-12: Traffic regimes under speed-flow diagram for the left lane

6.3.2. Traffic States on Each Lane

This section makes the differentiation between eight Traffic Regimes based on the traffic characteristics on each lane.Figure 6-13 presents statistics on variables indicating traffic states on both of the lanes, i.e. variables from *X3* to *X9* for the right lane and from *X10* to *X16* for the left lane.

Comparing variable *X3* (*HFlow*), Traffic Situations can be partitioned into three groups:

1) The group of low flow (*X3*<500vph) including regimes A, E, and F
2) The group of medium traffic (500vph ≤ *X3* ≤ 1100vph) including regimes B, C, and D.
3) The group of high traffic (*X3*>1100vph) including two regimes G and H.

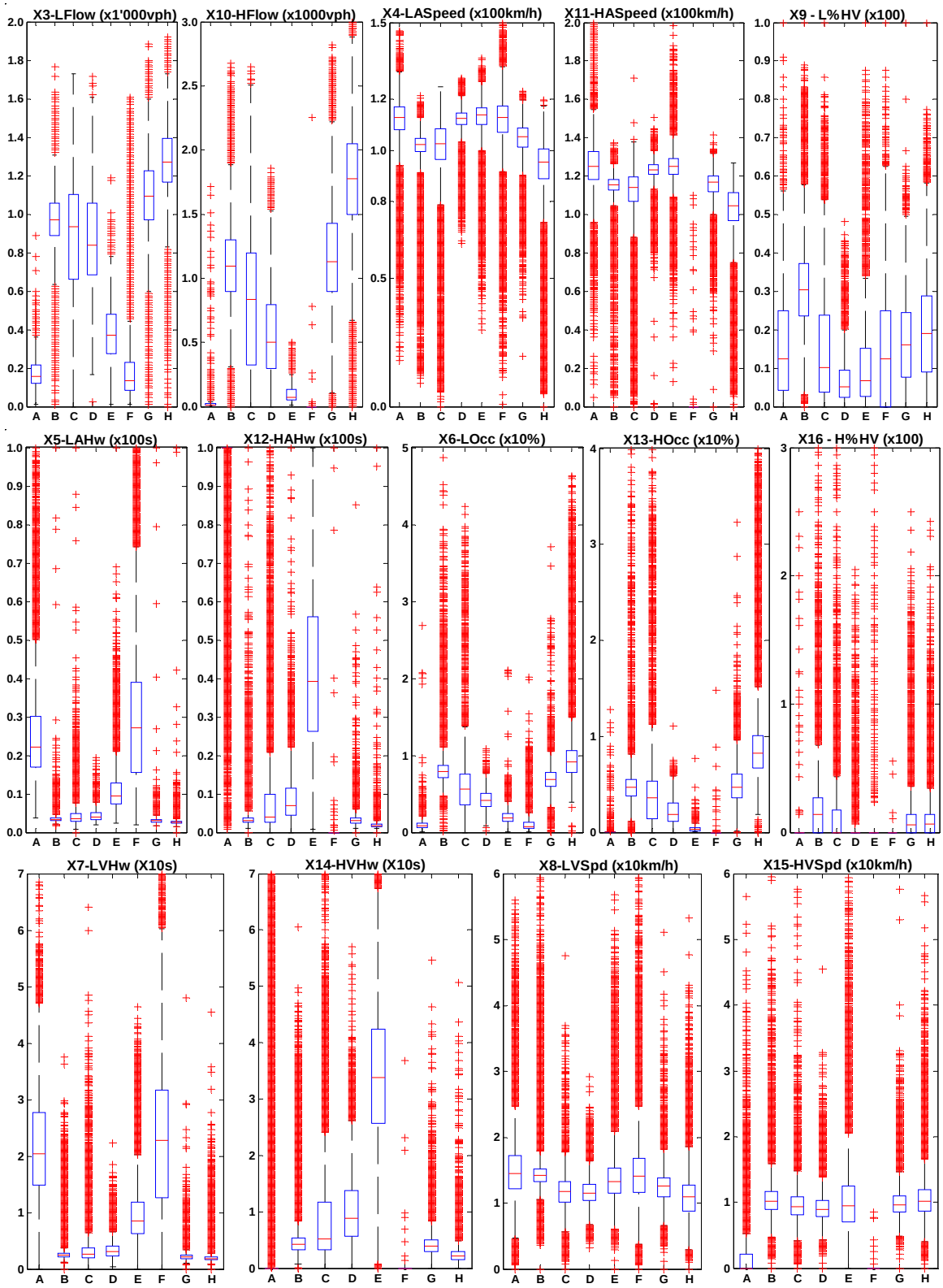

Figure 6-13: Statistics on variables representing lane states from *X3* to *X16*

Comparing variable *X10* (*HFlow*) between two regimes G and H, it can be stated that regime H represents high flow and congested conditions. Considering *X4* and *X11* (the average speeds on the right and left lanes), it is shown that the speed is lower under regime H than under any other regime. The high flow traffic under regime H is also reflected by the outstanding values of variables *X6* and *X13* (occupancy on the left and right lanes, respectively) compared to other regimes.

Regime G is the regime where traffic flows on both lanes (i.e. *X3* and *X10*) are the second highest (flows under regime H are the highest). According to Figure 6-11 and Figure 6-12, there is no congestion under regime G. Therefore, the average speeds under regime G are higher than under regime H. Regime B is also very similar to regime G by comparing variables from *X3* to *X8*, and *X10* to *X15*. The clearest difference between regimes G and B is the percentage of heavy vehicles on both lanes – *X9* and *X16*.

Regimes C and D are also similar by comparing most of variables representing lane states. Some significant difference between regimes C and D are:

- The average speeds on both lanes (*X4* and *X11*) are higher under regime D than under regime C.
- The occupancies on both lanes (*X6* and *X13*) are lower under regime D than under regime C.
- The percentages of heavy vehicles on both (*X9* and *X16*) are slightly lower under regime D than under regime C

Finally, using variables representing lane states allow differentiating the following traffic regimes or groups of traffic regimes:

- The free flow traffic conditions represented by regimes A, E, and F.
- The high flow and congested conditions represented by regime H.
- The high-medium traffic flow conditions represented by regimes B and G
- The low-medium traffic flow conditions represented by regimes C and D.

6.3.3. Traffic Variations and Non-traffic Characteristics

This section provides further data insights aiming to discover the characteristics making one regime different from the other regimes, especially to distinguish between regime B and regime G or between regime C and regime D. Figure 6-14 presents variables from *X17* to *X21* indicating the variations of traffic state across lanes and over two continuous traffic situations. Figure 6-14 also presents non-traffic variables *X1*, *X2*, and *X22*.

By comparing *X18* and *X20* (the changes of flow on the right and left lanes, respectively) of two regimes B and G, *X18* and *X20* under regime G are mostly negative which means that there are less vehicles for Traffic Situations under regime G compared to the traffic situations that precede. This reflects the reduction of traffic flow under regime G. That can be the reason explaining the increases of average speeds on both lanes (represented by *X19* and *X21*). For regime B, the inversed tendency occurs: the traffic flow is increasing and the average speed is reducing.

However, that tendency does not happen to regimes C and D. There is a slight increase of traffic flow under regime D compared to regime C. However, this change is not significant. It means that traffic characteristics under regimes C and D are similar.

Consider three non-traffic variables *X1*, *X2*, and *X22*, the most significant difference between regimes C and D is indicated by variable *X22* (the type of precipitation). More than 95% out of NTS belonging to regime C are under rainy conditions whereas only less than 1% of NTS under regime D are under rainy conditions. As *X22* is one of variables involved in clustering process, it contributes actively to the separation of traffic into regimes C and D. According to Figure 6-9, Risk Chance under regime C is higher than under regime D. According to Figure 6-13, the average speed is lower under regime C than under regime D. As most traffic characteristics are similar under regimes C and D, the type of precipitation is the reason for the risk chance difference.

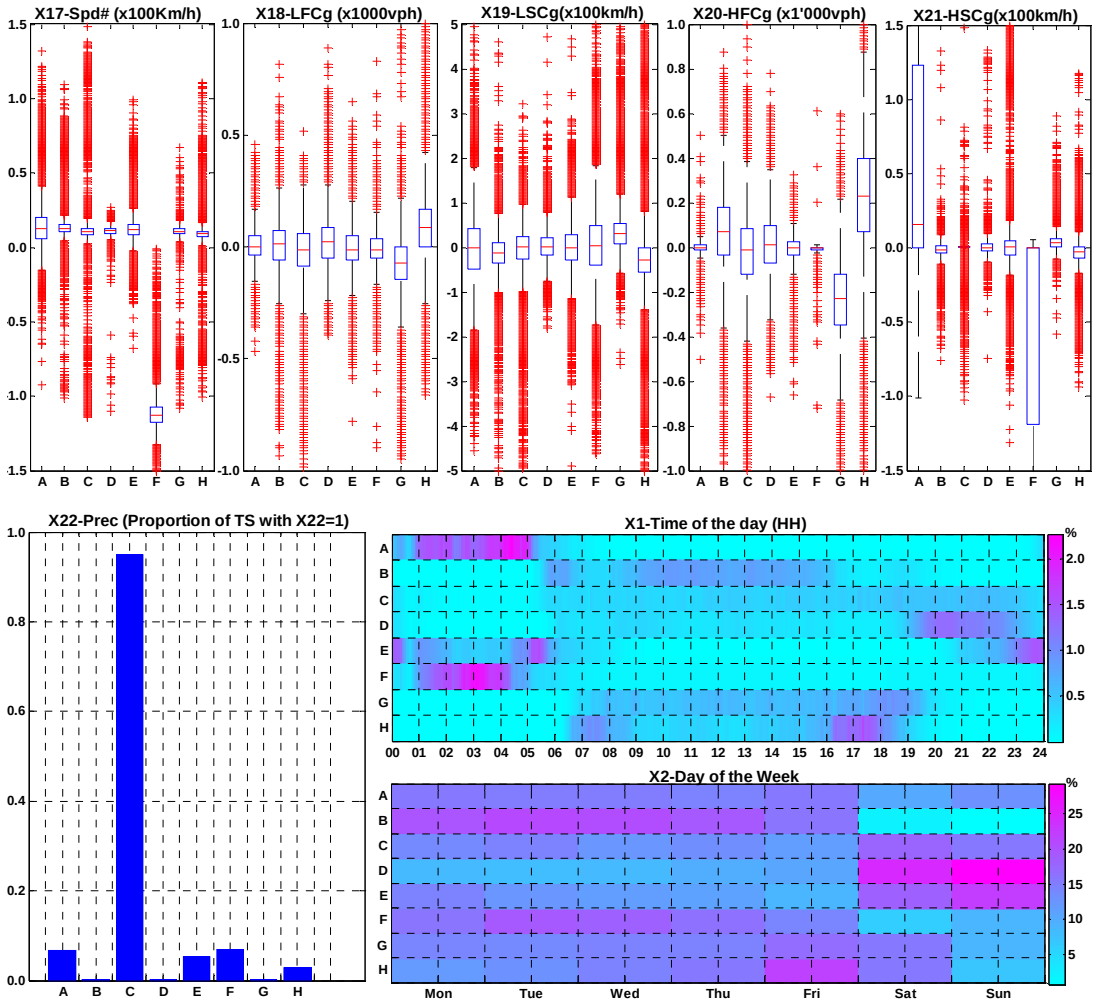

Figure 6-14: Statistics on variables from X17 to X21 and non-traffic variables X1, X2, and X22

Two categorical variables *X1* and *X2* are not used in data sampling process and therefore do not contribute to the formation of Traffic Regimes. However, X1 and X2 also characterizes the difference between Traffic Regimes. The following can be observed from Figure 6-14:

1) Regimes A, E, and F occur mostly during early morning before 6:00AM. Regime E occurs also after 9:00PM. Regimes A and F occur more often on weekday whereas; regime F occurs mostly during weekend.

2) Regime H occurs more often during rush hours. Regime H occurs rarely on Sunday yet very often on Friday.
3) Regime G is more frequent from 8AM to 7:00PM every day except Sunday when regime G is less frequent.
4) Regime B is most frequent from 5:30AM to 6:30AM and from 9:00AM to 4:00PM and from Monday to Friday. During weekends, regime B is rare.
5) Regime D can occur every day yet with high frequency on Sunday. High frequency of regime D's occurrences is also observed from 7:00PM to midnight.
6) Regime C is very similar to regime D by considering *X1* and *X2*.

6.3.4. Examples of Traffic Regimes

To provide a match between daily traffic and the use of traffic regimes, this section presents several examples of traffic regime evolution in one day traffic. Three dates are selected: a normal day (working, non-rain, and uncongested day), a rainy day, a congested day, and a weekend. In Figure 6-15, Figure 6-16, and Figure 6-17, variables *X3* (flow on the right lane) and *X4* (average speed on the right lane) for the whole day are presented and referenced on the left vertical axis. Traffic Regimes are referenced on the right vertical axis. There is also a period in each figure where there is no data due traffic detector reset. During this period, speed and flow are set to zero.

In general, on weekdays, starting from midnight the traffic volume reduces to its lowest level from 1:00AM to 4:00AM before increasing to morning peak at about 8:00AM. Thereafter, the traffic reduces and stays at high volume level before increasing from 3:00PM to afternoon peak at about 5:30PM. After that, traffic volume reduces until midnight.

Figure 6-15 presents traffic evolution for a normal working day with no precipitation (*X2*=2 - Tuesday and *X22*=0) and uncongested traffic. As it is non-rainy day the traffic did not come to regime C.

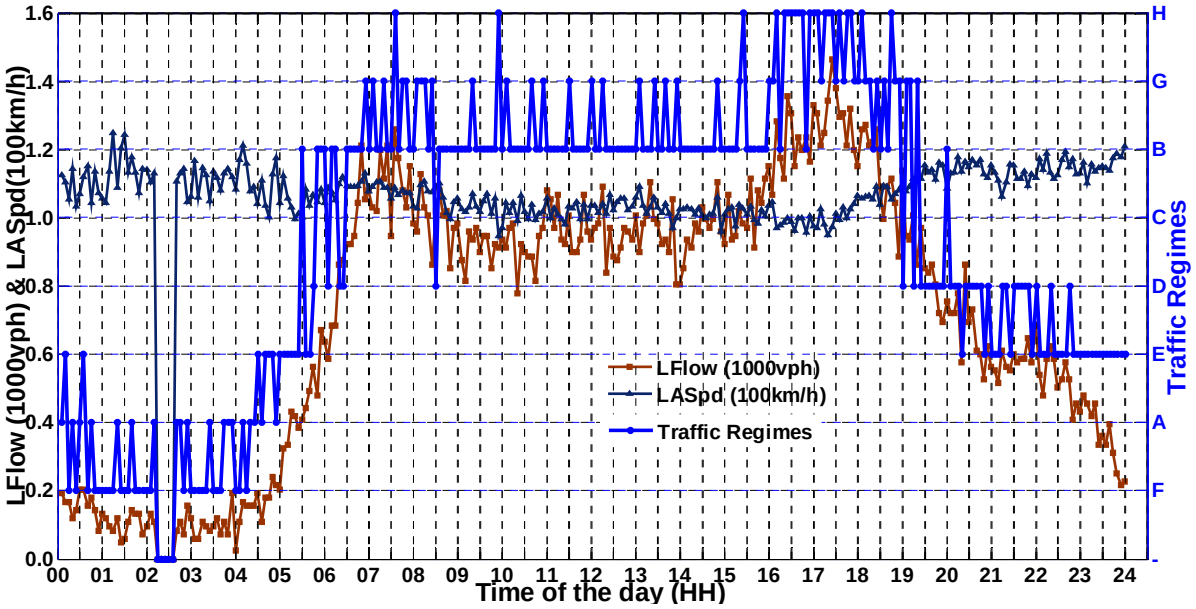

Figure 6-15: Traffic Regime evolution on Tuesday, April 25, 2006

Figure 6-16 presents traffic evolution for a weekend with some rainy period on Saturday, June 17, 2006. In early morning (from 1:00AM to 5:00AM), the traffic on Saturday is not as low as on working days. However, the traffic increases gradually to high flow at around 10:00AM. In the afternoon and evening, there are periods where the precipitation is positive: from 4:00PM to 4:30PM and some periods within 7:00PM to 8:00PM. During these periods, the traffic comes to regime C.

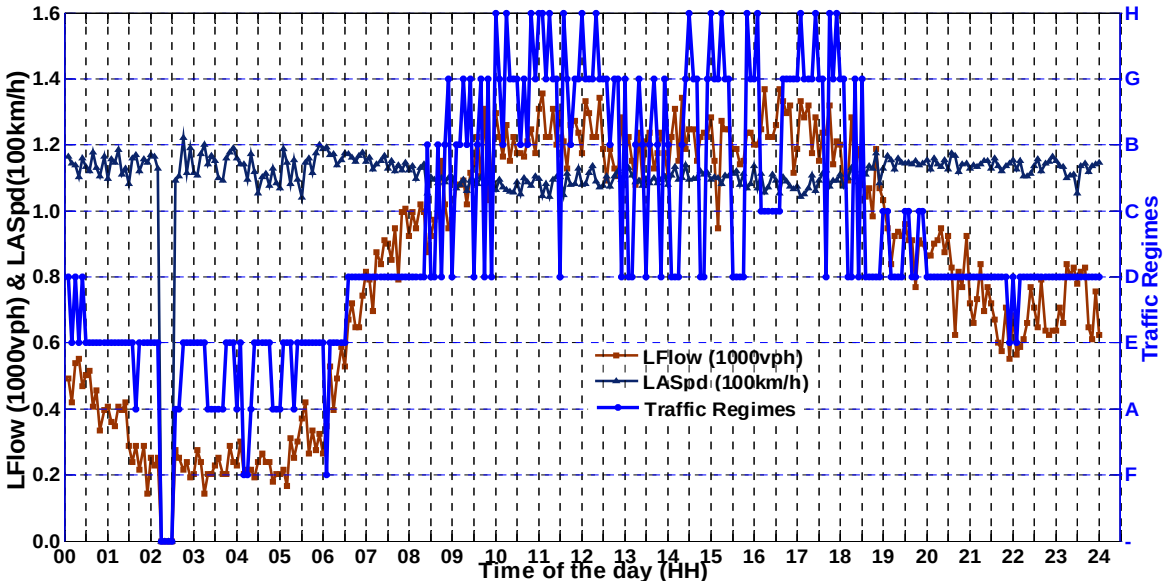

Figure 6-16: Traffic Regime evolution on Saturday, June 17, 2006

Figure 6-17 presents a working day traffic without precipitation and with congestion during morning peak (at around 7:00AM to 7:30AM). The congestion is indicated by the speed drop to below 40km/h.

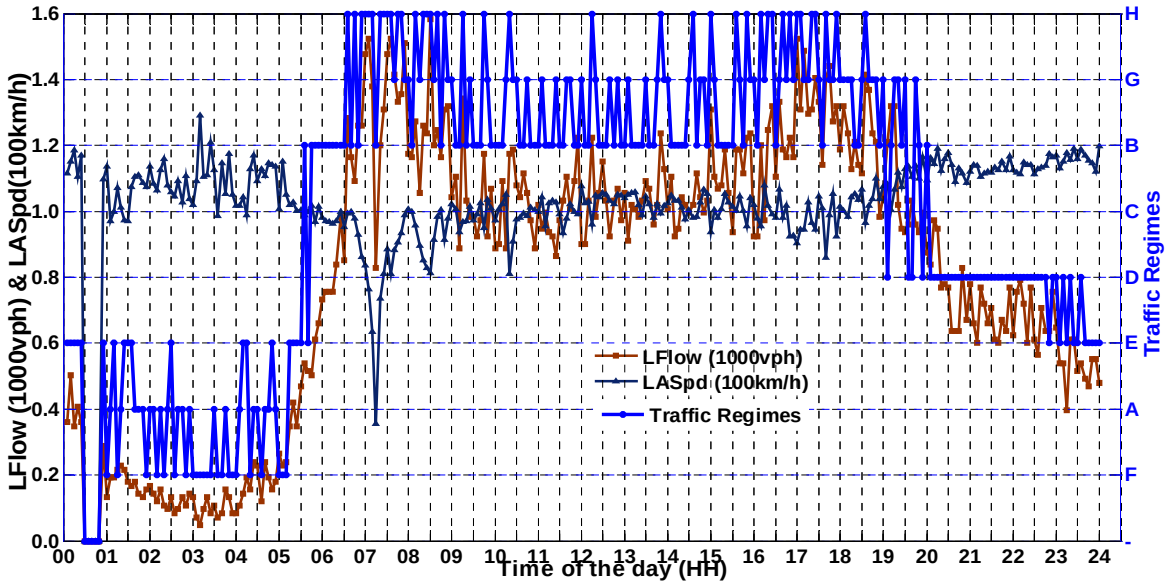

Figure 6-17: Traffic Regime evolution on Thursday, August 21, 2003

6.3.5. Summary

Three examples presented in Figure 6-15, Figure 6-16, and Figure 6-17 enforce the observations from 1) to 6) in section 6.3.3. Therefore, the following predeterminations can be conducted:

1) Regime B usually occurs when the traffic is increasing (*X18* and *X22* are usually positive). Therefore, it represents increasing traffic and it usually precedes high flow and congested traffic regime H.
2) Regime G usually occurs when the traffic is decreasing (*X18* and *X22* are usually negative). Therefore, it represents decreasing traffic and usually follows regime H.
3) When the traffic is already at regime B and does not significantly increase, it will come to regime G.

