國 立 交 通 大 學

交通運輸研究所

博士論文

No.056

事故鏈與因果分析

Accident Chain and Causality Analysis

指導教授：汪進財

研 究 生：鍾易詩

中 華 民 國 九 十 七 年 一 月

事故鏈與因果分析

ACCIDENT CHAIN AND CAUSALITY ANALYSIS

研 究 生：鍾易詩　　Student: Yi-Shih Chung

指導教授：汪進財　博士　　Advisor: Dr. Jinn-Tsai Wong

國立交通大學

交通運輸研究所

博士論文

A Dissertation

Submitted to Institute of Traffic and Transportation

College of Management

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

in

Management

January 2008

Taipei, Taiwan, Republic of China

中 華 民 國 九 十 七 年 一 月

事故鏈與因果分析

學　　生：鍾易詩　　指導教授:汪進財　博士

國立交通大學交通運輸研究所

摘要

分析事故因果關係為改善交通事故、提升交通安全的重要方法之一。本研究的目的在利用交通事故資料庫，憑藉先進之方法論與逐漸成熟之電腦計算能力，從事故鏈的觀點有效挖掘事故發生影響因子以及事故因果關係。研究中以粗略集合理論作為從橫斷面事故資料有效取得事故鏈之方法，該理論的優點在於可同時控制眾多影響變數，反應事故發生為眾多因子交互作用的本質；粗略集合理論規則的產生為比較事故個體差異的結果，可有效避免總計誤差在資料推論時可能造成之謬誤。

本研究以事故鏈為核心概念進行三項研究：首先，藉由系統性地導入不同組合之條件屬性，分析粗略集合理論解釋事故資料之能力，以及粗略集合規則解釋事故鏈之有效性。接著以粗略集合規則對應之事故發生頻率為指標進行資料分群，以事故鏈的觀點分析事故資料異質性。最後以成對比較粗略集合規則的方式，分析事故情境變動對事故後果可能之影響，藉以挖掘可能之事故因果關係。

本研究利用內政部警政署之事故資料庫，針對台灣地區小客車單一車輛事故進行實證分析。研究結果發現：粗略集合理論的上下界近似、近似精度、近似品質、規則產生數以及判中率，為比較不同事故種類發生過程的有效指標，粗略集合規則並可幫助研究者了解事故發生情境。在單一車輛事故中，衝撞道路設施為可預測性較高的事故種類，其與撞建築物、衝出路外與翻車事故之發生過程可能類似。另外，經常發生與稀少發生之事故型態特性確有明顯差異；前者為過去研究中常被指稱為高風險之駕駛族群，後者則與不良之駕駛環境連結。過去常以改善道路環境作為增進交通安全的方式，此等手段雖可有效降低中、低發生頻率情境之事故，但良好之道路環境可能間接鼓勵高風險駕駛人提高行駛速度。研究並發現事故的發生並非由單一因素造成，而是由一連串不利因素組合而成。若能在事故鏈中移除部分不利因素，有可能改變事故後果、降低事故嚴重度。

本研究以事故鏈為核心所主張之分析概念與架構，提供一個更貼近事故發生本質的分析方法，其中並對交通安全研究中常見之總計誤差、資料異質性、干擾因子等議題，進行深入探討。本研究所提之分析架構，可根據研究者手中資料完整性、對分析對象的了解程度，在事故鏈的大架構下作相對應之延伸。

關鍵字：事故鏈、交通安全、粗略集合、總計誤差、異質性

ACCIDENT CHAIN AND CAUSALITY ANALYSIS

Student: Yi-Shih Chung Advisor: Dr. Jinn-Tsai Wong

Institute of Traffic and Transportation
National Chiao Tung University

Abstract

Analyzing accident causality has been one of the many ways to enhance traffic safety. The objective of this research was to explore contributing factors and accident causality by utilizing crash databases with mature methodologies and powerful computational powers from chain perspective. Rough sets theory was adopted in this research to obtain accident chains from cross-sectional databases. This theory is advantageous due to its ability to simultaneously control numerous factors, which reflect the fact that the occurrence of accidents results from complex interactions of many contributing factors. The other advantage is that rough set rules are generated by comparing the individual differences, which would partially alleviate the issue of aggregation bias.

Three studies were conducted based on the concept of accident chains. The first study was to assess the ability of rough sets theory in explaining the underlying process of accident occurrence and in demonstrating accident chains by systematically loading combinations of condition attributes into rough sets. Second, the issue of data heterogeneity was examined from chain perspective by grouping accidents with the occurring frequency of rules. Finally, accident causality was addressed by comparing individual rules in pairs.

Taiwan's crash databases were adopted in the empirical study, where single auto-vehicle (SAV) accidents were chosen as the subject to analysis. It was found that lower/upper approximation, accuracy of approximation, quality of approximation, number of generated rules, and hit rates could effectively address the differences between accident types. The occurrence of crashes with facility may follow similar paths and is more predictable; these crashes have some similarities between the crashes with architecture, with facility, off-road and rollover types. Moreover, significantly different features were shown between frequently repeated and sparsely unique rules. The former rules linked to the characteristics of high-risk drivers shown in past studies while the latter was connected with poor road conditions. Providing better road environment has been considered as an effective way to improve traffic safety; however, better roads could encourage high-risk drivers to raise their driving speeds. Furthermore, instead of one single factor the combinations of unfavorable factors were found to be the causes leading to fatal accidents. If one or several undesired factors were removed from the chain, accident severity might be reduced.

The proposed approaches in the research provide a way to analyze accidents closer to

the essence of accident occurrence. Meanwhile, these approaches also provide alternative ways to alleviate issues often seen in safety research such as aggregation bias, heterogeneity of accident data, confounding factors, and so on. These approaches can be expanded based on analysts' on-hand data and their understanding of target subjects.

Keywords: Accident Chain, Traffic Safety, Rough Sets, Aggregation Bias, Heterogeneity

ACKNOWLEDGMENTS

I owe a debt of gratitude to many people whose help has been crucial to my success in completing this dissertation. First of all, I have been privileged to have the direction and guidance of one excellent mentor, Prof. Jinn-Tsai Wong. From the inception of this dissertation, Prof. Wong has been generous with his time and has provided eminently helpful advice on everything from the methodological design to the writing process. His ability to bring me back on track at difficult times has been a godsend. He also made invaluable contributions to the development of my ideas and the organization of the analysis. He has always motivated me to do my best work, offering timely feedback on each chapter and keen insight into the significance of my research. My other committee members, Prof. Hsin-Li Chang, Cherng-Chwan Hwang, Chi-Kang Lee, Pin-Yi Tseng, Ming-Chih Tsai and Li-Yen Chang, have also contributed insightful and critical comments and suggestions. I also received exceptionally fine guidance from teachers at each Ph.D. student seminar, including Prof. Lawrence Lan, Cheng-Min Feng, Jiuh-Biing Sheu, Mu-Chen Chen, and Yu-Chiun Chiou.

I have enjoyed the camaraderie of all the graduate students and friends at the Institute of Traffic and Transportation, NCTU, including Huk, Yi-Wen, Meng-You, Shih-Chang, Yi-San, Sharon, Simon, Ian, Jacky, Chao-Hung, Cheng-Hsieh, Anter, Chun-Ming, Beni, Jau-Rong, John Ko, Su-Ru and Wei Yu. I have benefited much from discussions with each of them at various times and places. My esteemed colleague Charles kindly offered his professional knowledge, practical experience and field data throughout the whole study. Without his help, this dissertation would never be finished.

Many other people have helped me weather the emotional storms and stress of graduate school, some of whom deserve special mention. Chin-Chih helped me to finish the editorial jobs for the Journal of the Chinese Institute of Transportation. To all the staffs on Taipei campus and friends at Yoga class, I am grateful for their companionship, diversions, and nourishment of body and soul.