The proofs in the next section would conclude these predeterminations.

6.4. TS Transitions

6.4.1. Introduction

TS transitions are the movements of the TS from one TR to another TR. The typical transitions of TR for one day via three examples presented in Figure 6-15, Figure 6-16, and Figure 6-17 is that the traffic starts from midnight in free flow regimes F, A, and E then stays in high density regime B, G, and H during daytime before coming back to regimes E, A, and F in the evening. Regimes C and D play intermediate role when the traffic changes from free flow to high density regimes.

This section aims to provide more insight about TS transitions for NTS and eventually for PTS. Because six PTS are used for each crash, the pattern of six TR transitions is considered here.

6.4.2. NTS Patterns

6.4.2.1. NTS Transitions

An NTS pattern is a TR transition pattern of NTS, i.e. the traffic conditions where there are no crashes recorded. Figure 6-18*a* and Figure 6-18*b* presents the proportions of NTS transitions among TR. In Figure 6-18*a* and Figure 6-18*b*, Traffic Regimes are grouped in three groups: group of low traffic (regimes A, E, and F), group of intermediate traffic (regimes D and C), and group of high traffic (regimes B, G, and H).

Consider two consecutive traffic situations TS_{t-1} and TS_t whose traffic regimes are α and β, respectively. Two regimes α and β can be the same or different and are among eight regimes from A to H. Call $trans(\alpha, \beta)$ the number of transitions from regime α to β in NTS data. Call $P_r(\alpha, \beta)$ is the proportion of NTS transitions from regime α to regime β compared to all NTS transitions from regime α to all regimes. $P_r(\alpha, \beta)$ is calculated as follows:

$$P_r(\alpha, \beta) = \frac{trans(\alpha, \beta)}{\sum_{\forall \beta} trans(\alpha, \beta)}$$

Similarly, $P_c(\alpha, \beta)$ is the proportion of NTS transitions from regime α to regime β compared to all NTS transitions from all regimes to regime β. $P_c(\alpha, \beta)$ is calculated as follows:

$$P_c(\alpha,\beta) = \frac{trans(\alpha,\beta)}{\sum_{\forall\alpha} trans(\alpha,\beta)}$$

Each cell in rows in Figure 6-18*a* represents $P_r(\alpha, \beta)$. For instance, as shown in Figure 6-18, more than 70% of transitions from regime D are destined for regime D, i.e. traffic situations remain in regime D and $\alpha = \beta$ = D. Each cell in columns in Figure 6-18*b* represents $P_c(\alpha, \beta)$. Among NTS transitions from (and to) regimes F, E, D, C, and B, more than 50% remain in (or originate) the same regimes.

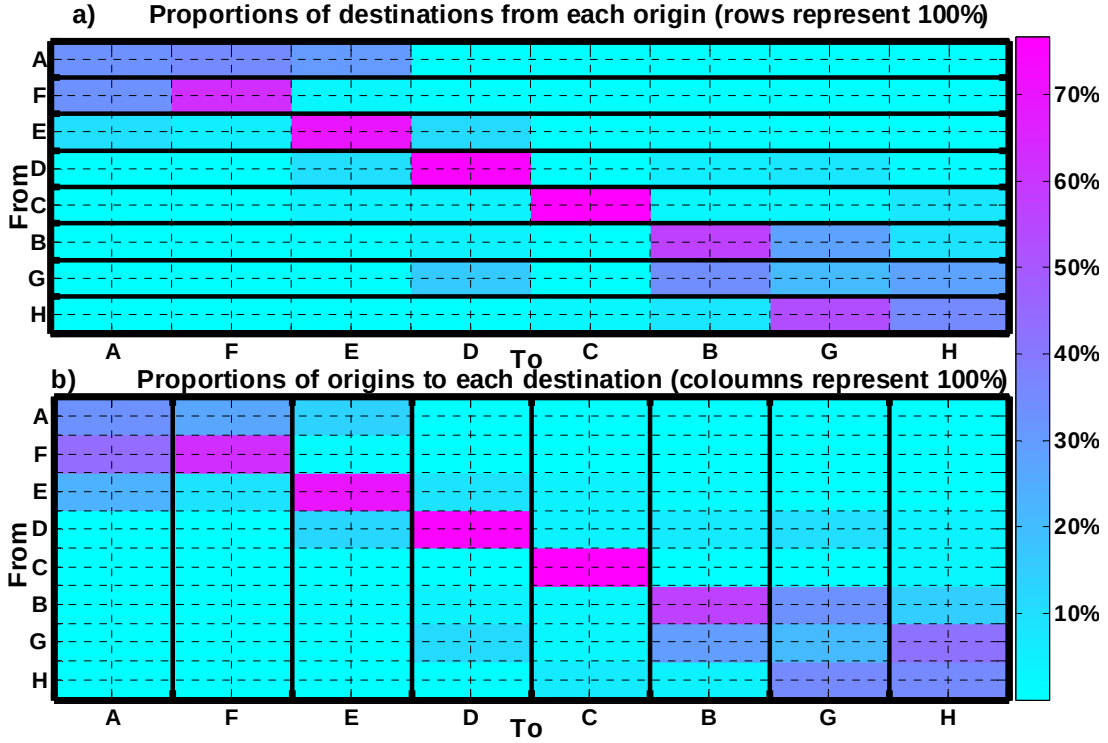

Figure 6-18: Proportions of NTS transitions from origins to destinations

Regime G is the most unstable regime because once the traffic is under regime G, it remains under regime G only 21% of the time. From regime G, about 34% of NTS transitions have regime B as destination and about 28% have regime H as destination. Among NTS transitions to regime G, transitions from H and B contribute 35% and 33%, respectively.

The high flow or congested regime H is also unstable with only about 35% of NTS transitions from regime H to regime H. Regime G exchanges the most with regime H: 53% of transitions from H is to G and 43% of transition to H is from G. Here high percentage of transitions from G to H among the transitions to H means that there is a fluctuation between G and H.

Regime B is rather stable with about 57% of NTS transitions from B remain in B. From B, about 28% of transitions are to G and 8% are to H. To B, about 29% are from G and 5% are from H. It means that the percentage of transitions from H to B is low.

The direct transitions between regimes B, G, and H and regimes A, E, and F are rare. The percentages of transitions between regimes B, G, and H or regimes A, E, and F with regimes C and D are low because regimes C and D are stable.

6.4.2.2. NTS Pattern Statistics

A TS pattern represents 30' evolution of TS under six traffic regimes. There are eight TR candidates (from A to H) for each of six traffic regimes. Therefore, there are totally 6^8= 1'679'616 possible patterns. There are transitions that can never occur such as from A to H to A. For this reason, the total number of NTS transition patterns during the whole study period is 11'782 among 1'050'372 available patterns (there must be at least 30' consecutive data to produce a pattern). There are patterns that repeat more than other patterns.

The pattern DDDDDD, i.e. the traffic remain under regime D for 30 minutes, repeats the most at 86'624 times, i.e. 8.25% of available patterns. After that, patterns EEEEEE, FFFFFF, BBBBBB, and CCCCCC contribute 4.30%, 2.40%, 2.28%, and 1.54% of available patterns. This result conforms to high percentages of transitions from and to one regime for regimes B, C, D, E, and F.

6.4.3. PTS Patterns

Figure 6-19 presents the evolution of traffic under traffic regimes before crashes used in this study. Each cell column represents a crash, each cell row represents the moment of the PTS before crashes. Each cell is a PTS and the color within each cell represents the traffic regime of that PTS. The top row represents the frequencies of corresponding patterns.

The PTS pattern HHHHHH repeats the most - eight times – which means that even the traffic is stable under regime H, the crash risk is still high. There are 64 crashes (i.e. 53.3%) for which the traffic comes only to regimes G and H such as the pattern 23, 25, 27-29, 36-38, 41, 46, 47, 56, 57, 58-64, and 66-72.

PTS patterns are also observed within NTS data. Figure 6-20 presents the repetition frequencies of PTS patterns and the pre-crash rate which is the ratio between the frequencies of PTS patterns in pre-crash data and PTS pattern in NTS data. The pattern indices in Figure 6-20 are the same to pattern indices in Figure 6-19. The PTS patterns BBBBBB and CCCCCC are the most popular in NTS data. After that, the frequencies of PTS patterns GBBBBB and HHHHHH are also high. It means that the chance for the traffic to end up with a crash after these patterns is not high.

There are patterns that can lead to higher pre-crash rate such as HGBCHH (one of four cases ended up with a crash), HHHBHB (two of ten cases ended up with a crash), CGHCHC (one of seven cases ended up with a crash), and GDDGHB (one of eight cases ended up with a crash).

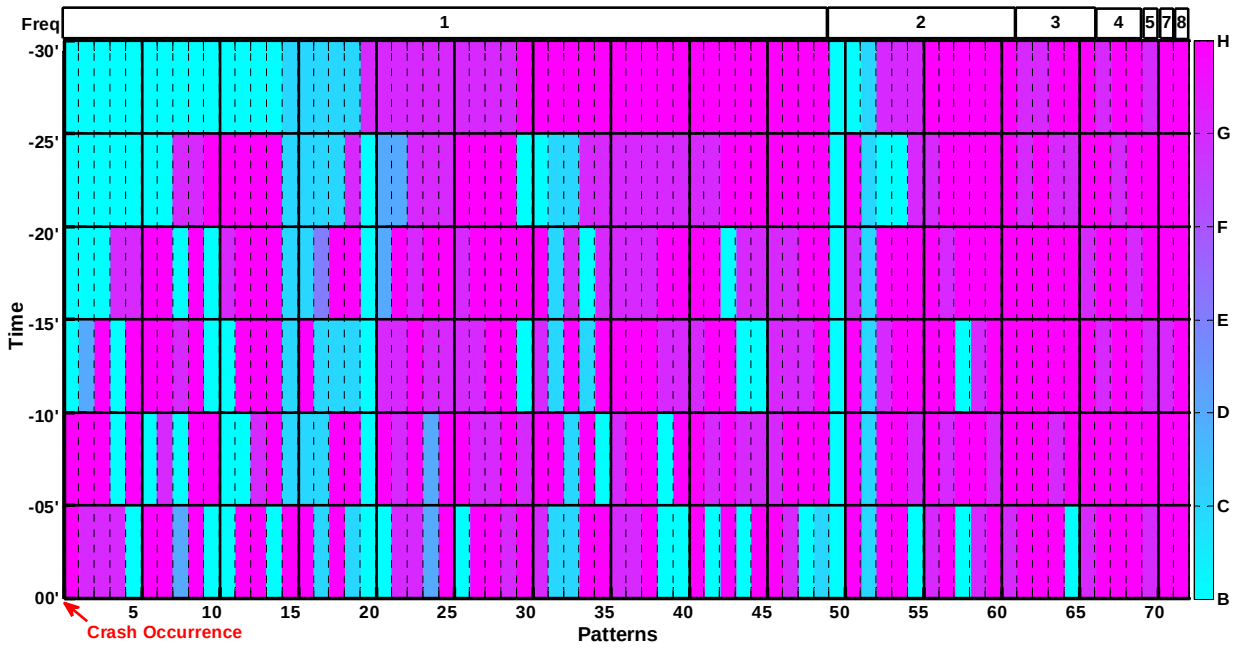

Figure 6-19: PTS Pattern repetitions

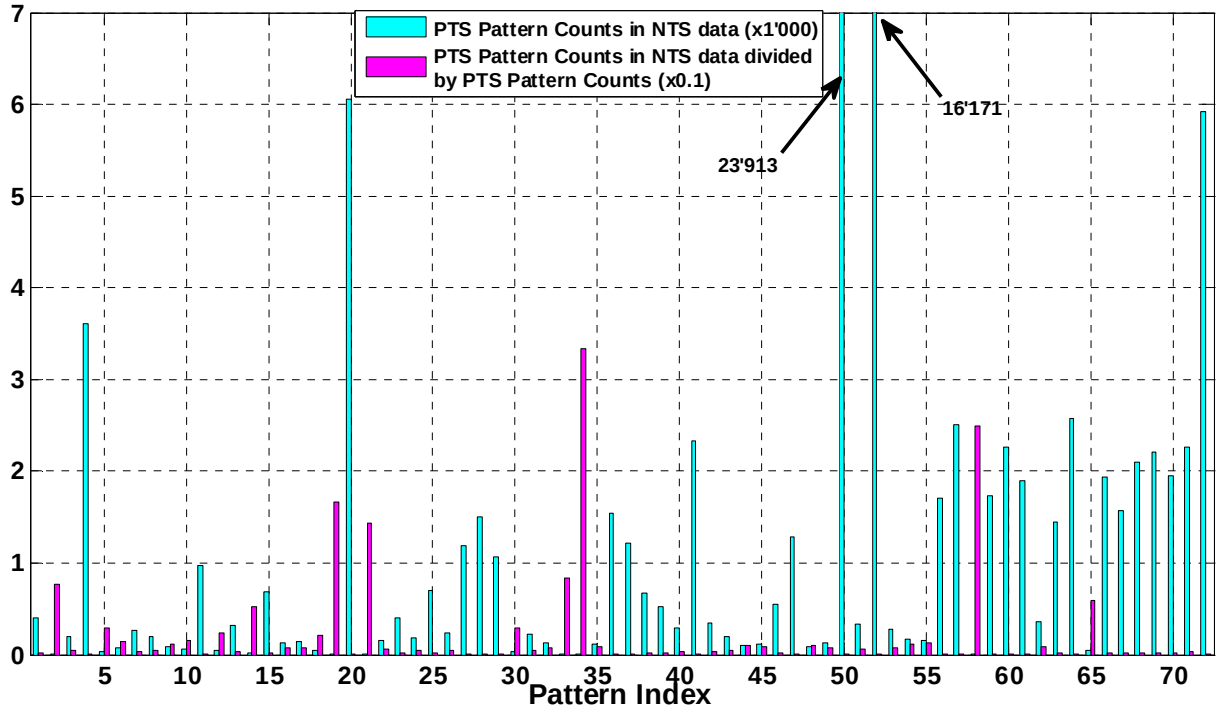

Figure 6-20: PTS patterns observed in NTS data

6.5. Summary

This chapter presents the decisions to sample NTS that are relevant to PTS and then concentrates on uncovering the characteristics of eight traffic regimes. The fundamental Speed/Flow diagrams on two lanes cannot properly characterize each of traffic regimes as only four variables can be presented and different regimes overlap in the diagrams.

Three groups of traffic regimes are observed: group of low traffic, group of medium traffic and group of high traffic. The typical one-weekday traffic evolution starts from midnight with the first group, then changes to the second group at around 5:00 to 6:00 AM before reaching the third group and remain in the third group during day time. The traffic in the evening changes to the second group and then the first group until midnight. Weekend traffic evolution is slightly different with the traffic coming at the third groups at some time near noon.

Within the first group, traffic under regime E is denser than under regimes A and F. Regime F represents special traffic conditions when there is only traffic on one lane.

Regimes C and D are classified into the second group and are similar in traffic characteristics. The only difference between C and D is that regime C occurs under rainy conditions whereas regime D occurs under normal weather conditions.

In the third group, regime H represents high flow or congested traffic conditions whereas regimes B and G represents traffic conditions leading to and following regime H, respectively. There is high fluctuation between regimes H and G and between regimes G and B.

Regarding to PTS evolution, it is observed that before almost of crashes, the traffic falls into regime H at least once (Regime H represents high flow or congested conditions).

In the next chapter, the development of risk identification models under regimes B, C, G, and H will be presented. No model is developed under regimes A, D, E, and F and whenever, a new traffic situation is classified into one of these regimes, it will be automatically declared as NTS.

Chapter 7 Real-time Risk Identification

This chapter dedicates to the identification of real-time traffic crash risks. The issues addressed here include selecting the most appropriate method for developing models capable of identifying traffic risk, improving the performance of developed models, interpreting results.

7.1. Overview

Relating to RIM development, Table 7-1 summarizes the results obtained from previous chapters.

Table 7-1: Results from previous chapters

Index	Parameter	Value
1	Number of Traffic Regimes considered	8, namely from A to H.
2	Traffic regimes with high Risk Chance	B, C, G, and H.
3	Number of variables used	22 (see Table 5-1 in Chapter 5)

According Figure 3-1, after defining NTS and PTS and sampling NTS to obtain Traffic regimes, RIM can be developed under each traffic regime. As presented in 3.4.2, the technique employed for RIM development should be a supervised learning technique. This is because the classes of outputs are known: NTS and PTS. Besides, the selected supervised learning method needs to match the following criteria:

1) Accept categorical variables such as time of the day and day of the week in inputs.
2) Should be resistant to the imbalance between NTS and PTS.
3) Should facilitate the interpretation of results

In Chapter 6, NTS and PTS are matched under Traffic Regimes such that NTS and PTS under each regime are the most comparable to each other. As Risk Chance under regimes A, D, E, and F is zero or almost zero, there is no need to develop RIM under such regimes. Under each of four regimes B, C, G, and H, a RIM is developing to differentiate between NTS and PTS. Given a new traffic situation, the role of RIM is to classify that traffic situation into one of two classes: NTS or PTS. It means that the traffic situation must have occurred for RIM to make the classification. It also means that RIM do not make any prediction yet just identify what has happened.

Based RIM, real-time prediction model is also established to provide short-term prediction

In the next sections, the selection of a supervised learning method is presented with respect to three criteria 1), 2), and 3) above. Thereafter, results of the application of the selected method are discussed. The result interpretation is also discussed with regard to the causality of crashes. Consequently, short-term crash risk prediction is discussed with applicability of the developed framework in reality.

7.2. Supervised Learning Method

7.2.1. Overview

Many methods can do supervised learning, especially when the outputs are binary such as NTS and PTS. The discussion in this section will focus on addressing the following questions:

1) As there are two approaches of supervised learning methods, i.e. regression and classification, which approach is more appropriate for the current study?
2) Among the methods of the selected approach (that can be regression or classification), how to the most appropriate method?

The answer to these questions can be found based on the criteria 1), 2), and 3) in section 7.1 and by testing different methods.

7.2.1.1. Data Settings

To make the results comparable, a single data set is used. Here, data under traffic regime H is used because PTS population and Risk Chance under regime H are higher than under any other regime. Risk Chance under regime H is 0.0043 and *IMRO* is 98'883:420≈235:1.

Let X be the matrix of TS (including NTS and PTS) whose rows are TS or observations and columns are variables. The number of variables M is 22 (M=22) according to Table 5-1. Therefore, $X=[X_1\ X_2\ X_3\ \dots\ X_{22}]$ where X_j ($1\leq j\leq M$) represents j-th column of matrix X. The number of observations P is 98'883+420=99'303 (P=99'303). Therefore, $X=[x_1; x_2; \dots; x_P]$ where x_i represents the i-th row of matrix X and $x_i = [x_{i1}\ x_{i2}\ \dots\ x_{iM}]$ where x_{ij} ($1\leq j\leq M$) represents j-th element of row vector x_i. Similarly, column vector $Y=[y_1; y_2; \dots; y_P]$ where y_i represents the i-th response of the original function f for the i-th observation x_i: $y_i = f(x_i)$. A function g is an estimated version of function f. The outputs $Y'=g(X)$ are the estimations of Y given by g. It is function g that we attempts to obtain and g should be as close to f as possible.

7.2.1.2. Supervised Learning Candidates

Among three criteria 1), 2), and 3) in section 7.1, the first and the third criteria can be used to limit the number of candidates. The candidates in the short list include Logistic Regression – LR (Pampel, 2000), Classification and Regression Trees – CART (Breiman et al., 1984), and Random Forests – RF (Breiman, 2001).

Logistic Regression (Pampel, 2000) is used for prediction of the probability of occurrence of an event by fitting data to a logit function. Logistic Regression is a generalized linear model used for binomial regression with the use of several predictor variables that may be either numerical or categorical. Interpreting results of logistic regression is facilitated thanks to its regression coefficients.

CART (Breiman et al., 1984) is a non-parametric learning technique, using the methodology of tree building by recursively partitioning data in two smaller data set. CART classifies objects or predicts outcomes by selecting from a large number of variables the most important one in determining the

outcome variable. Variables of inputs can be either numerical or categorical. Result interpretation is supported by the importance of variable on the built tree, which is estimated by the weighted impurity reduction obtained from all the nodes of the tree where that variable is decisive. Further explanation about CART can be found in section Appendix C.

Random Forest (Breiman, 2001) is an ensemble learning method that generates many classification and regression trees, trains the trees and aggregates their results. As RF is built based on trees, categorical or numerical variables are accepted. RF also offers the estimation on the important of variables used, which would facilitate the results interpretation.

Several other popular methods such as Neural Networks, Support Vector Machines, etc. are not in the short list because they are based on numerical calculations and interpreting their results is not an easy task.

With three selected methods (LT, CART, and RF) that can support categorical variables and facilitate result interpretation, the next task is to determine the one performing the best with data in the current study. The remaining of section 7.2 will discuss about that.