Finally, I owe much to my parents for always believing in me and encouraging me to achieve my goals. My father, to whom this dissertation is dedicated, provided me with rare opportunities to go abroad for obtaining a master's degree and to expand my horizons. My mother has always been there for me, and has never failed to do what she could to further my progress. Hui-Ling, the one I deeply love, helped me to focus on my academic pursuits with her compassion, generosity, and steadfast emotional support. She forever inspires me with her infinite capacity for love, joy, and faith.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 INTRODUCTION

The chapter consists of four sections. Section 1.1 addressed the principal concept on analyzing accident characteristics and causality in this study. The research problems, objectives, and framework were introduced in Section 1.2, 1.3, and 1.4, respectively.

1.1 Accident Causality Analysis from Chain Perspective

Exploring the causality of accidents is what transportation professionals and others have devoted themselves to. Understanding the causality of accidents can help us to know not only how accidents occur but also the possible ways to avoid accidents. To improve traffic safety, apprehending only correlations is not enough. Moreover, knowing distorted causal facts is even more dangerous. For example, the installation of street lights had been believed to increase safety, but it has been well known that the installation could result in higher driving speeds and may lead to more accidents (Elvik, 2004). Therefore, understanding the causality of accident occurrence is the best and may be the only way to effectively manage traffic safety.

The causes of an accident have usually been described with the closest-to-accident factors. Researchers, however, have tended to analyze accidents in a more thorough perspective – looking into not only an accident itself but also the activities and factors prior to and subsequent to the accident. Some accidents were found to be preventable not by correcting driving behaviors but by adjusting behaviors prior to driving (Eby *et al.*, 2000; Simoes, 2003). In other words, an accident may be prevented if one or more undesirable elements during in process were removed (Baker and Ross, 1961; Fleury and Brenac, 2001; Reason, 1997). Therefore, analyzing and preventing accidents from the chain perspective becomes an alternative approach to understanding accident causality.

Analyzing accidents via the chain concept should be taken along with two elements: the consideration of multiple factors and the ability to make causal inference between factors. The consideration of multiple factors reflects the fact that the generating process of accidents is complicated. Unless all important factors are accounted for, the confounding effects would bias the estimation results (Elvik, 2002). As for the relationships between factors and accident consequences, there should be of directional connections to show their causality. In short, the consideration of multiple factors and the causal relationship between factors are the two required elements in implementing traffic accident analyses and preventions from the chain perspective.

There have been two types of related research that apply to such an idea. One of them pre-specifies the contents of chains. The contents of an accident chain include contributing

factors and accident outcomes. For example, Elvik (2003) proposed to use a causal chain approach to reduce possible confounding effects on safety countermeasure evaluation studies. The approach was named a causal approach since the causality between factors and accident consequences was designated by professionals and treated as a true causality prior to data exploration. The sequence of factors was put into a logical and temporal order; the strength of links between factors was then estimated with data. Such an analysis is particularly useful to evaluate the effectiveness of safety countermeasures related to road improvement since engineering improvement usually follows physical laws; accordingly causal relationships between factors are concrete. The second type of research is to explore possible chains from data; the plausibility of possible causality is then judged. The causal chains derived from this approach are not limited to the evaluation of safety countermeasures. Instead, all possible causal chains in an accident database could be explored. For example, Chang and Wang (2006) adopted the classification and regression tree technique in analyzing the traffic injury severity in Taiwan. The population and conditions with higher risks of being injured was identified by observing the derived trees. Research adopting this approach usually interprets the outcomes from the correlation perspective rather than the causality perspective; the logical and temporal orders in the generating process of accidents are not always explicit.