7.2.2. Classification Approach

7.2.2.1. Problem Statement

Given X and Y, where Y contains labels *NTS* and *PTS* corresponding to x_i. Find function g that estimates function f: $f(X) = Y$. This is a classification problem as the response Y is categorical.

Rows in X and Y are randomly divided into two sets: training set and test set. The training set called *TrainingSet* contains indices of rows in X and Y used for training purpose. The test set called *TestSet* contained indices of rows in X and Y used for testing purpose. Here, 70% of rows (i.e. 69'438 rows including 69'218 NTS and 220 PTS) in X are used for training, 30% of data (i.e. 29'760 rows including 29'665 NTS and 95 PTS) in X are used for testing.

7.2.2.2. Result Comparison

Here, LR, CART, and RF are used as classification methods whose outputs are two classes NTS and PTS. Logistic Regression – LR is used as a classification method by maintaining an output threshold of 0.5. *TrainingSet* is used to develop the models. Once the models are developed, *TestSet* is input to test the prediction capacity of each model. Results of each model are summarized in Table 7-2.

The performance of each model related to each data set is given in detail in percentages of NTS and PTS correctly classified. This is important as NTS and PTS populations are imbalanced. If the performance of models with NTS and PTS is represented by the percentage of both NTS and PTS correctly classified, the performance on the minor class (i.e. PTS class) might be neglected. The performance of RF with NTS and PTS in *TestSet* is one example: percentage of PTS correctly classified is low (9.4%) yet the combined percentage of NTS and PTS correctly identified is 99.71% (9 PTS and 29'665 NTS correctly classified on the total of 29'760 data in *TestSet*).

Here, it is clear that the performance of models with PTS is really low, especially with PTS in *TestSet*. As one of the objectives in the current research is to identify traffic crash risk, such low performance of classification approaches related to PTS identification makes them undesirable to be selected.

Table 7-2: Performance (percentage of NTS or PTS correctly classified) of RL classification approach

Method	*TrainingSet*		*TestSet*	
	NTS	*PTS*	*NTS*	*PTS*
LR	100.00	0.50	100.00	0.00
CART	99.97	68.64	99.71	10.53
RF	100.00	100.00	100.00	9.40

7.2.3. Regression Approach

7.2.3.1. Problem Statement

Given X and Y, where Y contains numerical outputs y_i corresponding to x_i. y_i represents the probability for the i-th observation (i.e. x_i) to become pre-crash and called *pre-crash probability*. Pre-crash probability ranges from 0.000 to 1.000 with 0.000 representing pre-crash probability of an NTS and 1.000 representing that probability of PTS. There are traffic situations having pre-crash probability between 0.000 and 1.000. The traffic situations need to be classified as PTS or NTS. Therefore, a *pre-crash threshold* is defined to be a pre-crash probability value between 0.000 and 1.000 such that:

- A traffic situation having pre-crash probability greater than or equal to the pre-crash threshold will be classified as PTS.
- A traffic situation having pre-crash probability smaller than the pre-crash threshold will be classified as NTS.

Thereby, the problem in this case includes two tasks:

1) Developing a model g that estimates function f: $f(X) = Y$. This is a regression problem as the response Y is numerical.
2) Defining the pre-crash threshold for classifying a TS into NTS or PTS.

To solve the tasks, rows in X and Y are randomly divided into three sets: training, calibration and validation set. The training set called *TrainingSet* contains indices of rows in X and Y used for training purpose. The calibration set called *CalirationSet* contains indices of rows in X and Y used for calibrating pre-crash threshold. The valitation set called *ValidationSet* contained indices of rows in X and Y used for testing purpose. Here, the ratio between *TrainingSet, CalirationSet,* and *ValidationSet* is 6:2:2.

7.2.3.2. Mean Squared Errors

The performance of regression techniques are estimated by the mean squared errors of NTS and PTS in *TrainingSet* and *CalirationSet*. *TrainingSet* and *CalirationSet* are each divided into two parts for NTS and for PTS resulting four data sets: *TrainingSetPTS, TrainingSetNTS, CalirationSetPTS, and CalirationSetNTS.* The performance of the function g is summarized in four error terms: *TrPE, TrNE, TePE,* and *TeNE* corresponding to errors of g with four data sets *TrainingSetPTS, TrainingSetNTS, CalirationSetPTS, and CalirationSetNTS*, respectively. The four error terms are calculated according to Equation 9(a), (b), (c), and (d).

Equation 9: Four error terms

(a) $\mathrm{TrPE} = \frac{1}{\mathrm{TrainingSetPTS}} \sum_{i \in \mathrm{TrainingSetPTS}} \left(y_i - y'_i\right)^2$

(b) $\mathrm{TrNE} = \frac{1}{\mathrm{TrainingSetNTS}} \sum_{i \in \mathrm{TrainingSetNTS}} \left(y_i - y'_i\right)^2$

(c) $\mathrm{TePE} = \frac{1}{\mathrm{CalirationSetPTS}} \sum_{i \in \mathrm{CalirationSetPTS}} \left(y_i - y'_i\right)^2$

(d) $\mathrm{TeNE} = \frac{1}{\mathrm{CalirationSetNTS}} \sum_{i \in \mathrm{CalirationSetNTS}} \left(y_i - y'_i\right)^2$

Table 7-3 presents the four error terms given by three regression techniques. It can be seen that the imbalance of NTS and PTS data sets influences much to the differences between mean squared errors of NTS and PTS observations estimated by three regressions: the mean squared errors for PTS are much higher than for NTS in both *TrainingSetPTS* and *CalirationSet*. It means that the imbalance in data lowers the pre-crash probability for all TS, either NTS or PTS. Among three regression techniques, LR is influenced the most by the imbalance of data sets with highest mean squared errors for PTS. On the contrary, RF is also influenced by the imbalance of data but with lower mean squared errors for PTS.

Table 7-3: Mean squared errors of regression approaches

Regression technique	*TrPE*	*TrNE*	*TePE*	*TeNE*
LR	0.944	0.65×10^{-4}	0.954	0.76×10^{-4}
CART	0.334	5.63×10^{-4}	0.863	13.22×10^{-4}
RF	0.104	0.21×10^{-4}	0.848	1.78×10^{-4}

7.2.3.3. Pre-crash Threshold

The mean squared errors are highly influenced by the imbalance of data sets. The traditional probability of 0.5 corresponding to the squared error presented in Table 7-3 which works well with balanced data sets is not applicable for the developed models. For this reason, the pre-crash thresholds need to be tuned for the models to gain higher accuracy.

Once the pre-crash threshold for each model is set, any traffic situation having probability returned from the model greater than or equal to the pre-crash threshold is classified as pre-crash; otherwise the traffic situation is classified as non-crash.

In present chapter, the pre-crash threshold for a model is set in two steps:

- Develop the model with *TrainingSet*
- Test the developed model with *CalirationSet*. The threshold is set using *CalirationSet* based on some criteria (presented below). Note that in this chapter, there is no third data set to validate the prediction performance of the model together with the pre-crash threshold.

The pre-crash threshold to be set should satisfy the following criteria:

1) Percentage of PTS in *CalirationSet* correctly classified should be at least 70%. This is to guarantee that the model can detect the risk as one of the objectives of the current study.
2) Percentage of NTS in *CalirationSet* correctly classified should be at least 70%. This is to guarantee that the model is not trivial. There are cases where the criterion 1) is achieved yet in return the percentage of NTS correctly identified is low. Criterion 2) is to avoid these cases.
3) Among all the thresholds satisfying criteria 1) and 2), choose the threshold that gives maximum sum of percentages of NTS and PTS correctly classified.

Figure 7-1 illustrates how to determine the threshold. Horizontal axis represents all potential pre-crash thresholds which are zoomed in the interval from 0.00 to 0.20 as the model has low performance with thresholds than 0.20. Vertical axis represents percentages of NTS and PTS from *TrainingSet* and *CalirationSet* correctly classified and the sum of percentages of PTS and NTS from *CalirationSet*. Two curves representing percentages of PTS and NTS in *TrainingSet* are drawn in grey. Two curves representing percentages of PTS and NTS in *TrainingSet* are drawn in black. All the curves representing percentages do not lie above the horizontal line of 100%. The curve lie beyond the line of 100% represents the sum of percentages of NTS and PTS in *CalirationSet*. Applying criteria 1) and 2), the threshold interval is determined to be from 0.014 to 0.065. Among pre-crash threshold in that interval, the threshold of 0.038 is found to give maximum sum of percentages of NTS and PTS in *CalirationSet* (163.4). Therefore, the pre-crash threshold is set to be 0.038.

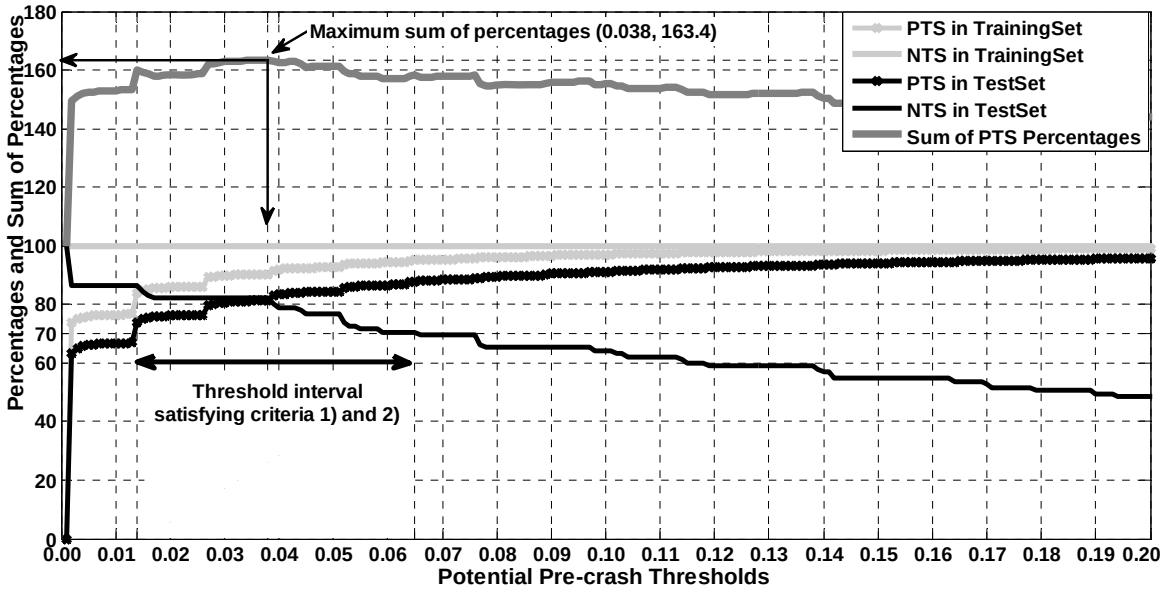

Figure 7-1: Pre-crash threshold determination using RF regression

With the pre-crash threshold of 0.038:

- percentage of PTS in *TrainingSet* correctly classified is 100%
- percentage of NTS in *TrainingSet* correctly classified is 90.19%
- percentage of PTS in *CalirationSet* correctly classified is 82.11%
- percentage of NTS in *CalirationSet* correctly classified is 81.30%

7.2.3.4. Classification Performance

Table 7-4 summarizes the pre-crash thresholds and the performance of three regression techniques: LR; CART; and RF. It is unable to set a threshold for CART as the criterion 1) is not satisfied. CART classifies all TS in *TrainingSet* correctly and classifies NTS in *CalirationSet* with high accuracy. However, CART cannot classify correctly more than 20% of PTS in *CalirationSet*. Therefore, according to criterion 1), no pre-crash threshold is set for CART.

LR has good performance and a pre-crash threshold can be set. However, RF outperforms LR for TS in *TrainingSet* and for PTS in *CalirationSet*. For *ValidationSet* representing new traffic situations as this data set is not used to train models or to define thresholds, RF creates the clear difference with LR in term of both percentages of NTS and PTS correct identified.

Table 7-4: Pre-crash thresholds and performance (%) of regression techniques

Technique	Pre-crash Threshold	*TrainingSet*		*CalirationSet*		*ValidationSet*	
		PTS	NTS	PTS	NTS	PTS	NTS
LR	0.007	68.18	85.68	72.63	85.31	55.43	67.31
CART	-	-	-	-	-	-	-
RF	0.006	100.00	93.59	85.44	76.09	70.35	75.31

Therefore, RF is selected as supervised learning method that will be used for developing risk identification models in the current research.

7.2.4. Summary

Section 7.2 is dedicated to the choice of a supervised machine learning method which is the most appropriate to the current study. Finally, Random Forests Regression method is selected because it can:

- Work with both categorical and numerical variables.
- Facilitate result interpretation with its internal evaluation of importance of variables while developing models (See Appendix C).
- Prove its better performance with the data in the current study in comparison with the performance of other techniques (See section 7.2.3.4).

7.3. TR-based Risk Identification Models

7.3.1. Overview

In this section, RIM are first developed using all 22 variables to attain certain accuracy. Thereafter, the developed models are refined to reduce the number of variables making the models more compact while maintaining the accuracy.

7.3.2. Random Forests

Random Forest (Breiman, 2001) is an ensemble learning method that generates many classification and regression trees (CART), trains the trees and aggregates their results. Successive trees do not depend on earlier trees - each is independently constructed using a bootstrap sample of the data set. Here, the summary of Random Forests is presented. More detailed explanation can be found in Appendix C.

According to Breiman, (2001), the motivation for inventing RF is that CART is unstable together with its moderate accuracy. Maximum trees usually work well with training data but have low performance with test data. Tree pruning can improve CART performance with test data and result in trees with relatively higher accuracy. However, CART is unstable as even a small change in training data could also lead to totally different trees which make tree interpretation become problematic. Therefore, the main idea of RF is to create many maximum trees such that there is no correlation between any pair of trees and then aggregate trees' results. With regards to the current research, Random Forests can do the following:

- Provide classification and regression of traffic situations with high accuracy
- Provide the estimation of variable importance for the variables used to define traffic situations.

More details on Random Forest are presented in Appendix C.

7.3.3. RIM Performance

Applying Random Forests Regression to train TR-based RIM as presented in section 7.2.3, four models are obtained corresponding to four highly risky traffic regimes B, C, G, and H. Under each regime, the developed RIM is used to test six data sets: NTS and PTS for training, calibration, and validation, the results are summarized in Table 7-5. The results for all four regimes are calculated based on data sets clustered and classified into four regimes B, C, G, and H.

Table 7-5: Summary of RIM's results

Traffic Regime	Likelihood Threshold	Training (%)		Calibration (%)		Validation (%)	
		NTS	PTS	NTS	PTS	NTS	PTS
B	0.0045	98.94	100	97.64	80.00	88.77	83.33
C	0.0002	95.64	100	92.01	100.00	78.91	90.00
G	0.0003	93.37	100	89.76	91.89	85.45	87.80
H	0.0060	93.59	100	85.44	76.09	70.35	75.31
All four regimes		95.67	100.00	91.93	84.77	82.83	83.62

Identifying PTS under regime H is more challenging as the percentage of PTS (for validation) correctly identified is the lowest, i.e. 75.31%. Especially, this percentage is achieved only when the percentage of NTS correctly identified is at lowest level – 70.35%, at the limit for a trivial model according to criterion 2) of pre-crash threshold. It means that about 25% out of PTS under regime H cannot be correctly identified and about 30% of NTS are wrongly considered as PTS.

Under the other regimes B, C, and G, percentages of NTS and PTS (for validation) correctly identified are much higher than under regime H. However, the percentages of wrongly identified cases are still high.

Totally, more than 80% out of NTS and PTS under four traffic regimes B, C, G, and H are correctly identified. For PTS, 83.62% is the percentage of all PTS correctly identified. For NTS, 82.83% is just the percentage of NTS correctly identified under four regimes B, C, G, and H. This is because any traffic situation falling into regimes A, D, E, and F is automatically identified as NTS. Therefore, in the whole NTS population, the percentage of NTS correctly identified is much higher (about 91%), i.e. there are about 10% of NTS wrongly identified as PTS.

In reality, if an alarm is raised each time a traffic situation is identified as PTS, there will be the wrong alarms at 10% of the times – which is a huge amount. Together with high percentage of PTS incorrect identified as NTS, this is the biggest issue for the developed model to be applied in reality.

However, traffic situations are considered so far just like independent observations. Fortunately, as presented in section 6.4.3, PTS seem to move together according to certain patterns. Therefore, analyzing the patterns of PTS correctly identified would be the path to increase the performance of the overall model.

7.3.4. RIM Refinement

The objective of RIM refinement is to make the developed models more compact by reducing the number of variable used while simultaneously maintaining the accuracy of the models at similar level as presented in Table 7-5. The refinement is undertaken independently under each regime.

In Random Forests Regression, one of outputs is the internal estimation of importance of the variables used in developing the models (See Appendix C). Figure 7-2 presents the normalized importance of the variables. The normalization is applied to each regime and is simply the division of the original variable importance to the maximum variable importance under the regime. The normalization aims to project all positive importance of variables into the same scale of 0.0 to 1.0.

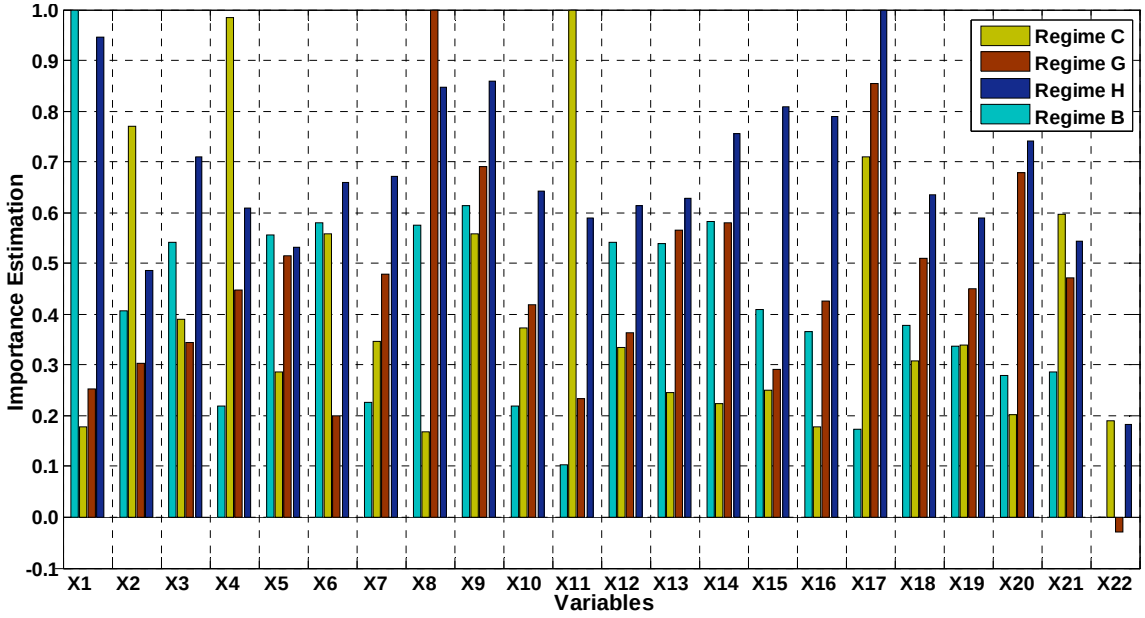

Figure 7-2: Importance of variables under four regimes B, C, G, and H

The importance of a variable is estimated based on a concept called Out-Of-Bag (OOB) error: a variable is more important if it provokes higher OOB error and vice versa (More detail is available in section 8.5.3.b). Therefore, based on variable importance, the following decisions are taken to refine the models:

- Variable *X22(Prec)* under regime G should be removed as its importance is negative which means that the presence of *X22* values increases the OOB error, i.e. the relationship between *X22* and crash occurrences is random.
- Refine the models by iteratively removing variables having low importance (by applying refinement algorithm presented in Appendix C)

Under regime B, variable *X22* has zero importance and can be removed. The performance of revised model under regime B with 21 variables is similar to the performance of the model with 22 variables. Further elimination of any other variable reduces the performance of the model.

Under regime C, the variables having lowest importance estimation are *X8 (LVSpd)*, *X1(TDay)*, *X16 (H%HV)*, *X22 (Prec)*, *X20 (HFCg),* and *X5* (*HAHw*). These variables are in turn removed from variable sets to develop new models (by applying refinement algorithm presented in Appendix C). However, the performance of obtained models is much lower than the model presented in section 7.3.2.

Under regime G, new model with *X22* excluded is developed. As expected, the new model performs better than the model with all 22 variables. The importance of the remaining 21 variables is also re-estimated and presented in Figure 7-3. When more variables are excluded, the developed model performs worse than the model with 21 variables (only *X22* is excluded).

Regime H is similar to regime B with potentially removable variables are *X22 (Prec)*, *X2 (WDay)*, *X5 (LAHw)*, and *X21 (HSCg)*. However, by removing one of those variables, the corresponding models have the reduced performance compared to the performance of the model with 22 variables.