Three opportunities appear gradually providing the potential to overcome the aforementioned shortcomings, which include the becoming comprehensive accident databases in Taiwan, powerful computational capabilities, and mature methodologies. The accident databases in Taiwan have been built and maintained by National Police Agency, Ministry of the Interior. Although not as complete as Fatality Analysis Reporting System (FARS, the accident database of United States), after several revisions and improvements, the current Taiwan accident databases provide some important factors for analyses. Moreover, the AI and data mining methodologies have been growing in recent years. Although some techniques are black-box types, others are easier to understand and have good performance as well. In addition, the evolution of computational power provides the opportunity to calibrate parameters with complex forms. Grabbing these opportunities may provide the potential to explore accident causality from cross-sectional databases and thus motivate this research.

1.2 Research Problems

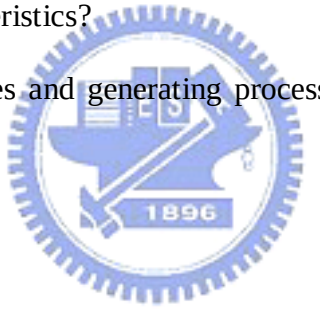
Learning accident patterns is one of the many ways to demonstrate accident chains. Different accident cases could be represented with different contributing factors, interactive relationships, and activity chains. However, paths leading to an accident are countless and

complicated if all details are concerned; it would be technically impossible to analyze accidents in such a detail. A compromise way is to classify data based on either prior knowledge or on statistical information extracted from data, which is called a pattern. An accident pattern describes a typical condition of accident occurrence such as driver characteristics, vehicle types, weather conditions or road conditions. When similar conditions occur, similar accident consequences would be expected.

The first problem this research desired to address was whether those accident patterns significantly exist or accidents just occur without patterns. If accident patterns significantly exist, how their characteristics could be explored. Accident patterns are expected to consist of most important factors in accident occurrence, so it might be possible to identify the causes of accidents and quantify them.

All in all, this research was trying to explore accident causality by examining the following problems in sequence:

- Do accident patterns significantly exist?
- If yes, what are their characteristics?
- Can the corresponding causes and generating processes be identified, quantified and analyzed?



1.3 Research Objectives

There were two primary objectives in this research:

1. Propose an approach for identifying accident patterns and exploring their characteristics:
   The approach proposed in this research was aimed to eliminate the effects of confounding factors and reveal individual differences of accidents. Given this objective, two types of methodologies were employed. One was the classification methodology which was adopted to eliminate the confounding factors among entities. The second one was statistical methodologies including descriptive statistics and regression-type techniques. They were adopted to explore the characteristics of the derived accident patterns.

2. Propose an approach for examining accident causality:
   Accident patterns represent accident chains, and accident chains demonstrate the circumstances under which accidents occur. The second objective was aimed to examine accident causality based on the derived accident patterns.

1.4 Research Framework

Given the objectives, the research framework was illustrated in Figure 1-1. Prior to analyzing accidents, framework of driving safety was built up as the basis to select appropriate contributing factors, to develop suitable approaches, and to judge the validity of derived accident patterns. Meanwhile, the connections among accident data, accident analyses, and countermeasure development were discussed to help define the research scope. Three studies were then conducted based on the coverage of data. The first one was a generalized two-step approach for exploring accident characteristics from chain perspective. Based on this, the second study was undertaken for analyzing the heterogeneity of accident data, a phenomenon usually shown on accident data especially on cross-section data. The third study was conducted to examine accident causality by comparing rules in pairs. Empirical studies were presented in the following chapter. The related issues were discussed in Chapter 5 and the conclusions and recommendations were drawn in Chapter 6.

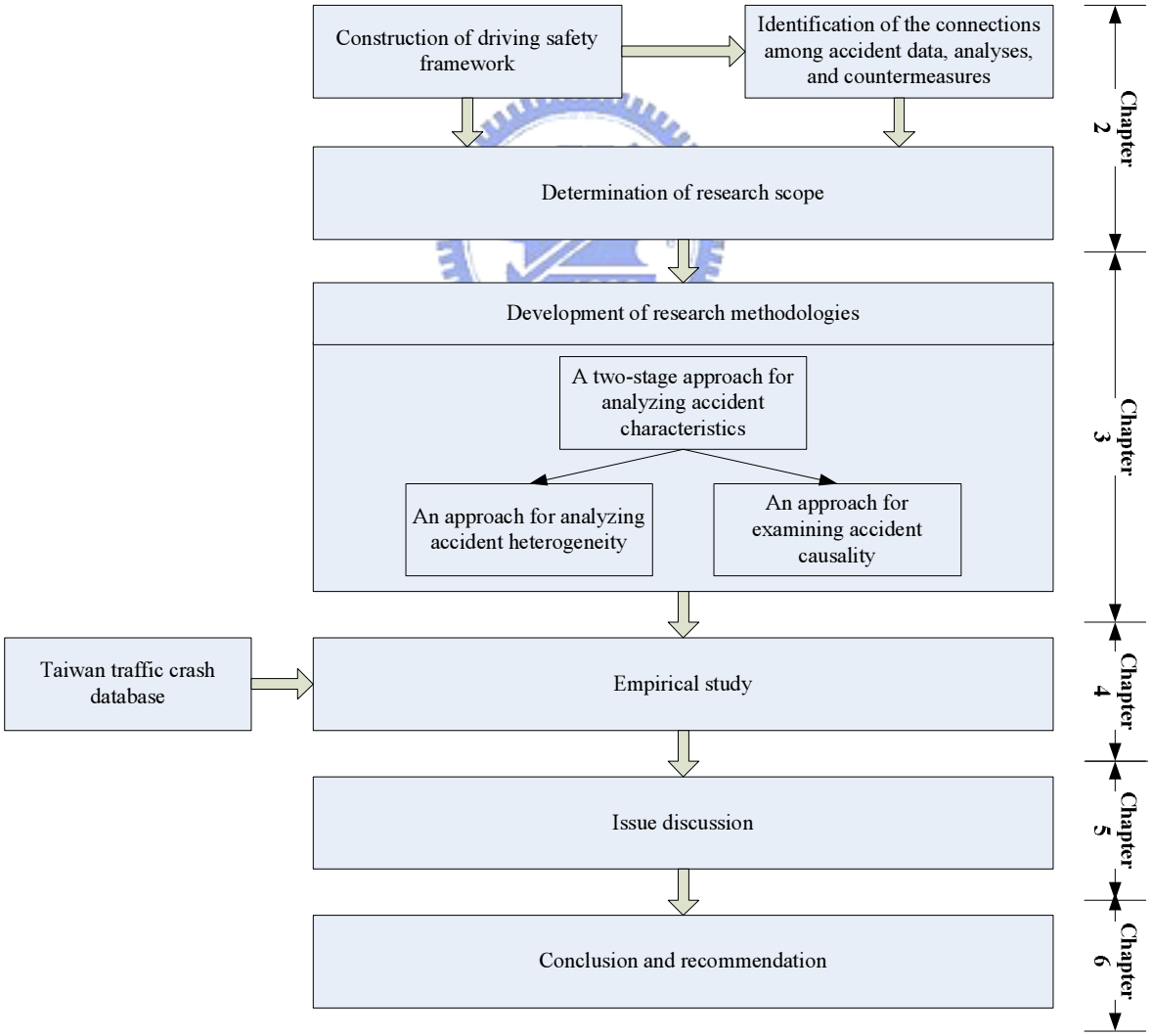


FIGURE 1-1 Research framework.

Chapter 2 CONCEPTUAL FRAMEWORK OF DRIVING SAFETY

The aim of this chapter was to build a conceptual framework which explains the generating process of accidents from chain perspective. The necessity and advantages of applying the chain concept on analyzing and preventing accidents were revealed from the built framework.