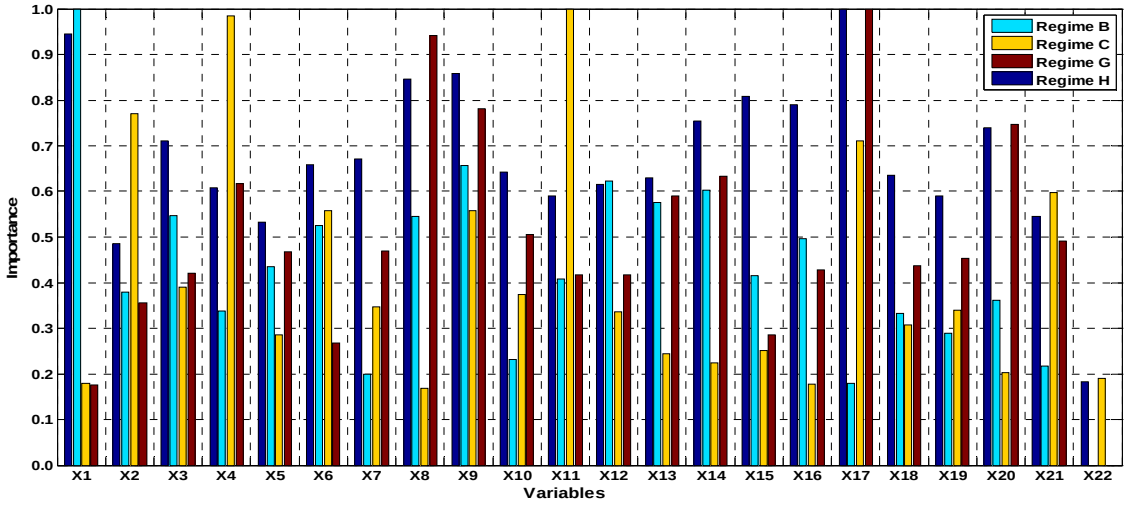

Figure 7-3: Revised importance of variables under four regimes B, C, G, and H

Finally, the refinement by removing variables is only applied to models under regimes B and G with the elimination of variable *X22* (the precipitation).

The performance of refined models is summarized in Table 7-6 where the results are unchanged for RIM under regimes C and H and slightly changed for RIM under regimes B and G. The importance of variables is also re-estimated and presented in Figure 7-3. For regimes B and G, *X22* is not used and therefore has zero importance.

Table 7-6: Summary of revised RIM's results

Traffic Regime	Likelihood Threshold	Training (%)		Calibration (%)		Validation (%)	
		NTS	PTS	NTS	PTS	NTS	PTS
B	0.0045	99.35	100	98.26	80.00	91.53	83.33
C	0.0002	95.64	100	92.01	100.00	78.91	90.00
G	0.0002	92.70	100	88.87	94.59	84.05	87.80
H	0.0060	93.59	100	85.44	76.09	70.35	75.31
All four regimes		95.67	100.00	91.93	84.77	82.83	83.62

7.3.5. Critical Factors

7.3.5.1. Overview

Each variable plays certain role in the performance of TR-based RIM. The role of a variable in a model is quantified by the importance of that variable. Under each regime, several variables such as *X4* and *X11* contribute more than other variables such as *X1*, *X8*, and *X16* as *X4* and *X11* has higher importance than *X1*, *X8*, and *X16*.

According to section 8.5.3.b, the importance of a variable is estimated via the OOB error that the variable produces when its values in OOB data are permuted. That variable becomes more important if the OOB error caused by the permutation of its values becomes high. Variables such as *X22 (Prec)* under regimes B and G have non-positive importance and do not contribute to the classification process.

The presence of variables in Figure 7-3 indicates that each variable is used in revised models because they have certain importance in improving the models' performance. However, the performance of the models depends heavily on some most important variables: the performance of a model reduces dramatically if one of those variables is unused. Those variables are called *Critical Factors – CF*.

It is worth noting that finding Critical Factors in the current study should not be confused with feature selection that attempts to select variables for developing models. Model refinement as presented in section 7.3.4 can be considered as form of feature selection.

Critical Factors are the variables who contribute the most to the differentiation between NTS and PTS. Therefore, results can be interpreted based on Critical Factors to identify the main causality of crashes. Knowing Critical Factors would help to prevent the traffic from developing further and ending up with crashes. Preventive measures can also be designed aiming to change the values of Critical Factors to redirect the traffic in the way that conditions leading to crashes are cleared.

Critical Factors are decided based partly on variable importance estimated by Random Forests. Each variable is also tested in following scenarios:

- Scenario 1: Developing single variable models. The importance of a variable is reflected by the performance of the corresponding model. A variable which is more important than another variable is the one whose corresponding model has higher performance than the model of the other variable.
- Scenario 2: The importance of a variable is reflected by the performance reduction of the model when that variable is removed. A variable is more important than another variable if the performance reduction caused by that variable is higher than the performance reduction cause by the other variable.

7.3.5.2. Results

In both scenarios 1 and 2, the performance used for comparison is the sum of percentages of NTS and PTS correctly identified. All the models are developed using Random Forests and by following the algorithm presented in Appendix C.

A variable that is selected as a Critical Factor if it satisfies the following criteria:

1) It is one of 5 most important factors estimated by Random Forest presented in Figure 7-3.
2) It is one of 10 most important factors estimated according scenario 1.
3) It is one of 10 most important factors estimated according scenario 2.

The first criterion is the most important as the importance of variables is estimated by RIM. The other two criteria are used as additional justification for the variables that have satisfied the first criterion.

Table 7-7 presents the list of critical factors selected under each regime based on the criteria above. Although non-traffic factors are not useable for developing preventive measures, they are necessary for improving the performance of the developed models.

Table 7-7: List of Critical Factors under each regime

Regime	Critical Factors	
	Traffic Factors	Non-traffic factors
B	*X3 (LFlow)*, *X8 (LVSpd)*, *X9 (L%HV)*, *X11 (HASpd)*	*X1 (TDay)*
C	*X4 (LASpd)*, *X11 (HASpd)*, *X17 (Spd#)*, *X21 (HSCg)*	*X2 (WDay)*
G	*X8 (LVSpd)*, *X9 (L%HV)*, *X17 (Spd#)*, *X20 (HFCg)*	
H	*X8 (LVSpd)*, *X9 (L%HV)*, *X17 (Spd#)*, *X20 (HFCg)*	*X1 (TDay)*

According to Table 7-7, the following points are observed:

1) Most of Critical Factors relate to traffic status on the right lane: *X3*, *X4*, *X8*, *X9*, and *X17*.
2) Many Critical Factors are different representations of speed: *X4*, *X8*, *X11*, *X17*, and *X21*.

Regime B represents increasing traffic with *X18* (*LFCg*) and *X20* (*HFCg*) (i.e. the flow change compared to the previous traffic situation on both lanes) being positive most of the time. According to section 6.4.2.1, the traffic from regime B usually comes to regime H (5%), regime G (29%), and for most of the time remains in regime B (57%). As traffic flow is increasing the average speed is decreasing, vehicles try to change from the slow lane to the faster lane. However, this change is not easy as the occupancy on the slow lane is high (about 8.0-9.0%, see *X6 – LOcc* Figure 6-13) due to the high percentage of heavy vehicles on the slow lane (the percentage of heavy vehicles on slow lane under regime B is the highest). Car drivers who would like to change lanes under regime B find themselves blocked on the slow lane. As the traffic is increasing still try to quit this situation and lose the necessary attention to front vehicles. Crashes might have occurred in such scenarios. By examining crashes of which corresponding PTS were just before the crash occurrence, the main cause mentioned in the crash records is *inattention*.

Regime C represents the most fluid traffic among four traffic regimes B, C, G, and H. According to Figure 6-18, the traffic under regime C is stable, i.e. it remains under regime C most of the time (about 70%). Besides, the traffic usually falls into regime C during weekend or public holidays. One of traffic-related Critical Factors under regime C is abnormal (too high or too low) in comparison with other traffic regimes. Yet, one important factor found in crash records is that for crashes related to traffic regime C (i.e. there is at least one PTS falling into regime C), most of drivers are not professional. However, this predetermination needs to be examined more as there are only 9 such crash.

Traffic-related Critical Factors under regimes G and H are the same. However, the trend of traffic under these two regimes is opposite: the traffic on both lanes under regime H is increasing while decreasing under regime G. In both cases, it means that there is a big change of traffic. In addition, traffic transitions between regimes G and H are very frequent (53% of the transitions from H are to G and 43% of transitions to H are from G). This fluctuation can be observed before many crashes (see section 6.4.2.2). In general, traffic flow under both regimes G and H is high and occur mostly during day time and on weekdays. Road users during these periods are experienced commuters who would like to avoid congestion and might use their experience to quit the location before congestion is formed. If many drivers react at the same time, traffic will become fluctuated. In this case, if drivers remain in their current

state (no lane changing, no overtaking, etc.), the fluctuation will be reduced and hence, the chance to avoid crashes will be high.

7.4. Real-Time Motorway Traffic Risk Identification Model (MyTRIM)

7.4.1. Overview

Although TR-based RIM have high performance in identifying NTS and PTS as presented in Table 7-6, that performance is not sufficient for a real-life application as the wrong identification rate remains high. As discussed in section 7.3.3, TR-based RIM just consider traffic situations as independent observations which might be the reason for limiting their performance.

Therefore, the objective of this section is to improve the overall performance based on results obtained from TR-based RI with taking into account the fact that PTS usually come together in pattern before turning into crashes. As such, the evolution of six PTS before each crash (as the pre-crash zone of a crash include 6 PTS, see section 5.4) is examined. This is done in model called *Real-Time Motorway Traffic Risk Identification Model – MyTRIM*. The target of MyTRIM is, one the one hand, to identify the fact that crashes will occur shortly and, on the other hand, to reduce the wrong identification rate of crash occurrences.

Here, we distinguish two concepts in a real-time framework: *warning* and *alarm*. A warning is raised when the crash risk is identified and is outputs of TR-based RIM. Warnings can be given after every time a traffic situation is identified as PTS. An alarm is raised when a crash is going to occur. If a warning is an instant signal and dependent on one traffic situation, an alarm should the results of several consecutive traffic situations.

A warning can fall into one of two states: correct and false. A correct warning is raised when a PTS is correctly identified as PTS. A false warning is raised when an NTS is incorrectly identified as PTS. Besides, the term “missed warning” is also used to indicate the fact that a PTS is incorrectly identified as NTS (the PTS was missed).

Similarly, an alarm can fall into one of two states: correct and false. A correct alarm is raised before a crash occurrence. A false alarm is raised when there is no crash occurring after that. The term “missed alarm” is used to indicate the fact that there is no alarm before a crash occurrence.

In this section, the relationship between warnings and alarms is analyzed. As the percentage of false warnings, correct warnings and missed warnings are known (that is the percentages of NTS identified as PTS, of PTS correctly identified, and of PTS identified as NTS, respectively), analyses are undertaken to find the pattern of warnings that can trigger an alarm such that the percentage of correct alarms is maximized whereas the percentages of missed alarms and of false alarms are minimized.

For 30 minutes before crashes, six TS of 5-minute intervals are classified into PTS and NTS. To identify crash occurrences, MyTRIM memorizes historical risk status called *length of risk memory* - L_{rm}. However, questions can be posed as presented in Figure 7-4.a) and Figure 7-4.b).

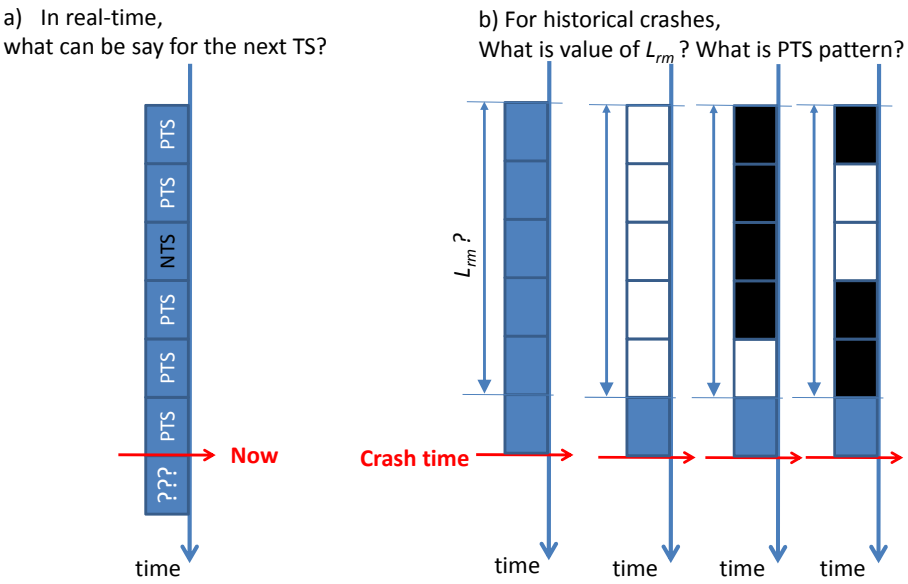

Figure 7-4: Historical risk status

Figure 7-4.a) presents the need to predict the status of the next traffic situations given the status of several last traffic situations identified by TR-based RIM. If there is a way to identify the status of the next traffic situation with high accuracy, counter measures can be implemented to change traffic situations with risky status. To verify that idea, historical crashes need to be analyzed.

Figure 7-4.b) presents several examples where traffic risk status is identified for the first five traffic situations before crashes. If crash occurrences cannot be identified before the sixth traffic situation, there will be no time left to avoid the crashes. Moreover, Figure 7-4.b) also presents the need to determine the value of L_{rm} and the pattern of risk status that can indicate crash occurrences in the next traffic situation.

The pattern of risk status is decided to be consecutive traffic situations identified as PTS. This decision is to restrain the false percentage of alarms - which is the motivation for developing MyTRIM. Depending on the value of L_{rm}, this pattern can be PTS, PTS-PTS, PTS-PTS-PTS, PTS-PTS-PTS-PTS, or PTS-PTS-PTS-PTS-PTS which correspond to value of L_{rm} ranging from 1 to 5.

7.4.2. False Alarm & Missed Alarm

7.4.2.1. Definition

The proportion of false alarms among all non-crash traffic conditions is called *false alarm rate*. As false alarms are wrong decision, false alarm rate should be minimized. Because false alarm is only related to non-crash conditions, NTS data are exclusively used to analyze the false alarm rate.

The proportions of missed alarms and correct alarms among all crashes are called *missed alarm rate* and *correct alarm rate*, respectively. As missed alarms are wrong decisions, the missed alarm rate should be minimized. To calculate the missed alarm and true alarm rates, only PTS are used.

7.4.2.2. False Alarm

The procedure to test the false alarm rate includes:

1) Select all non-crash data, i.e. NTS.
2) Classify each NTS into one of Traffic Regimes.
3) Classify each NTS into one of two classes: NTS or PTS.
4) Check false alarm rate with different lengths of risk memory, L_{rm}=1, 2, 3, 4, and 5.

In step 3), if the NTS belongs to one of regimes A, D, E, or F, it is automatically identified as NTS. If the NTS belongs to one of regimes B, C, G, or H, the corresponding RIM is used to classify the NTS into NTS or PTS. Finally, after step 3), the NTS is classified into NTS or PTS class.

In step 4), each NTS is regarded as the current traffic situation – cTS. L_{rm} values are counted from cTS to the NTS before cTS and so on. As the NTS counted in L_{rm} are consecutive in term of time, any missing NTS among L_{rm} will stop processing that NTS pattern.

Figure 7-5 presents the false alarm frequencies according to different lengths of risk memory, L_{rm}. If the alarm is raised after each warning (i.e. L_{rm}=1), the alarm frequency will be high, resulting high false alarm rate. The false alarm rate reduces quickly when the length of risk memory increases. To obtain the low false alarm rate, high value of L_{rm} is recommended.

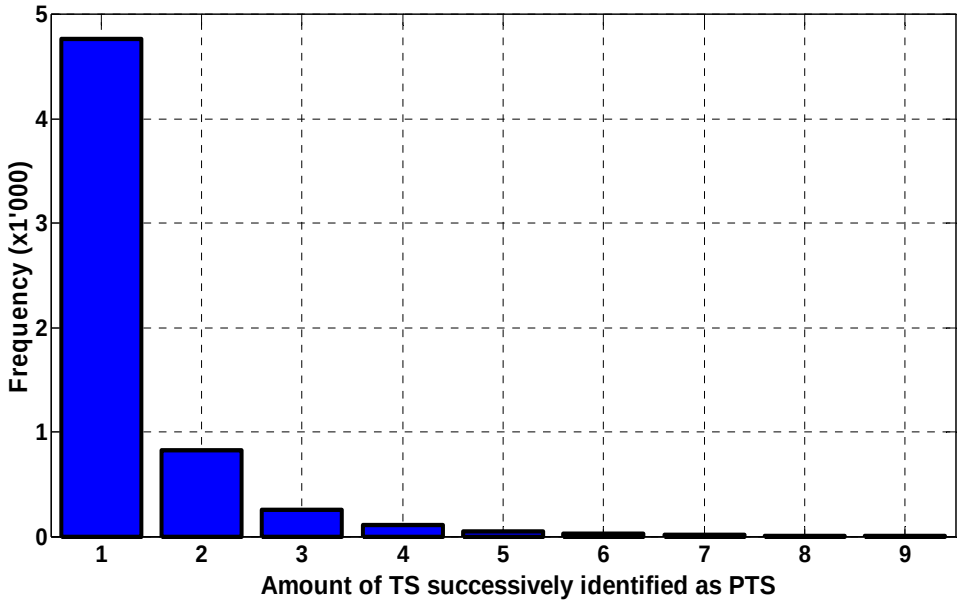

Figure 7-5: False alarm frequencies as function of L_{rm}

To prevent crash risks, the maximum length of risk memory is 5 as the risks are assumed to be present as early as at 30 minutes before crashes.

7.4.2.3. Missed Alarm

The procedure to test the missed alarm rate includes:

1) Select all pre-crash data, i.e. PTS
2) Classify each PTS into one of Traffic Regimes.
3) Classify each PTS into one of two classes: NTS or PTS.
4) Check missed alarm rate with different length of risk memory, L_{rm}=1, 2, 3, 4, and 5.

Steps from 1) to 3) are similar to the procedure applied to test false alarm rate presented in section False Alarm. In step 4), a missed alarm for a crash is counted if there are not enough L_{rm} consecutive PTS identified as PTS in the period from 5 to 30 minutes before the crash. It is worth reminding that all PTS are extracted in pre-crash zones, i.e. 30 minutes before crashes.

Figure 7-6 presents the PTS identification before all crashes. Each vertical line represents the risk evolution identified by RIM for each crash from 30 minutes before crashes to the crash occurrences. Black cells represent PTS incorrectly identified (i.e. identified as NTS). White cells represent PTS correctly identified. For example, for the first crash, all six PTS are correctly identified by RIM. For crashes 60 and 94, only two PTS are correctly identified whereas; 4 remaining PTS are not correctly identified (i.e. identified as NTS). Crash 1 is preventable using any L_{rm} value from 1 to 5 as it is possible to raise an alarm at least 5 minutes before the crash. Crash 60 is only preventable if L_{rm} is set to 1 so that preventive measures are implemented at 25 minutes before the crash. If L_{rm} is greater than 1, there are not enough L_{rm} consecutive PTS correctly identified and crash 60 becomes unpreventable. Similarly, crash 94 is preventable at 15 minutes before the crash only with L_{rm} =1 and is unpreventable with L_{rm} >1.

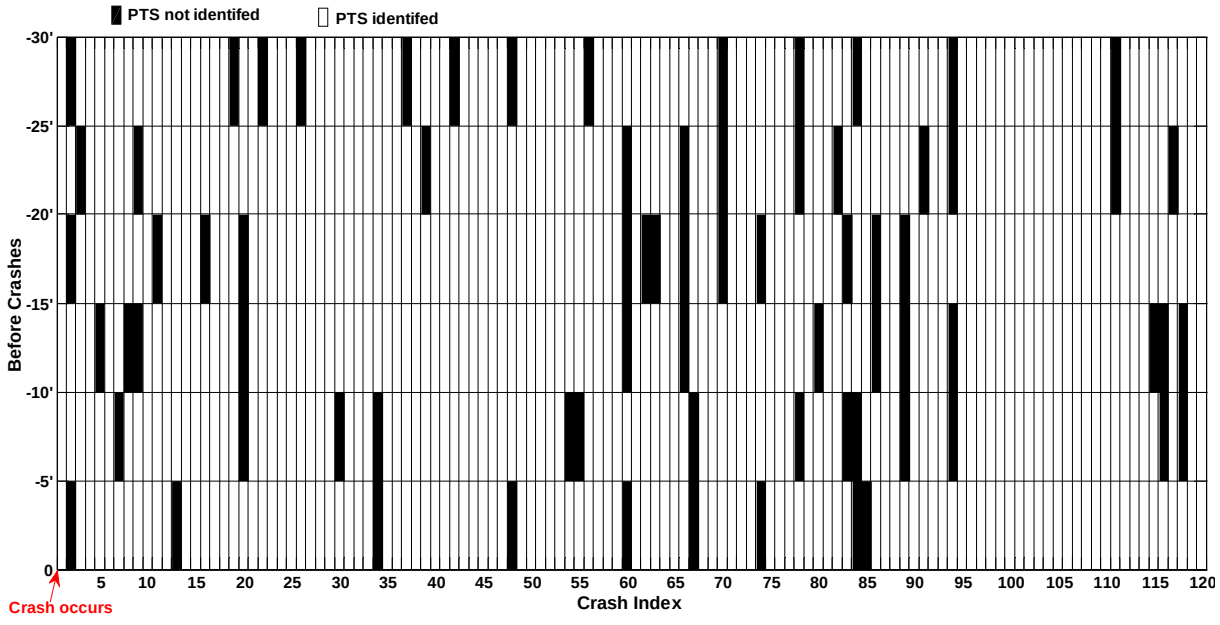

Figure 7-6: Evolution of risk identified before crashes

Figure 7-7 summarizes the frequencies of missed alarm according to five different L_{rm} values (from 1 to 5). The missed alarm frequency increases quickly as L_{rm} increases. Therefore, to minimize that missed alarm rate, it is necessary to maintain shorter length of risk memory.

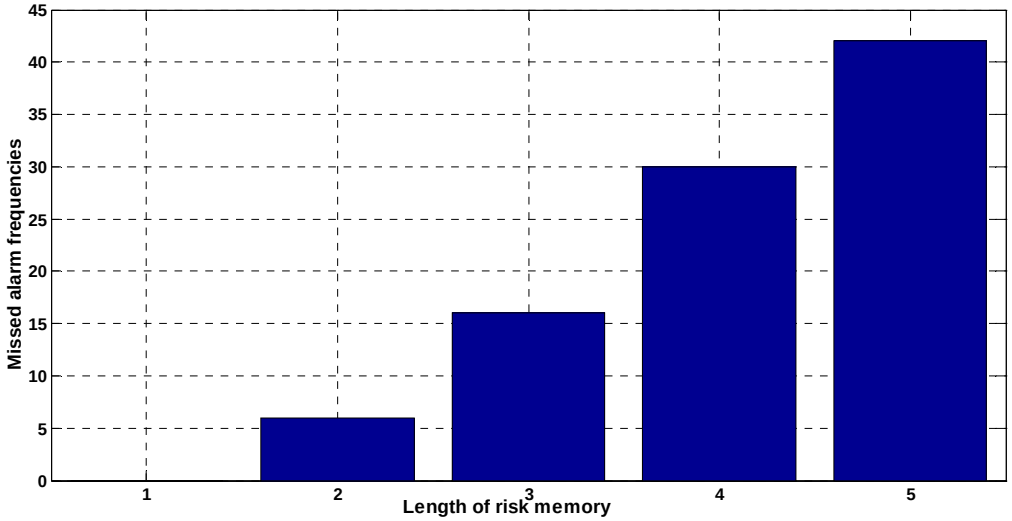

Figure 7-7: Missed alarm frequencies as function of L_{rm}

7.4.2.4. False– Missed Tradeoff

The results obtained from sections 7.4.2.2 and 7.4.2.3 are in contrast: to reduce false alarm rate, increasing the length of risk memory is needed whereas to reduce the missed alarm rate, reducing the length of risk memory is necessary. Therefore, there is no length of risk memory that minimizes both missed and false alarm rates.

Figure 7-8 presents the missed and false alarm rates as functions of the length of risk memory. Due to the low number of crashes, the missed alarm rate increases quickly. On the contrary, high number of non-crash cases results in high number of false alarms although the false alarm rate is low. The choice of the length of risk memory can be decided by traffic operators. The cost for making all crashes preventable, i.e. L_{rm}=1, is the false alarm rate at 3.7%. This means that there is about 53 minutes per day in average that the traffic is put in alarm state. If the false alarm rate is minimized, i.e. only about 0.2% or 2.88 minutes per day with L_{rm}=5, the missed alarm rate will be high at 35%. The false alarm rate of 0.2% is also the minimum rate, which indicates that positive false alarm rate is unavoidable.

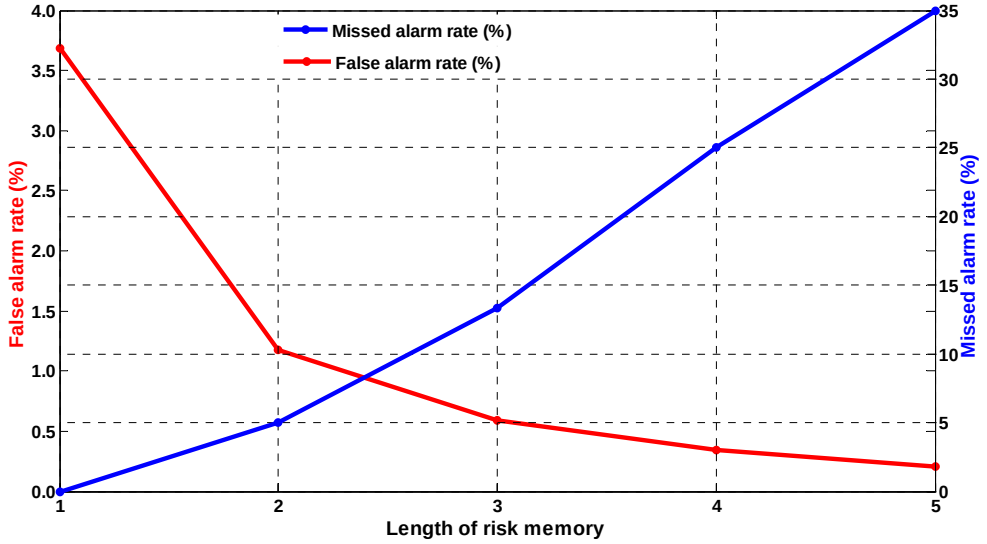

Figure 7-8: False and missed alarm rates as function of L_{rm}

Finding the best value of L_{rm} is challenging. However, depending on the priority of traffic operators, following scenarios might be considered.

- *Scenario 1*: If the crash rate of the considered road section (this is a supposed location, not the case of the study site of the current research) is very high, traffic operators might want to do a campaign to make drivers pay more attention to the danger of the location. Then, L_{rm}=1 is recommended. Human lives are the most important. With this value, the chance to prevent all rear-end and sideswipe crashes is high, i.e. the chance to save human lives is high. Although the false alarm rate is high, this might be a good way to draw the attention of drivers.
- *Scenario 2*: In general, high false alarm rate is undesirable. In the end, it is the false alarms that disturb drivers and traffic operators. If MyTRIM is applied in real life, current traffic conditions will be compared to traffic conditions before the application of MyTRIM in term of the disturbance it causes and the utility it brings. Therefore, L_{rm}=5 would make the influence of MyTRIM in term of disturbance more transparent whereas the chance to save human lives remains high (about 65% of crashes). In fact, L_{rm}=5 is the best choice because crashes are rare events. High false alarm rate would gradually make drivers ignore the presence of alarms while for a long time, crashes do not occur.

Other lengths of risk memory, i.e. L_{rm}=2, 3, or 4, can also be used. However, these values are more suitable for scenario 1. This is because false alarm rates corresponding to these values are high.

7.4.3. Applicability

In a real-time framework, traffic data are constantly collected and processed to generate traffic situations for the last aggregation interval, called the *current traffic situation* or *cTS*. Thereafter, cTS is classified into class of NTS or NTS.

If MyTRIM is used by traffic operators, what is the information to be sent to traffic operators?

There can be two cases discussed in sub-sections below.

7.4.3.1. Binary Outputs

If traffic operators are only interested in whether there is crash occurrence during the next traffic situations or not, the outputs of MyTRIM can be binary with two values corresponding to answers "yes" and "no" from MyTRIM to the operators.

In this case, the value of L_{rm} needs to be defined in advance.

Figure 7-9 presents an example of binary outputs returned by MyTRIM to traffic operators. At current moment (*Now*), traffic operators receive signal from MyTRIM saying that there will be a crash occurrence during the next traffic situation.

Figure 7-9: An example series of binary outputs returned by MyTRIM

Binary outputs are more suitable for cases where L_{rm} value is low, i.e. L_{rm} =1, 2, or 3 and traffic operators wait to activate preventive measures.

7.4.3.2. Multiple Level Outputs

If false alarm rate is more sensitive to traffic operators, the five outputs of MyTRIM corresponding to five L_{rm} values with respect to cTS can be combined and provided to traffic operators. Five outputs can be combined because the output value "yes" of higher L_{rm} also implies the output value "yes" of lower L_{rm}. Therefore, the combined output is the output "yes" of the highest L_{rm}.

Figure 7-10 presents four examples of multi-level outputs. In Case 1, no risk is identified at the beginning. At the second period, risk is identified, the output of MyTRIM is "yes" with L_{rm}=1, "no" with other L_{rm} values, the combined output is "yes" at level 1. At the third period, risk is identified; the output of MyTRIM is "yes" with L_{rm}=1 and 2 and "no" with other L_{rm} values, the combined output is "yes" at level 2 and so on. At the sixth period, risk is identified; the output of MyTRIM is "yes" with L_{rm}=1, 2, 3, 4, and 5, the combined output is "yes" at level 5.

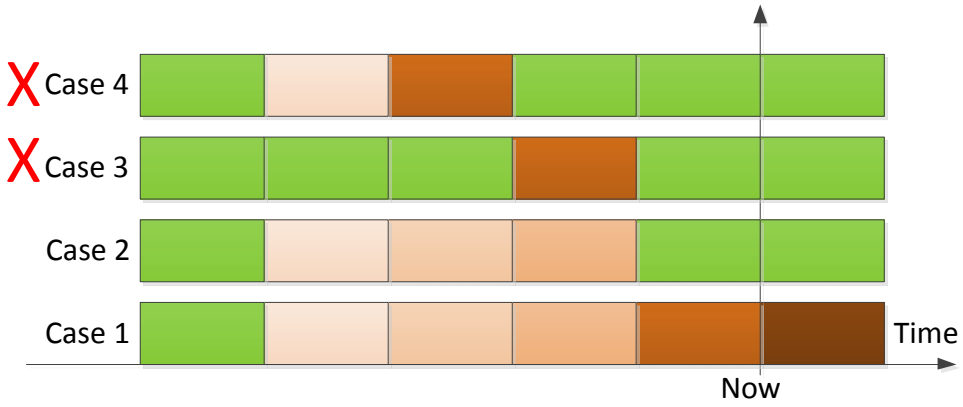

Figure 7-10: Examples of multi-level outputs

Similarly in Case 2, the risk develops from the second to the fourth period up to level 3. Thereafter, the risk disappears and the combine output become "no".

Case 3 can never occur as the risk jumps from "no" to "yes" at level 4. Case 4 can never occur either as the risk jumps from "yes" level 1 to "yes" level 4.

7.5. Summary

This chapter discusses about the development of Traffic Regime – based Risk Identification Models – TR-based RIM under highly risky traffic regimes and Motorway Traffic Risk Identification Model - MyTRIM. The obtained RIM allow identifying high percentages of NTS and PTS in validation data sets which indicates the high accuracy of the models in identifying new traffic situations into one of two classes: NTS and PTS. RIM are also refined such that variables having no impact on models' performance can be eliminated.

The developed RIM also identify variable importance. The most important variables are detected and called Critical Factors. The importance of variables is estimated by several approaches to find the correct Critical Factors. Critical Factors are useful in understanding causality of crashes and provide preliminary idea on developing preventive measures.

Also the performance of RIM is high, the overall false warning rate remain very high. This is why MyTRIM is developed. MyTRIM functions based on a parameter called length of risk memory - L_{rm} whose value ranges from 1 to 5. L_{rm}=5 is the most desirable as the false alarm rate is at acceptable level whereas the correct alarm rate remains high.

The applicability of MyTRIM is also discussed with respect to the form of MyTRIM's outputs which can be binary or multi-level. While binary output is more appropriate in term of crash prevention, multi-level output might be better for traffic operators who would like to observe the evolution of traffic risk in real-time.

Chapter 8 Conclusions

8.1. Summary

The current research investigates the motorway traffic crash risk identification by elaborating a methodology aimed at developing models capable of identifying real-time, called *Motorway Traffic Risk Identification Models - MyTRIM*. With a selected study site, the proposed methodology is implemented to exploit individual vehicle traffic data combined with meteorological as well as crash data, aiming to differentiate traffic conditions leading to crashes from other traffic-related conditions.

The imbalance between pre-crash and non-crash cases is one of the problems addressed in the current research (according to the rarity of crashes on motorways). A methodology for sampling non-crash data relevant to pre-crash data is proposed in the current research, and aims at avoiding an arbitrary selection of non-crash data (in comparison with pre-crash data). Results of data sampling process are clusters called Traffic Regimes.

Different classification and regression approaches were tested in order to choose the most suitable one - Random Forests Regression - that is capable of handling the data imbalance problem and have good performance with data in the current research. Under each Traffic Regime, a TR-based Risk Identification Model is developed to differentiate between NTS and PTS. Critical factors are also identified, which is useful in understanding causality of crashes and provides preliminary idea on developing preventive measures

Also the performance of RIM is high, the overall false warning rate remain very high. This is why MyTRIM is developed. MyTRIM functions based on a parameter called length of risk memory - L_{rm} whose value ranges from 1 to 5. L_{rm}=5 is the most desirable as the false alarm rate is at acceptable level whereas the correct alarm rate remains high.

The applicability of MyTRIM is also discussed with respect to the form of MyTRIM's outputs which can be binary or multi-level. While binary output is more appropriate in term of crash prevention, multi-level output might be better for traffic operators who would like to observe the evolution of traffic risk in real-time.

In the next section, the contributions of the current research are presented. Thereafter, the applicability of the obtained results is discussed in section 8.3. Section 0 examines the potential improvements in terms of performance of MyTRIM. Subsequently, possible future research directions are discussed in section 0.

8.2. Research Contributions

Here, the main contributions of the current investigation are discussed in details.

8.2.1. New Methodology for Modeling Risk Identification

Table 8-1 summarizes the differences in the methodology and compares them to the methodologies proposed in previous studies in term of raw data and variable use.

Table 8-1: Difference between risk identification modeling methodologies

Studies	Traffic variables from raw data	Agg. Interval	Single station?
Oh et al, 2001	Lane-based volume, occupancy, and average speed	10 sec	**Yes**
Golob et al., 2008	Lane-based volume and occupancy	30 sec	No
Lee et al., 2003	Lane-based volume, occupancy, and average speed	20 sec	No
Abdel-Aty et al., 2005	Lane-based volume, occupancy, and average speed	30 sec	No
Hourdakis et al., 2006	Individual vehicle data	-	No
Hossain et al., 2010	Station-based volume and average speed	5 min	No
Current research	**Individual vehicle data**	**-**	**Yes**

Most of the studies presented in the literature consider multiple traffic detector stations as a way to create new variables. The common trend of those studies, especially in studies by Hossain et al. and Abdel-Aty et al., is to perform tests using a different number of detector stations upstream and downstream crash locations. The only single station-based study was developed by Oh et al. However, none of these studies discusses potential variables that can be created by means of analysis. As a consequence, an important variable, the speed difference between lanes (identified in the current research as one of main factors contributing the high crash risk), was therefore ignored.

Besides, individual vehicle data is used by Hourdakis et al., 2006 and by the current research. However, Hourdakis et al., 2006 developed their model based on multiple traffic detectors.

8.2.2. Methodology for sampling non-crash traffic data

The work presented in this section is motivated by the rarity of motorway crashes, leading to the low performance of machine learning techniques, as illustrated in section 3.3.1. Four methodologies for sampling non-crash data are testedas follows:

i) S1 by Oh et al, 2001: non-cash cases at 30 min before and pre-crash cases right before crashes,

ii) S2 by Abdel-Aty et al, 2008: : matched case control – controlling the time of the day and the day of the week as well as weather conditions,

iii) S3 by Pande et al, 2007: random selection, and

iv) S4: Methodology proposed in the current research.

The methodology for the test includes the following steps:

- Define pre-crash cases. Pre-crash cases are the same when the last three sampling methodologies (S2, S3, and S4) are applied. When applying S1, only one pre-crash case is accounted for right before the crash. The remaining cases are considered non-crash cases for S1, S2, and S3.

- Sample non-crash data. Apply non-crash data sampling methodologies.
- Develop a risk identification model using Random Forest regression based on pre-crash data and on sampled non-crash data.

As a Random Forest regression is applied, the data is divided into three sub-sets: training, calibration, and validation data. Table 8-2 presents the results of the developed models using four non-crash data sampling methodologies (S1, S2, S3, and S4) for validation.

The methodology S1 presents a low performance as non-crash data is simply selected by taking data at 30 minutes before crashes. The application of S3 improves the model's performance, yet it is still low. This happens because the chance for irrelevant non-crash data, selected for comparison with pre-crash data, is equal to the chance for relevant non-crash data. The performance improves by applying S2; meaning by controlling the time of the day, the day of the week, and meteorological conditions; thereby the search space is reduced. However, the developed model does not perform well with new traffic conditions, which are not accounted for by such controls. With the proposed methodology S4, new traffic conditions are classified into existing traffic regimes. In several regimes, new conditions can be immediately declared as non-crash because crashes (rear-end or sideswipe) difficultly occur under those regimes. Under different circumstances, new traffic conditions are tested and classified into pre-crash or non-crash. Therefore, the performance of models developed using the proposed non-crash data sampling methodology S4 is much improved.

Table 8-2: Performance of data sampling methodology

Data sampling methodology	NTS (%)	PTS (%)
S1	51.00	39.00
S2	68.00	67.00
S3	58.45	61.45
S4 (Proposed method)	**89.83**	**83.62**

8.2.3. Improvement of Risk Assessment Accuracy

Risk assessment relates to the capacity of models to correctly identify pre-crash and non-crash traffic conditions. As the data used in other studies is not available, there is no mean to verify the accuracy of such models. Table 8-3 presents the summary of the best accuracy reported in those studies.

The missed alarm and the false alarm rates presented in Table 8-3 relate to the respective percentages of pre-crash and non-crash cases incorrectly identified as non-crash and pre-crash cases, respectively. A model is more explicative if it has lower missed and false alarm rates.

For each study, there is at least one data set that is used to develop the risk identification model, called training data set. Moreover, depending on the learning method (classification or regression), one or two other data sets can be used. The two other data sets are calibration and validation data sets. The accuracy presented in Table 8-3 is applicable to validation data sets (i.e. data sets that represent new data).

Table 8-3: Stated accuracy of relevant studies

Studies	Missed Alarm (%)	False Alarm (%)	Note
Oh et al, 2001	-	-	One data set for Training, calibration, and validation
Lee et al, 2003	-	-	One data set for Training, calibration, and validation
Hourdakis et al, 2006	41.67	6.81	One data set for calibration and validation
Abdel-Aty et al, 2005-2008	26.10	30.00	Two data sets for training and validation
Pande et al, 2005-2007	26.00	34.00	Two data sets for training and validation
Hossain et al, 2010	36.67	20.00	One data set for calibration and validation
Proposed RIM	**10.27**	**16.38**	**Three different data sets**

Two studies by Oh et al. 2001 and Lee et al. 2003 cannot be compared to other studies as no validation data sets were used. Two more studies, by Hourdakis et al. 2006 and Hossain et al. 2010, combine calibration and validation data sets in one single set and employ regression methods. Therefore, the developed models are not assessing new traffic data. Only two data sets are used in the studies of Abdel-Aty et al. and Pande et al., as the methods used were based on classification. Therefore, among the previous studies, only the results by Abdel-Aty et al. and Pande et al. are validated. However, the accuracy reported in those studies is much lower than the accuracy obtained by applying the methodology proposed in the current research.

8.2.4. Crash Risk Prediction

In previous studies, traffic crash risk assessment was well studied. A further improvement was suggested by Abdel-Aty et al. and Pande et al. in preventing crash risks: once the risk is identified, the incoming traffic conditions are also at high risk and preventive measures such as variable speed limits are immediately activated. In this case, the activation of preventive measures is dependent on the performance of risk identification models, again rather low (see Table 8-3).

In the current study, crash risk prediction is undertaken based on the test of several consecutive time intervals L_{rm}, called *length of risk memory*. The future crash risk is more certain if traffic conditions during those time intervals are identified as risky. As illustrated in Figure 7-8, false alarm and missed rates cannot be altogether minimized, yet an optimal value of L_{rm} can be selected based on the location of the study site.

It is worth noting that by fixing L_{rm}=1, MyTRIM works exactly as the model suggested by Abdel-Aty et al. and Pande et al.: whenever a risk is identified, the traffic crash risk is predicted to occur during the next time interval.

8.3. Applications

As discussed in 7.4.3, MyTRIM can be used by traffic operators. MyTRIM can provide binary or multiple level outputs. Besides, with the current version of MyTRIM, the following immediate applications can be considered in the context of the selected study site:

1) To provide directly to drivers the real-time traffic crash risk status (warning, alarm or both) via a real-time traffic information system, such as Variable Message Signs. This helps drivers to take the necessary precautions when they are within risky traffic conditions. Drivers have though no obligation to change their driving behavior.
2) To be used as a tool for evaluating traffic crash risks at the same road location before and after a change at the corresponding road section. The change can be, for instance, the opening/closure of a new on-ramp upstream or of a new off-ramp downstream. It is advisable that there should not be any infrastructural change at the location of the traffic detectors where the data MyTRIM is applied for the evaluation.

8.4. Potential Improvement

The performance of developed models in the current research is high although the model development process can be improved. Figure 8-1 presents the potential areas of improvement to optimize the obtained results. With the developed models, different design choices in methodology can be selected aiming to optimize performance.

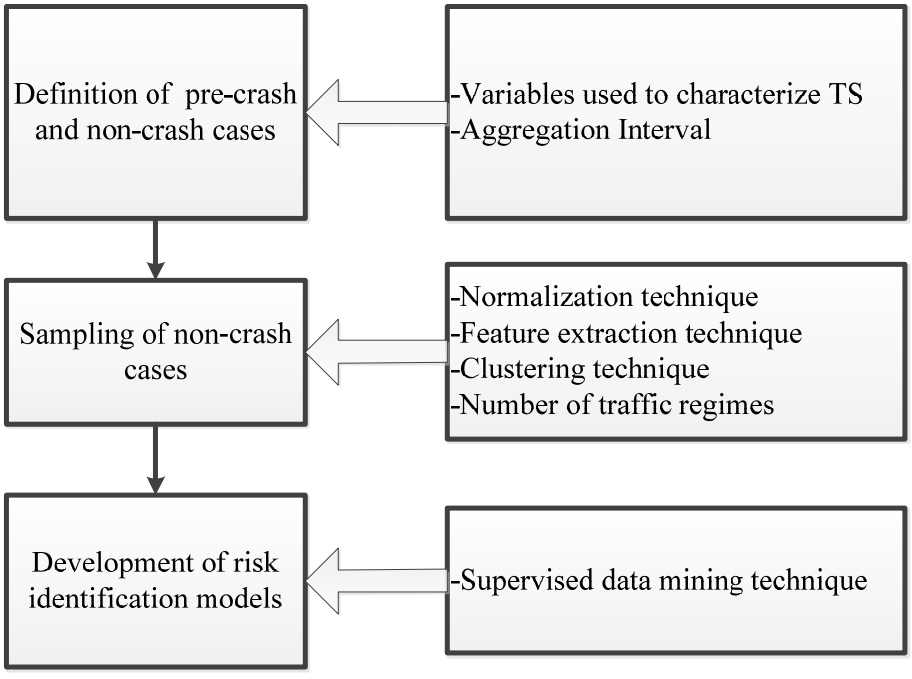

Figure 8-1: Potential improvement for optimizing performance

Several improvements can be implemented when defining NTS and PTS. Most of traffic related variables characterizing traffic situations in the current research are related to speed and flow. These variable choices facilitate the result interpretation as the variables are fundamental. The use of other variables such as risk indicators (for instance, TTC, PBTR, MARS, etc.) can be tested. The duration of the aggregation can also be optimized. Here, 5-minute intervals are used. However, this duration can be optimized with respect to the prediction performance of MyTRIM.

Similarly, in sampling non-crash cases and in developing risk identification models, the choices of normalization, feature extraction, clustering, and supervised data mining techniques can also be optimized.

8.5. Future Research Directions

Starting from the current study, following research directions can be targeted:

- Extending to a larger study site/road network
  - Use of a series of detectors
  - Use of floating car data.
- Extending to other data types such as Vehicle-To-Vehicle - V2V data/Vehicle-To-Infrastructure - V2I data
- Developing algorithms to manage motorways with objective based on Risk Indicators

8.5.1. Extensions to Larger Study Sites

So far, MyTRIM is developed for Swiss 2x2 motorways (i.e. motorways with two lanes per direction on the Swiss motorway network). The methodology for developing MyTRIM is flexible and can be applied to different study sites with similar road designs.

Depending on the availability of traffic detectors installed at the study sites and the number of lanes per direction, more variables can be added as inputs to develop the model. If there are many traffic detectors and the spacing between the detectors is low enough, variables representing traffic variation between neighbor detector stations (VT4 variables, see section 2.4.2) can be used.

Environmental factors (if available) can also be used as inputs for the model. Potential meteorological variables include:visibility, temperature, wind direction and speed, etc.

To this extent, data from floating cars can be used to characterize traffic situations. Using this type of data would produce another type of variables (other than VT1, VT2, VT3, and VT4) that is not based on traffic detectors.

8.5.2. Extensions to Other Traffic Data Types

Other traffic data types such as V2V or V2I are other alternatives providing traffic characteristics. The developed models are based on centralized traffic data collectors (i.e. traffic detectors). V2I data, with the infrastructure playing the role of centralized data collector, can be used to characterize the traffic at the infrastructure location.

In a V2V cooperative system the processors are every individual vehicle, there is no centralized processor. In this case, V2V data alone cannot provide the overview of traffic conditions of whole road sections. However, local traffic characteristics for each vehicle can be available. Thereby, it is possible that local risk identification models are developed at each vehicle.

8.5.3. Risk-based Motorway Management Algorithm

In the current study, VSL is the only active traffic management strategy employed as a countermeasure for preventing traffic crash risk when the risk is predicted to occur. The application of ramp metering, dynamic lane marking, managed lanes or the combination of them can potentially be used to prevent crash risks.

Based on the developed models, algorithms for managing risky traffic conditions can be elaborated aiming to make the traffic safer.

References

Abdel-Aty, M. & Pande, A. (2005), Identifying Crash Propensity Using Specific Traffic Speed Conditions, *Journal of Safety Research*, **36**(1), 97-108.

Abdel-Aty, M., Pande, A., Das, A. & Knibbe, W. (2008), Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems, *Transportation Research Record: Journal of the Transportation Research Board*, **2083**(-1), 153-161.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F. & Hsia, L. (2004), Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression, *Transportation Research Record: Journal of the Transportation Research Board*, **1897**(-1), 88-95.

accident, a. C. i. N. (2011), A Crash Is No Accident.

Aguero-Valverde, J. & Jovanis, P. (2008), Analysis of Road Crash Frequency with Spatial Models, *Transportation Research Record: Journal of the Transportation Research Board*, **2061**(-1), 55-63.

Aguero-Valverde, J. & Jovanis, P. P. (2006), *Spatial Analysis of Fatal and Injury Crashes in Pennsylvania*, Elsevier, Oxford, ROYAUME-UNI.

Allen, B. L., Shin, B. T. & Cooper, P. (1978), Analysis of Traffic Conflicts and Collisions, *Transportation Research Record*(667), 67-74.

Andrey, J. (2010), Long-Term Trends in Weather-Related Crash Risks, *Journal of Transport Geography*, **18**(2), 247-258.

Archer, J. (2001), Traffic Conflict Technique: Historical to Current State-of-the-Art.

Arnedt, J. T., Wilde, G. J. S., Munt, P. W. & MacLean, A. W. (2001), How Do Prolonged Wakefulness and Alcohol Compare in the Decrements They Produce on a Simulated Driving Task?, *Accident Analysis & Prevention*, **33**(3), 337-344.

Blum, J. J. & Eskandarian, A. (2006), Managing Effectiveness and Acceptability in Intelligent Speed Adaptation Systems, *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, pp. 319-324.

Boschung. (2010), Measuring Stations.

BPU. (2011), Statistics on Non-Occupational Accidents and the Level of Safety in Switzerland. Status 2010: Road Traffic, Sports, Home and Leisure.

Breiman, L. (2001), Random Forests, *Machine Learning*, **45**(1), 5-32.

Breiman, L. & Cutler, A. (2004), Rftools for Predicting and Understanding Data, *INTERFACE WORKSHOP*.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Belmont C.A. Wadsworth.

Chawla, N. V., Japkowicz, N. & Kotcz, A. (2004), Editorial: Special Issue on Learning from Imbalanced Data Sets, *SIGKDD Explor. Newsl.*, **6**(1), 1-6.

Chen, C., Liaw, A. & Breiman, L. (2004), Using Random Forest to Learn Imbalanced Data.

DfT. (2009), Reported Road Casualties Great Britain: 2008. Annual Report, Departement for Transport, London.

Ding, C. & He, X. (2004), K-Means Clustering Via Principal Component Analysis, pp. 225-232.

Donnell, E. T. & Mason, J. J. M. (2006), Methodology to Develop Median Barrier Warrant Criteria, *Journal of Transportation Engineering*, **132**(4), 269-281.

Edwards, J. B. (1999), Speed Adjustment of Motorway Commuter Traffic to Inclement Weather, *Transportation Research Part F: Traffic Psychology and Behaviour*, **2**(1), 1-14.

ETSC. (2006), Road Safety Performance Index Flash1: Pinning Them Down on Their Promise, *Road Safety PIN Flash*.

ETSC. (2008), Road Safety Performance Index - Flash 8: Reducing Deaths on Motorways.

EuroRAP. (2009), European Campaign for Safe Road Design.

FEDRO. (2009), Instruction: Traffic Counting Stations (Title Translated from French), Swiss Federal Roads Office.

FEDRO. (2009), Roads & Traffic - Facts & Figures 2009 : Annual Publication of the Swiss Federal Roads Office., Swiss Federal Roads Office

FSO. (2005), Road Traffic Accidents - Instructions (Tittle Translated from French).

FSO. (2005), Road Traffic Accidents - Survey Form (Title Translated from French), Federal Statistics Office.

FSO. (2010), Definitions (Title Translated from French), Federal Statistics Office.

Geurts, P. (2010), Bias Vs Variance Decomposition for Regression and Classification, in Maimon, O. & Rokach, L. (eds.), *Data Mining and Knowledge Discovery Handbook*, Springer US, pp. 733-746.

Golob, T. F., Recker, W. & Pavlis, Y. (2008), *Probabilistic Models of Freeway Safety Performance Using Traffic Flow Data as Predictors*, Elsevier.

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.

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.

Haj-Salem, H. & Lebacque, J.-P. (2009), Risk Index Modeling for Real-Time Motorway Traffic Crash Prediction, pp. 55-64.

Hayton, J. C., Allen, D. G. & Scarpello, V. (2004), Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis, *Organizational Research Methods*, **7**(2), 191-205.

Hayward, J. C. (1971), Near Misses as Ameasure of Safety at Urban Intersections, *The Pensilvania State University, Department of Civil Engineering*.

Hossain, M. & Muromachi, Y. (2010), Evaluating Location of Placement and Spacing of Detectors for Real-Time Crash Prediction on Urban Expressways, *89th TRB Annual meeting*, Washington DC.

Hourdakis, J., Garg, V., Michalopoulos, P. & Davis, G. (2006), Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations, *Transportation Research Record: Journal of the Transportation Research Board*, **1968**(-1), 83-91.

Jain, A. K. (2010), Data Clustering: 50 Years Beyond K-Means☆, *Pattern Recognition Letters*, **31**(8), 651-666.

Jolliffe, I. T. (2002), Principal Component Analysis, Springer, NY.

Kaiser, H. (1960), The Application of Electronic Computers to Factor Analysis, *Educational and Psychological Measurement*, **20**(1), 141-151.

Klingender, M., Ramakers, R. & Henning, K. (2009), In-Depth Safety Impact Study on Longer and/or Heavier Commercial Vehicles in Europe, *Power Electronics and Intelligent Transportation System (PEITS), 2009 2nd International Conference on*, pp. 368-373.

Kohonen, T. (1982), Self-Organized Formation of Topologically Correct Feature Maps, *Biological Cybernetics*, **43**(1), 59-69.

Lee, C., Hellinga, B. & Saccomanno, F. (2003), *Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic*, National Research Council, Washington, DC, United States.

Lee, C., Saccomanno, F. & Hellinga, B. (2002), *Analysis of Crash Precursors on Instrumented Freeways*, National Research Council, Washington, DC, ETATS-UNIS.

Lord, D. & Mannering, F. (2010), The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives, *Transportation Research Part A: Policy and Practice*, **44**(5), 291-305.

Minderhoud, M. M. & Bovy, P. H. L. (2001), Extended Time-to-Collision Measures for Road Traffic Safety Assessment, *Accident Analysis & Prevention*, **33**(1), 89-97.

Mouzon, O. d., Faouzi, N.-E. E., Pham, M.-H. & Chung, E. (2008), Road Safety Indicators: Swiss Results in Vaud Canton, *Advances in Transportation Studies an International Journal, Section B*, 81-96.

Mulder, M., Abbink, D. A. & Boer, E. R. (2008), The Effect of Haptic Guidance on Curve Negotiation Behavior of Young, Experienced Drivers, *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pp. 804-809.

NHTSA. (1997), Crashes Aren't Accidents.

OECD. (2004), "Glossary of Statistical Terms".

Oh, C., Oh, J.-S., Ritchie, S. G. & Chang, M. (2001), Real-Time Estimation of Freeway Accident Likelihood, *80th Annual Meeting of the Transportation Research Board, Washington, D.C., 2001*.

Oh, C., Park, S. & Ritchie, S. G. (2006), A Method for Identifying Rear-End Collision Risks Using Inductive Loop Detectors, *Accident Analysis and Prevention*, **38**(2), 295-301.

Pampel, F. C. (2000), *Logistic Regression: A Primer*, Sage Publications.

Pande, A. (2005), Estimation of Hybrid Models for Real-Time Crash Risk Assessment on Freeways, *Department of Civil and Environmental Engineering*, University of Central Florida, Orlando, FL 32816-2450, United States.

Pande, A. & Abdel-Aty, M. (2005), Identification of Rear-End Crash Patterns on Instrumented Freeways: A Data Mining Approach, *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, pp. 337-342.

Pande, A. & Abdel-Aty, M. (2006), Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions, *Accident Analysis & Prevention*, **38**(5), 936-948.

Pande, A. & Abdel-Aty, M. (2006), *Comprehensive Analysis of the Relationship between Real-Time Traffic Surveillance Data and Rear-End Crashes on Freeways*, National Research Council, Washington, DC, United States.

Pardillo Mayora, J. M. & Jurado Piña, R. (2009), An Assessment of the Skid Resistance Effect on Traffic Safety under Wet-Pavement Conditions, *Accident Analysis & Prevention*, **41**(4), 881-886.

Parker, D., Reason, J. T., Manstead, A. S. R. & Stradling, S. G. (1995), Driving Errors, Driving Violations and Accident Involvement, *Ergonomics*, **38**(5), 1036-1048.

RoadPeace. (2011), Road Crash - Not Road 'Accident'.

Rumar, K. (1985), The Role of Perceptual and Cognitive Filters in Observed Behavior, *Human Behavior and Traffic Safety*, Plenum Press, New York, pp. 151–165.

Rzepecki-Smith, C. I., Meda, S. A., Calhoun, V. D., Stevens, M. C., Jafri, M. J., Astur, R. S. & Pearlson, G. D. (2010), Disruptions in Functional Network Connectivity During Alcohol Intoxicated Driving, *Alcoholism: Clinical and Experimental Research*, **34**(3), 479-487.

SAN. (2011), Trial Driving License Et 2-Phases Training (Translated from French), Service des automobiles et de la navigation, Lausanne.

Satterthwaite, S. P. (1976), An Assessment of Seasonal and Weather Effects on the Frequency of Road Accidents in California, *Accident Analysis & Prevention*, **8**(2), 87-96.

Schapire, R. (1990), The Strength of Weak Learnability, *Machine Learning*, **5**(2), 197-227-227.

Stine, J. S., Hamblin, B. C., Brennan, S. N. & Donnell, E. T. (2010), Analyzing the Influence of Median Cross-Section Design on Highway Safety Using Vehicle Dynamics Simulations, *Accident Analysis & Prevention*, **42**(6), 1769-1777.

Umedu, T., Isu, K., Higashino, T. & Toh, C. K. (2010), An Intervehicular-Communication Protocol for Distributed Detection of Dangerous Vehicles, *Vehicular Technology, IEEE Transactions on*, **59**(2), 627-637.

Wang, C., Quddus, M. A. & Ison, S. G. (2009), Impact of Traffic Congestion on Road Accidents: A Spatial Analysis of the M25 Motorway in England, *Accident Analysis & Prevention*, **41**(4), 798-808.

WHO. (2004), World Report on Road Traffic Injury Prevention: Summary.

Yang, Q. & Wu, X. (2006), 10 Challenging Problems in Data Mining Research, *International Journal of Information Technology & Decision Making*, **5**(4), 597-604.

Appendix A: List of Abbreviations

Abbreviation	Full Phrase
AADT	Annual Average Daily Traffic
AL	Average over Lanes
CART	Classification and Regression Tree
CF	Critical Factors
cTS	Last/current Traffic Situations
DETEC	Federal Department of the Environment, Transport, Energy and Communications
DS	Differentiation between Stations
ETSC	European Transport Safety Council
EU	European Union
FEDRO	Swiss Federal Roads Office
FSO	Swiss Federal Statistics Office
IMRO	IMbalance RatiO
imp()	Impurity
IR	Impurity Reduction
km/h	Kilometer per hour

LB	Lane - Based
LR	Logistic Regression
L_{rm}	Length of Risk Memory
m/s	Meter per second
MeteoSwiss	Swiss Federal Office of Meteorology and Climatology
MPL	Multi-Layer Perceptron
MyTRIM	MotorwaY Traffic Risk Identification Model
NTS	Non-crash Traffic Situation
OD	Origin-Destination
OECD	Organization for European Economic Cooperation
OOB	Out-Of-Bag
PC	Principal Component
PCA	Principal Component Analysis
PET	Post-Encroachment Time
PTS	Pre-crash Traffic Situation
RF	Random Forest
RIM	Risk Identification Model
TCT	Traffic Conflict Technique

TET	Time Exposed Time-to-collision
TIT	Time Integrated Time-to-collision
TR	Traffic Regime
TS	Traffic Situation
TTC	Time-To-Collision
V2I	vehicle to infrastructure
V2V	vehicle to vehicle
Varim	Variable important of CART
VIM	Variable IMportance
VMS	Variable Message Signs
vph	Vehicles per hour
VSL	Variable Speed Limits
VTx	Variable Type x (x=1 to 4)
WHO	World Health Organization
WL	Weak Learners

Appendix B: Crash Declaration

In case of road traffic crash, there is a form standardized by Swiss Federal Statistics Office (FSO) for the police to fill in. FSO also provides a guideline explaining the meaning of each field in the form and the steps to follow to correctly identify the crash type. The form is a two-side paper illustrated in Figure B-1 and Figure B-2. The guideline can be found in (FSO, 2005).

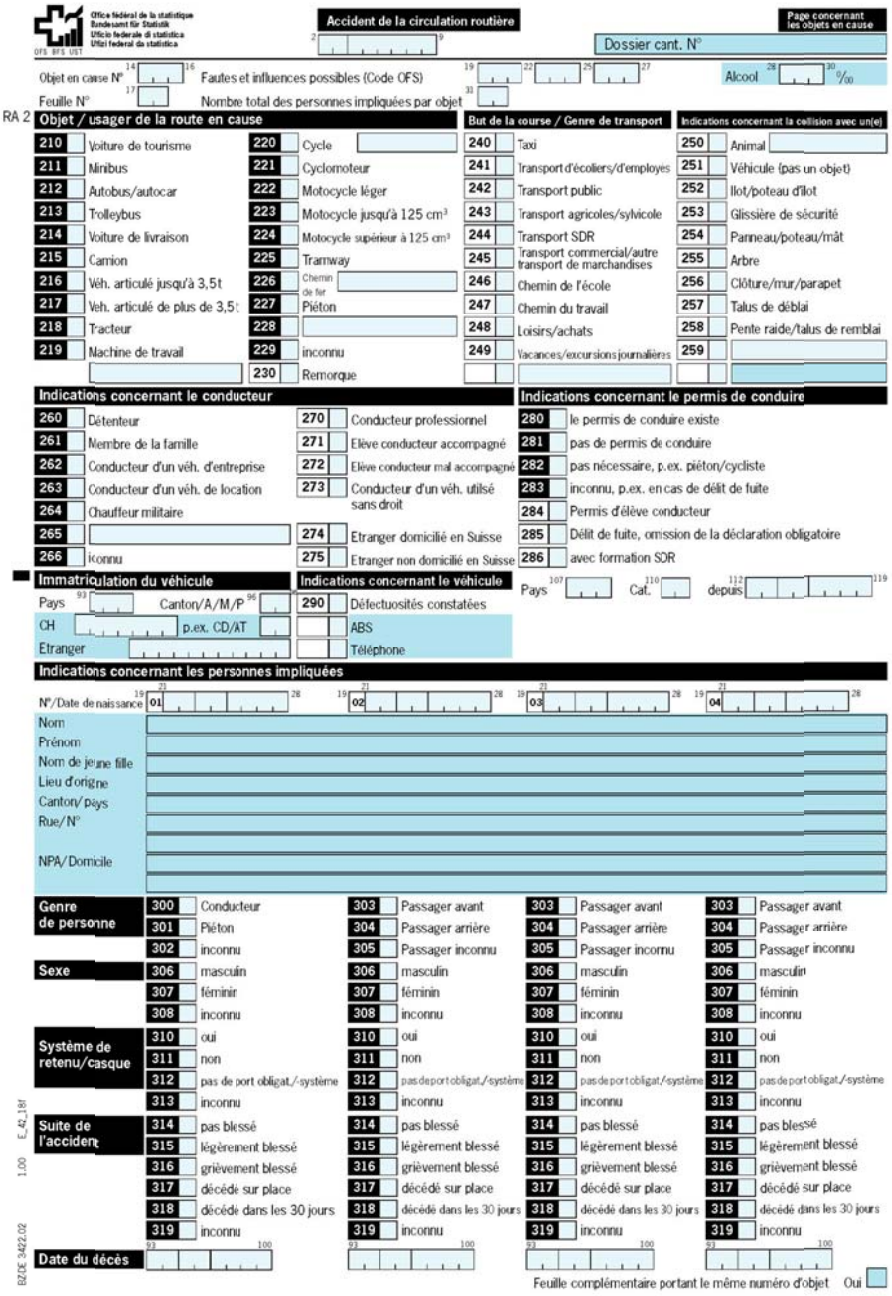

Figure B-1: Page 1 of the accident declaration form. Source: (FSO, 2005)

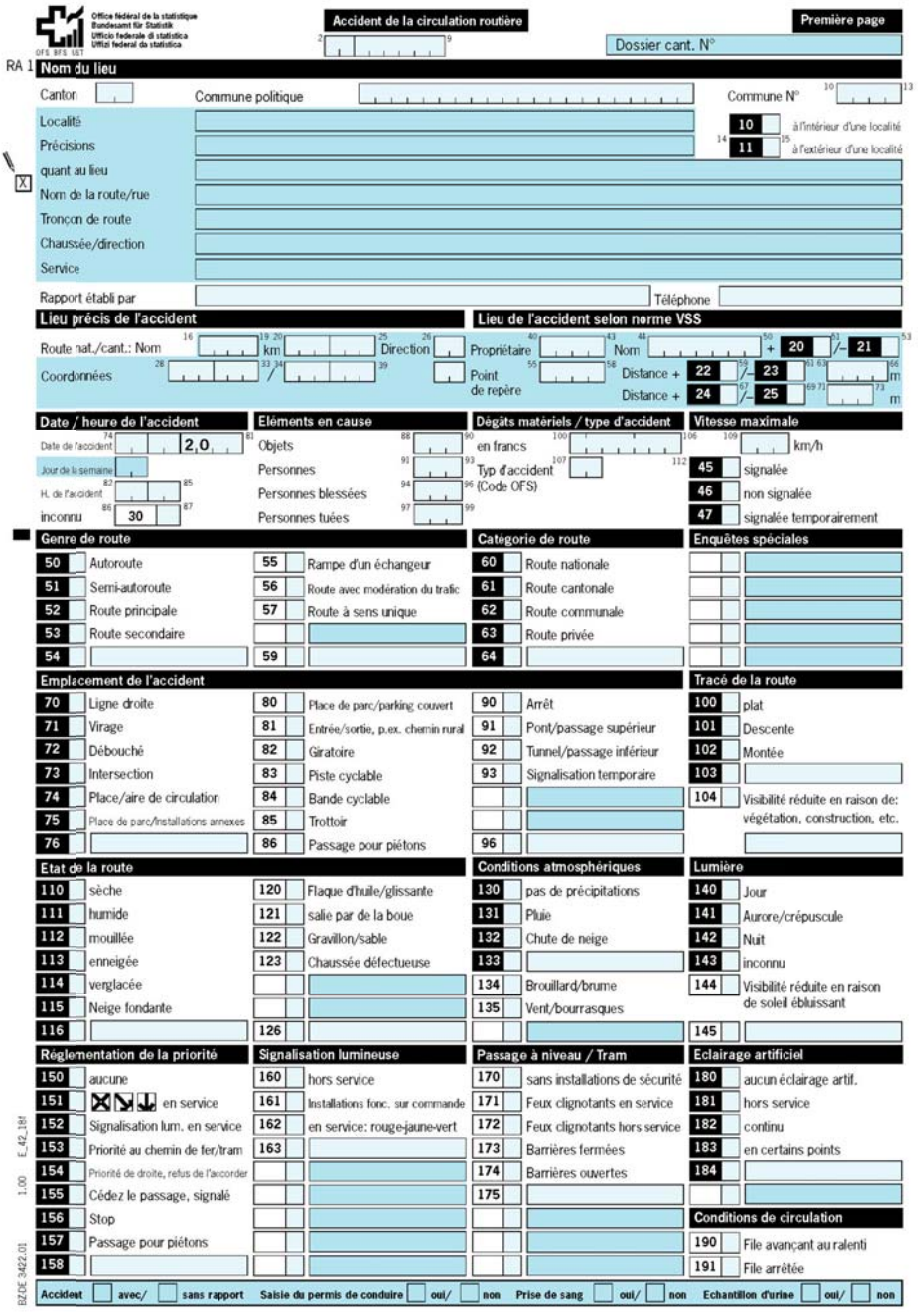

Figure B-2: Page 2 of the accident declaration form. Source: (FSO, 2005)

Figure B-3 presents the definition of 10 categories of crashes in Switzerland. Under each category, there are particular definitions for different crashes. Category A, for instant, defines rear-end crashes and includes two crash situations: front vehicle is static and front vehicle is moving.

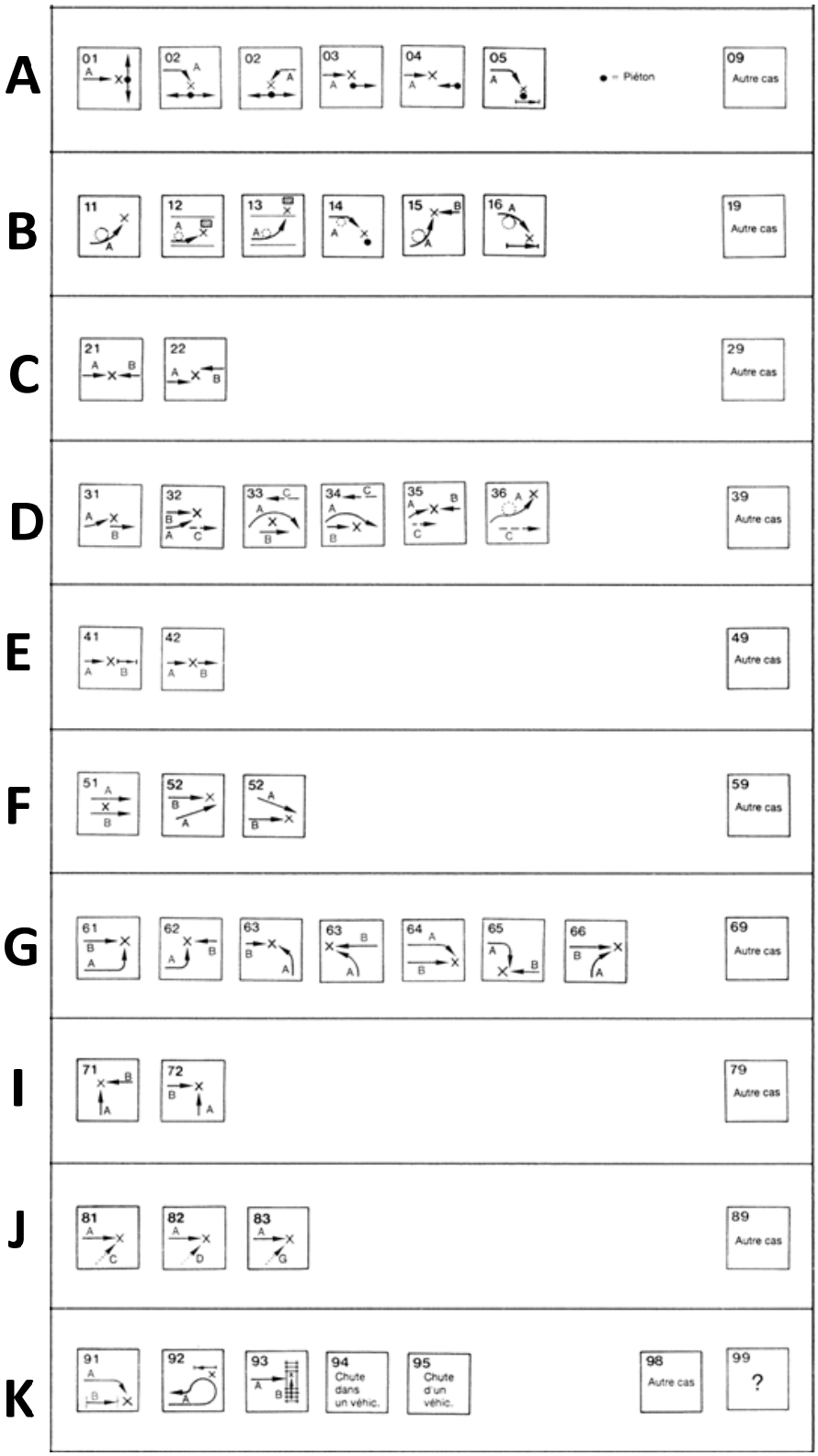

Figure B-3: Ten crash types. Source: (FSO, 2005)

Appendix C: Random Forests

C.1. Overview

Random Forest – RF is the main machine learning method used in the current research. In this chapter, the detailed of this method is introduced.

As RF is built based on Classification and Regression Trees - CART, it is essential that CART is also presented to provide step by step understandings of Random Forests.

C.2. Classification and Regression Trees (CART)

1. Introduction

CART (Breiman, Friedman, Olshen and Stone, 1984) is a non-parametric learning technique, using the methodology of tree building by recursively partitioning data in two smaller data set. CART classifies objects or predicts outcomes by selecting from a large number of variables the most important one in determining the outcome variable. To lower the training error presented in, CART reduces the training error at each partitioning step.

CART starts with the whole training data set, divides it into two subsets according to a partitioning criterion. The procedure repeats for each subset until the subset is not divisible according to a certain stopping criterion. Recursive partitioning creates a binary tree structure. Each partitioning is undertaken at a *node* of the tree. When a subset is not *divisible*, the corresponding node is called a *leaf* of the tree.

When a subset is *divisible*, it is partitioned into two smaller subsets. The partitioning is undertaken based on a *split* of a *decisive variable*. The decisive variable and the split are selected from *candidate combinations* of each variable and each of its *potential splits*. For a numerical variable, all of its values found in the data set under consideration are sorted. Any mean between two consecutive sorted values is a potential split value and the split based on that value is a potential split. Therefore, if there are h sorted values for variable X_j, there will be h-1 potential split values for X_j. For a categorical variable, any partitioning of categorical values into two disjointed subsets is a potential split. A partitioning of categorical values is undertaken by finding one subset of the values and then the other subset includes the remaining values unselected for the first subset. Therefore, if there are h categorical values, there will be $\frac{1}{2}\sum_{l=0}^{h}\prod_{m=0}^{l}(h-m)$ potential splits.

Consider a node called *NodeF* where F is the data set corresponding to that node. A variable X_j is called *decisive* at a node *NodeF* if the split at *NodeF* is based on that variable. Data set F is partitioned into two subsets F_{left} and F_{right} corresponding to $NodeF_{left}$ and $NodeF_{right}$, respectively, which are children of *NodeF*. The criteria for partitioning F into F_{left} and F_{right} are based on a measure called *impurity* that represents the homogeneousness of classes in data set F. If there is only one class in F, the impurity of *NodeF* is zero. At *NodeF*, the impurity is *imp*(*NodeF*). The pratitioning at NodeF is based on the split that gives the maximum impurity reduction *IR*(*NodeF*) calculated according to Equation 10.

Equation 10: Impurity Reduction at *NodeF*

$$IR(NodeF)=imp(NodeF)- imp(NodeF_{left})- imp(NodeF_{right})$$

2. Learning Algorithm

a. Impurity Measures

CART proposes both approaches: classification and regression depending on the outputs of the approach. The impurity measure also depends on the approach.

For the classification approach, Gini index is the most widely used as impurity measure *imp*(*NodeF*) calculated as presented in Equation 11, where p_{NTS} and p_{PTS} are proportion of *NTS* and *PTS* in *F*, respectively. *imp*(*NodeF*) is maximum if the proportions of NTS and PTS are equal.

Equation 11: Gini index measuring impurity in classification trees

$$imp(NodeF) = 1 - (p_{NTS}^2 + p_{PTS}^2)$$

For the regression approach, the sum of squared errors in *F* is used to estimate the *imp*(*NodeF*).

Equation 12: Sum of squared errors as impurity measure in regression trees

$$imp(NodeF) = \sum_{i \in F} (y_i - c_F)^2 \; where \; c_F = \sum_{i \in F} y_i$$

b. CART Growing Algorithm

Figure C-1 presents the algorithm for training CART. Let *I* is the set of data set. The algorithm starts with the data set containing all training data called *TSSet* and *I*={*TSSet*}. Thereafter, the algorithm comes to a loop to process all the data sets available in *I*.

For any data set *F* available in *I*, *F* is picked up from *I* (i.e. *F* is removed from *I*). For each variable X_j, find all possible splits and find the split IR_j that gives maximum reduction of impurity. For all variables, find IR_{max} which is the maximum reduction of impurity among all IR_j. Thereafter, F is partitioned based on the split giving IR_{max} into F_{left} and F_{right}.

Relating to the stop criteria, F_{left} and F_{right} are considered to be whether divisible. If F_{left} (or F_{right}) is divisible, it is put in *I* (i.e. *I*=*I* U { F_{left} (or F_{right})}. If F_{left} (or F_{right}) is not divisible, the node corresponding to F_{left} (or F_{right}) become a leaf of the tree.

There are many criteria to consider whether *F* is divisible. One of the criteria is to check whether there is one class or one observation in data set *F*. The tree obtained based on this criterion is the maximum tree.

The maximum trees have one class at each leaf (for classification trees) or one observation at one leaf (for regression trees) and therefore according to Equation 5(*l*), have zero bias.

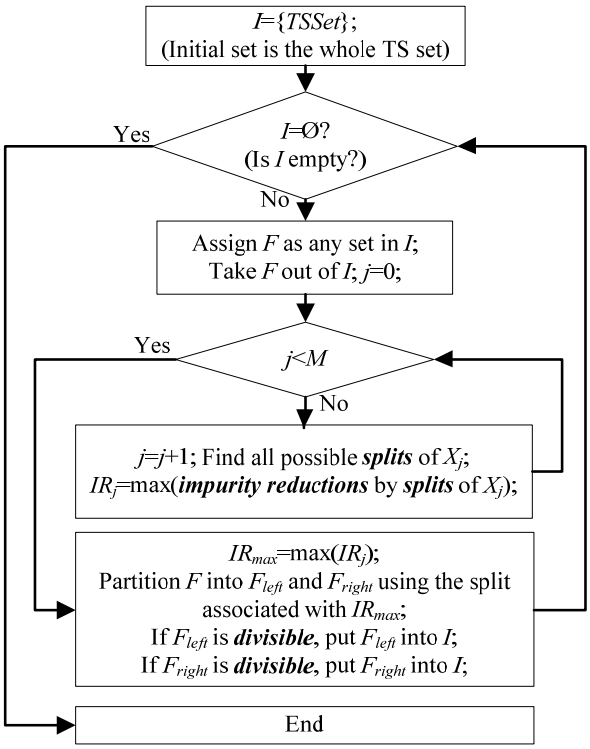

Figure C-1: CART training algorithm

c. Tree Pruning

One of issues of CART is that the maximum CART fits very well the training data yet has low performance with test data (i.e. the generalization of maximum CART is poor). For this reason, the obtained tree is pruned to improve the generalization. However, in this study, tree pruning is not of interest because the main purpose is to introduce Random Forest who makes use of maximum trees. The readers who are interested in CART pruning can have more information in (Breiman, Friedman, Olshen and Stone, 1984).

3. Variable Importance

The importance of a variable on a tree is estimated by the weighted impurity reduction obtained from all the nodes of the tree where that variable is decisive. At *NodeF* which is not a leaf, the importance of the decisive variable X_j is the impurity reduction obtained at *NodeF* weighted by the proportion of observations branching to *NodeF* compared to the whole training data set *TSSet*. Let *Varim*(*Xj*) the importance of variable X_j for the whole tree, $Varim_F(X_j)$ the importance of X_j at *NodeF*, $Prop_F$ the proportion of observations split into data set *F* compared to the whole data set *TSSet*. $Varim_F(X_j)$ is calculated based on Equation 13.

Equation 13: Importance of decisive variable at a node

$$Varim_F(X_j) = IR(NodeF)Prop_F \; where \; Xj \; is \; decisive \; variable \; at \; NodeF$$

Varim(*Xj*) is calculated based on Equation 14.

Equation 14: Importance of variable X_j for the whole tree

$$Varim(X_j) = \sum_{All\ NodeF\ where\ X_j\ is\ decisive} Varim_F(X_j)$$

Intuitively, a variable can become more important if:

It is decisive for many times, i.e. it reduces much the impurity on the tree.
It is decisive at the nodes at or near the root of the tree where the proportion of observations compared to the whole data set is high.

A variable having zero importance is the variable that is not decisive at any node of the tree. That variable does not participate or contribute to the classification or regression process.

C.3. Random Forests (RF)

1. Introduction

Random Forest (Breiman, 2001) is an ensemble learning method that generates many classification and regression trees (CART - (Breiman, Friedman, Olshen and Stone, 1984)), trains the trees and aggregates their results. Successive trees do not depend on earlier trees - each is independently constructed using a bootstrap sample of the data set.

According to Breiman, (2001), the motivation for inventing RF is that CART is an unstable together with its moderate accuracy. Maximum trees usually work well with training data but have low performance with test data. Tree pruning can improve CART performance with test data and result in trees with relatively higher accuracy. However, CART is unstable as even a small change in training data could also lead to totally different trees which make tree interpretation become problematic. Therefore, the main idea of RF is to create many unstable trees (i.e. the maximum trees) fitting very well the training data such that there is no correlativity between any pair of trees and then aggregate trees' results.

2. Weak Learners (WL)

According to Schapire, (1990), a weak learner (WL) is *the learner that can produce an hypothesis that performs only slightly better than random guessing*. The author also concluded that it was possible to convert *a mediocre learning algorithm into one that performs extremely well*.

According to Breiman and Cutler, (2004), a weak learner *is a prediction function that has low bias* which comes at the cost of high variance. Breiman and Cutler also demonstrate the idea stated in (Schapire, 1990) that converting weak learners in some way can generate a learn having high prediction power.

Maximum trees are good example of weak learners as they fit very well with training data. Figure A-2 presents the performance of four weak learners which are regression trees (CART) in predicting outputs of the sine function $y=sin(2\pi x)$. Training and test data are generated using the given sine function. Four training data sets are randomly generated to fit four weak learners. The test data set is generated and represented in Figure A-2 as *Original Function*. Thereafter, the test data are input to four weak learners. The outcomes of the test are illustrated in Figure A-2 as "*Weak learner 1*", "*Weak learner 2*", "*Weak learner 3*", and "*Weak learner 4*". It is clear that there is high variance between the predicted outputs of weak learners and the expected outputs from the original function. However, if outputs of weak learners for each input are averaged, the averaged output is the better estimation of the original output. According to (Breiman, 2001), if the number of weak learners come to infinitive, the average outputs are precisely the outputs of the original function. The converted learner, which is averaging in this example, is called an *ensemble learner*.

3. Randomization

In reality, it is not always possible to get training data such as in the example illustrated in Figure A-2. In many cases, there is only one training data set. Therefore, it is necessary to use the training data set effectively. By applying maximum trees as WL, another issue is that for one training data set, there can be only one maximum tree generated.

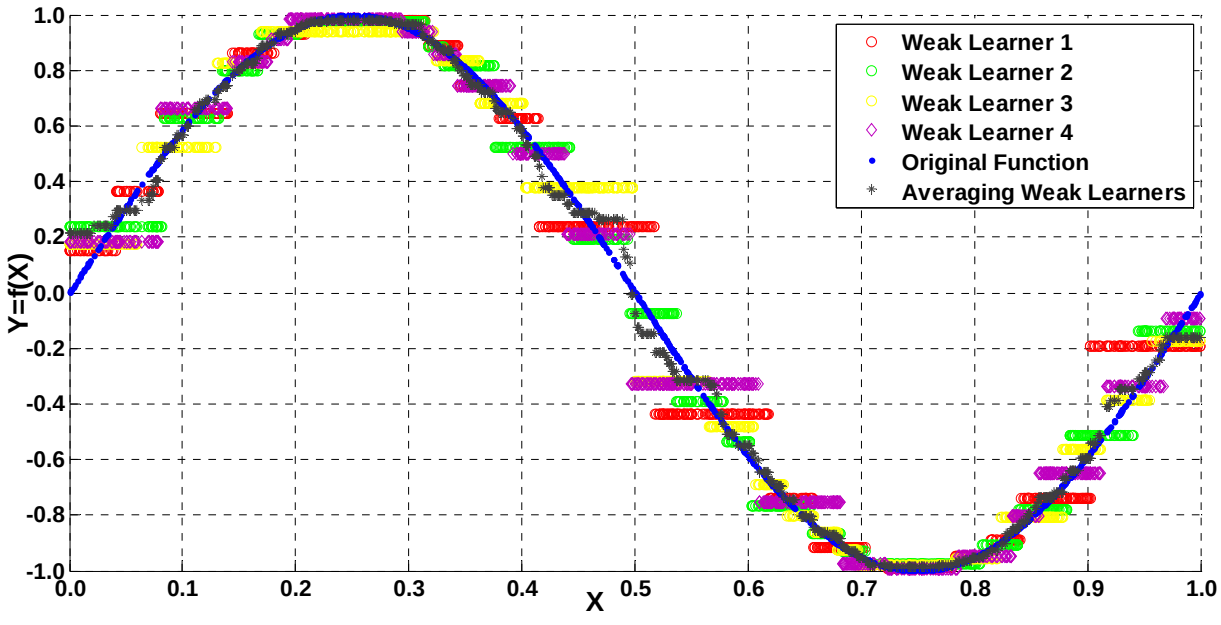

Figure C-1: Maximum regression trees as weak learners and averaging weak learners

Therefore, (Breiman and Cutler, 2004) prove that if the independent, identically distributed randomization of weak learners is used, the bias of the ensemble learner is approximately unchanged compared to the

bias of weak learners whereas; the variance is significantly reduced. The authors also point out that generated WL should have low correlation to get higher performance of the ensemble learner that aggregates WL.

The independent, identically distributed randomization is used to generate the *t*-th tree in two steps below:

1) Generate a bootstrap sample of the training set called B_t where t=1:N_{tree} (the total number of trees).
2) Grow the maximum tree using generated data set such that:
   a) At each node, *m* variables are selected at random out of *M* variables.
   b) The split used is the best split on these *m* variables.

A tree is obtained from two randomizations: the training data set for that tree and the selection of variables at each node of the tree. B_t is the set of indices of observation selected for the *t*-th tree and is sampled by replacement from *TrainingSet*.

4. Learning Algorithm

The algorithm for building RF is presented in Figure A-3*(a)*. Trees in RF are trained in the similar manner with training CART (see Section b). The algorithm for growing trees in RF is presented in Figure A-3(*b*).

The inputs for training RF include:

- The training data set *TSSet* which is a part of the whole available data set contained in matrix *X*. *TSSet* is also equal to *TrainingSet* mentioned in Section 0.
- The number *m* of variables randomly selected at each tree node.
- The number N_{tree} of trees grown in RF.

According to (Breiman, 2001), *m* is recommended to be squared root of *M* (which is the total number of variables in matrix *X*) and N_{tree} is selected by trying different values.

Growing a tree in RF based on the data set L_t is similar to growing CART using L_t presented in Figure A-3. The main difference is that at each node of the tree in RF, the split is determined among potential splits given by *m* randomly selected variables whereas the split in CART is determined among all potential splits given by all *M* variables.

Besides growing trees, there are other processing steps as presented in Figure A-3*(a)* such as calculating error E_t for the *t*-th tree, calculating Out-Of-Bag error for the whole RF, and evaluating variable importance. These processing steps are discussed in sections 5 and 6 of this chapter.

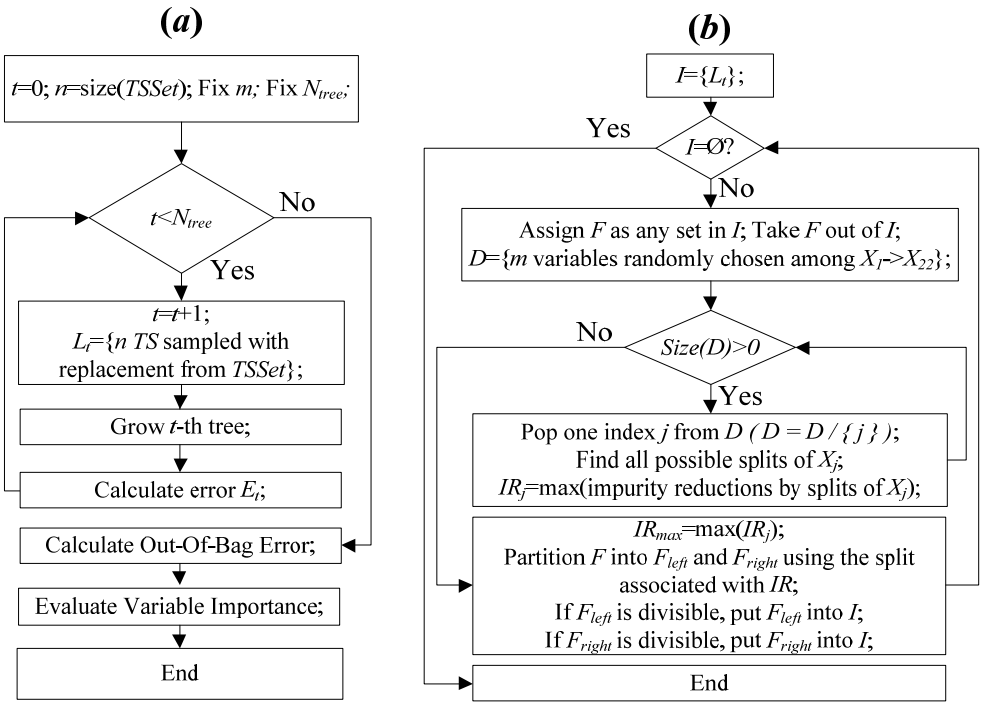

Figure C-2: RF training algorithm

When all trees are grown, RF aggregates the results. For RF regression, there are N_{tree} trees grown and the likelihood for the i-th observation x_i (i.e. Traffic Situation) given by the t-th tree is $T_t(x_i)$, the likelihood for the i-th observation by RF is presented in Equation 15.

Equation 15: Aggregating trees to generate outputs of RF regression

$$RF(x_i) = \frac{1}{N_{tree}} \sum_{t=1}^{N_{tree}} T_t(x_i)$$

For RF classification, the class for the input x_i is given by the major vote of trees' outputs.

5. Out-Of-Bag Data and Errors

The size of data for training RF is n. By sampling with replacement, the probability for a datum to be selected is 1/n and the probability for the datum not to be selected is (n-1)/n. The size of sampled data is also equal to n. Therefore the probability for a datum not to be selected in the sample data is $\left(1-\frac{1}{n}\right)^n$. If n is large then $\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = e^{-1} \approx 0.37$ (or 37%) is the proportion of data not sampled from original training data. These data are called *Out-Of-Bag* (*OOB*) data for the tree grown base on the data sampled. The index of observations in *OOB* data for the t-th tree is contained in OOB_t data set.

For the *t*-th tree in RF, observations in OOB_t are not used for training the tree. Therefore, OOB_t can be used to test the prediction performance of the tree as new data. The term E_t presented in **Error! Reference source not found.**(*a*), called *Out-Of-Bag error* of the *t*-th tree, is the test error of *t*-th tree for observations corresponding to OOB_t. E_t is calculated according to Equation 16.

Equation 16: Out-Of-Bag error of the *t*-th tree

$$E_t = \sum_{i \in OOB_t} \left(y_i - Tree_t(x_i)\right)^2 \; ; \; where \; Tree_t(x_i) \; is \; the \; output \; of \; t-th \; tree \; for \; x_i$$

For an observation x_i of *TSSet*, there is a set called $OOBTree_i$ containing the indices of all the trees, for which x_i is in the OOB data. Let OE_i the mean squared error of all the trees whose indices are in $OOBTree_i$ for x_i as input. OE_i is calculated according to Equation 17.

Equation 17: Out-Of-Bag error of *i*-th observation (x_i) in *TSSet*

$$OE_i = \frac{1}{size(OOBTree_i)} \sum_{t \in OOBTree_i} (y_i - Tree_t(x_i))^2$$

Finally, let *OE* the Out-Of-Bag error of RF, *OE* is estimated according to Equation 18*(a)* or Equation 18*(b)*

Equation 18: Out-Of-Bag error of Random Forests

(a) $OE = \sum_{i=1}^{P} OE_i$

(b) $OE = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{size(OOBTree_i)} \sum_{t \in OOBTree_i} (y_i - Tree_t(x_i))^2$

6. Variable Importance

Let $VIM(X_j)$ the variable importance of variable X_j. There are two different methods for estimating the $VIM(X_j)$. These two methods usually give similar estimation of variable importance. Two methods: using CART-like technique and using value permutation error are described in this section.

a. Using Mean VIM over Trees (CART-like Technique)

Equation 14 estimates the importance of variable using CART (Section 3). Although trees are grown in RF in a slightly different way compared to growing CART, the principle for evaluating variable importance for CART is still applicable for each tree of RF. $VIM_t(X_j)$ is the importance of variable X_j estimated by the *t*-th tree in RF. $VIM_t(X_j)$ can be estimated by following Equation 14. $VIM(X_j)$ is estimated by following Equation 19.

Equation 19: Estimation of variable importance by CART-like technique

$$VIM(X_j) = \sum_{t=1}^{N_{tree}} VIM_t(X_j)$$

b. Using Value Permutation

The *t*-th tree has a error E_t for its OOB_t data set. For each variable, it values in OOB_t are permuted over all observations in OOB_t while keeping values of other variables unchanged. Put down the tree the observation in OOB_t. As the relationship between variables in the observations is broken, the OOB error E_t is expected to increase to become E_{tj} (i.e. OOB error of *t*-th tree when values of variable X_j are permuted). The average increase of OOB error of a variable over all the trees divided by the standard deviation of the increase is the estimation of the variable importance as presented in Equation 20. A variable whose permuted values cause high increase of OOB error is more important than a variable whose permuted values cause lower increase of OOB error.

Equation 20: Variable importance by using OOB error

(a) $avgVIM(X_j) = \sum_{t=1}^{N_{tree}} E_{tj}$

(b) $stdVIM(X_j) = \sum_{t=1}^{N_{tree}} (E_{tj} - avgVIM(X_j))^2$

(c) $VIM(X_j) = \frac{avgVIM(X_j)}{stdVIM(X_j)}$

Appendix D: RIM Refinement

D.1. Overview

This appendix is the complement to section 7.3.4 discussing on the refinement of regime-based Risk Identification Models. This appendix also attempts to prove that the models presented in section 7.3.4 are better than other models.

D.2. Refinement Algorithm

Given a set *X* of variables *X*={*Var_1*, *Var_2*, …, *Var_n*} and six data sets: NTS and PTS for training, NTS and PTS for calibration, and NTS and PTS for validation. The process to train regime-based RIM is used to develop RIM and is summarized below:

- Models are developed using training data sets.
- Pre-crash thresholds are set using calibration data sets. Threshold criteria are applied.
- Model performance is evaluated using validation data sets.

When a model is developed and there is no pre-crash threshold satisfying the first two threshold criteria (see section 7.2.3.3), the model is call *inappropriate* which means that the model will not be further considered. The first two threshold criteria require that a threshold to be set should identify at least 70% of NTS and 70% of PTS in the calibration data sets.

Figure D-1 illustrates the recurrent function to find *good* models – the models which are not inappropriate and having high performance. The function *Refine* takes as input the set of initial variables and returns the set of models. In the body of *Refine* function, there are calls to four functions: *DevelopModel, SetThreshold, Validate* and *Refine* (self-call). The function *DevelopModel* takes as input the set of variables and develops a model with the variables using training data sets. The call to function *Refine* inside its body is a recurrent call to develop models with smaller sets of variables. For instant, the set *X'* of variables contains 1 variable less than the set *X*. From the set X', models with subsets of variables can be developed. However, the call *Refine (X')* is applicable only if model *M* - result of *DevelopModel*(*X'*) is not inappropriate – which means that a model is not refined if the model cannot satisfy the first two threshold criteria.

If the model *M* is not inappropriate, its pre-crash threshold is set using the third threshold criteria via function *SetThreshold*. Thereafter, model *M* is validated using validation data sets via function *Validate*. Model *M* is considered a good model if it can satisfy the following two validation criteria:

1) 70% of PTS in validation data set identified
2) 70% of NTS in validation data set identified

The model having the best performance should satisfy the validation criteria and have the greatest total sum of percentages of NTS and PTS in validation data sets identified.

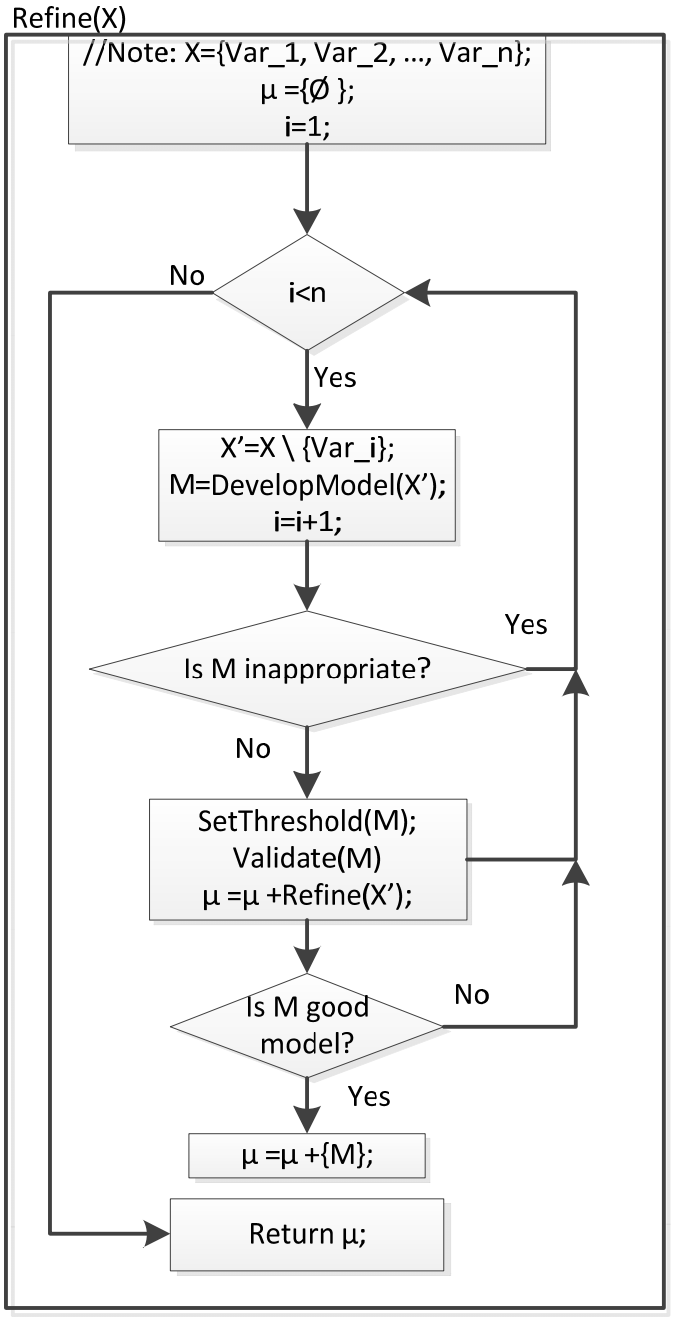

Figure D-1: Recurrent function for refining models.

Curriculum vitae

Minh Hai PHAM

Phone: +41 (0) 21 693 06 03

Email: minhhai.pham@epfl.ch; pminhhai@gmail.com

Date of Birth: 05st August 1978

Nationality: Vietnamese

Education

Master of Computer Sciences, Institut de la Francophonie pour l'Informatique, Hanoi, Vietnam	*Oct 2002 - Sept 2004*
Engineer of Information Technology, Hanoi University of Technology, Hanoi, Vietnam	*Sep 1996 - May 2001*

Major research activities

Doctoral dissertation *"Methodology and Application of Real-Time Motorways Traffic Risks Identification Models (MyTRIM)".*	*Dec 2006- Feb 2011*
COST-TU0702 research project: Network Vulnerability (EU COST action project).	*Since 2010*
NEARCTIS research project: Network of Excellence for Advanced Road Cooperative Traffic management in Information Society (EU-FP 7 project)	*Since 2010*
FUSAIN research project: FUSion of SAfety INdicators. Application of data fusion techniques to improve the performance of safety indicators (Swiss research project)	*Since Jul 2007*
COST 352 project: Improvement of Individual Driver Behavior Model for the Safety Assessment of Traffic Flow Simulation (EU COST action)	*Jun 2006-Dec 2008*

INTRO (INTelligent ROad): Study the influence of weather conditions on traffic safety (European project)	*Jun 2006-Jun 2008*

Teaching activities

Preparation of lectures and tutorials for following topics at graduate, master and doctoral level course, EPFL, Switzerland**:**	
Traffic flow theory; Statistical data analysis; Traffic safety.	*Since 2010*
Practical assignments: Tutor for master students to understand and use traffic simulation software AIMSUN**.**	*Since 2008*
Lecture Assistant: Database Management Systems, Thang Long University, Hanoi, Vietnam	*2002-2006*
Online academic course management using MOODLE.	*Since 2008*

Other professional activities

International Symposium on Traffic Simulation 2006 (ISTS06), *Lausanne, Switzerland* Member of local organizers team of the ISTS06 symposium.	*September 2006*

Invited speaker

TRANSINFRA, Fribourg, Switzerland. "Une stratégie pour la gestion active et sensible aux risques du trafic sur autoroute"	*March, 2010*
Journée Technique LAVOC, EPFL, Switzerland. "Motorway traffic crash risks".	*September 2009*

List of publications

Journal papers

[1]Mouzon, O. d., Faouzi, N.-E. E., **Pham, M.-H.** & Chung, E. (2008), Road Safety Indicators: Swiss Results in Vaud Canton, Advances in Transportation Studies an International Journal, Section B, 81-96.

[2] **Pham, M.-H.**, Chung, E., Mouzon, O. d. & Dumont, A.-G. (2007), Season Effect on Traffic: A Case Study in Switzerland, SEISAN KENKYU, 59(3), 214-216.

Book Chapters

[1]**Pham, M.-H.** & Chung, E. (2009), Chapter 4: Risk Indicators - Unsafe Traffic Conditions, Cost Action 352 - the Influence of in-Vehicle Information Systems on Driver Behaviour and Road Safety, Czech Republic: COST Office, pp. 44-54.

[2] **Pham, Minh-Hai** ; Bernhard, D. ; Diallo, G. ; Messai, R. et al. SOM-based Clustering of Multilingual Documents Using Ontology. In: Data Mining with Ontologies: Implementations, Findings and Framewoks, 2008, p. 65-82. Hershey, USA: Information Science Reference, 2008.

Conferences

[1]**Pham, M.-H.**, Faouzi, N.-E. E. & Dumont, A.-G. (2011), Real-Time Identification of Risk-Prone Traffic Patterns Taking into Account Weather Conditions, 90th Transportation Research Board annual meeting, Washington DC.

[2]**Pham, M.-H.**, Bhaskar, A., Chung, E. & Dumont, A.-G. (2011), Methodology for Developing Real-Time Motorway Traffic Risk Identification Models Using Individual Vehicle Data, 90th Transportation Research Board annual meeting, Washington DC.

[3]**Pham, M.-H.**; Bhaskar, Ashish ; Chung, Edward ; Dumont, André-Gilles Presented (2010). Random Forest Models for Identifying Motorway Rear-End Crash Risks Using Disaggregate Traffic Data and Meteorological Information 13th ITSC, Madeira Island, Portugal, 19 – 22 September 2010.

[4]**Pham, M.-H.,** Bhaskar, A. & Dumont, A.-G. (2010), Vers Un Modèle Proactif Pour Identifier Des Risques De Trafic, Conférence Sécurité routière: prévention des risques

et aide à la conduite, PRAC2010 Paris, France.

[5]**Pham, M. H.**, Bhaskar, A., Chung, E. & Dumont, A. G. (2010), Towards a Pro-Active Model for Identifying Motorway Traffic Risks Using Individual Vehicle Data from Double Loop Detectors. Road Transport Information and Control Conference and the ITS United Kingdom Members' Conference (RTIC 2010) - Better transport through technology, IET, pp. 1-9.

[6]**Pham, M.-H.,** Chung, E. & Dumont, A.-G. (2009), Methodology for Traffic Risks Identification, Swiss Transportation Research Conference Ascona, Ticino, Switzerland.

[7]**Pham, M.-H.,** Mouzon, O. d., Chung, E. & Faouzi, N.-E. E. (2008), Sensitivity of Road Safety Indicators in Normal and Crash Cases, 10th AATT Conference, Athens, Greece.

[8]**Pham, M.-H.**, Mouzon, O. d., Chung, E. & Dumont, A.-G. (2008), Sensitivity of Risk Indicators under Motorway Traffic Regimes Clustered by Self-Organizing Map, 7th European Congress on ITS, Geneva, Switzerland.

[9]Mouzon, O. d., **Pham, M.-H.**, Faouzi, N.-E. E. & Chung, E. (2008), An Effective Real-Time Proactive Road Traffic Risk Indicator Based on Traffic Data: Compensated Platoon Braking Time Risk (Cpbtr), 7th European Congress on ITS, Geneva, Switzerland.

[10] **Pham, M.-H.**, Mouzon, O. d., Chung, E. & Dumont, A.-G. (2007), Applicability of Road Safety Indicators to Assess Driving Risks under Swiss Road Conditions, Swiss Transportation Research Conference Ascona, Ticino, Switzerland.

[11] Mouzon, O. d., **Pham, M.-H.**, Faouzi, N.-E. E. & Chung, E. (2007), Road Safety Indicators: Swiss Results in Vaud Canton, Road Safety and Simulation (RSS), Roma, Italy.