2.1 Coverage of Driving Safety

There has been some research proposing frameworks and models to explain driving behavior and its connections between accidents. Few of them were built from the chain perspective but focused on a certain issue (Elvik, 2003; Juarez *et al.*, 2007); others interpreted driving behaviors yet usually put most of their attentions merely on the driving stage (Fuller, 2005; Sümer, 2003; Wilde, 2001). In this study, some of the cases were extended and integrated as a more general conceptual framework of driving safety.

To understand the causes of an accident, analyzing only the behaviors at the driving stage is not enough. Juarez *et al.* (2007), for example, proposed a multilevel model to prevent death among minority young drivers from motor vehicle crashes. They suggested that effective prevention should cover the whole driving processes instead of focusing merely on the driving stage. The whole driving process includes the prior-to-driving environment factors, the driving behaviors, and the crash outcomes. In particular, the prior-to-driving environment factors are those which may affect the young driver's choice on seat belt use or vehicle choice. Fleury and Brenac (2001) also suggested analyzing accidents through looking into the whole driving process. They proposed to analyze accidents at five stages: the situation prior to driving, the driving situation, the discontinuity situation, the emergency situation, and the collision situation. The conditions of one stage are affected by its previous stage and affect its subsequent stage. Both researches indicate that driving behaviors as well as the occurrence of crashes should not be fully determined by local factors, i.e. only factors at the driving stage. Consequently, the construction of the chain framework should be built first from the factors prior to driving until the factors representing the end of the event.

Numerous factors are involved in the chain. Some research proposed to explicitly partition them into several stages such as Fleury and Brenac (2001); other research, however, such as Juarez *et al.* (2007) and Sümer (2003) who presented a contextual mediated model which divided factors into distal and proximal context, did not. Fleury and Brenac (2001) proposed to divide factors into a distinctive five stages since their approach was proposed to conduct an in-depth study; therefore, detailed and required information for each stage would

be collected. On the other hand, Sümer's approach (2003) was to analyze the relationship among personality, driving behaviors, and accidents. Since the focus was put on linking the connections between psychological factors and resulting driving behaviors, only two levels of connections were represented (i.e. the connection between psychological factors and driving behaviors, and the connection between behaviors and accident outcomes) although psychological factors could affect the activities prior to driving and then affect the driving behaviors. In brief, the partitions of factors along the chain should depend on the available data and the purposes of the analyses. Nonetheless, the clearer sequential connections are the factors, the more solid the results.

It is assumed that our proposed framework is to be adopted in the research with an accident database. An accident database usually consists of three types of data: person, vehicle, and accident characteristics. Although the sequences for all factors can not be fully determined, a rough partition can be achieved. For example, mode choice must be made prior to driving. Therefore, the numerous factors provided by an accident database can be divided at least four stages: prior to driving, driving, incidents or accidents, and rescue.

2.2 Construction of Driving Safety Framework

At the prior-to-driving stage, the decision of the trip characteristics is the critical factor, affecting safety-related trip characteristics like when to drive, which route to take, or whether to take passengers or not – should be considered. Elder drivers, for example, are found to develop more driving strategies than youngsters (Eby *et al.*, 2000; Simoes, 2003). The strategies include not driving after dark, reduce going on freeways, driving only in familiar areas, planning routes where protected left turns can be made and driving with a co-pilot; all of which fit to compensate their physical impairment (Eby *et al.*, 2000; Simoes, 2003). Therefore, the age factor should be represented at this stage. With similar deduction, numerous factors can be found at the prior to driving stage. To organize these factors, the multilevel model proposed by Juarez *et al.* (2007) is adopted and modified. The trip characteristics are mainly determined by four sets of factors: driver characteristics, vehicle characteristics, local laws and enforcement, and passenger characteristics. Of which, driver characteristics are further affected by social context, national/regional culture, family, and peers; driver and vehicle characteristics are both further affected by public polices such as driver education and the required safety equipments. The necessity of these factors at the driving stage has been declared by Juarez *et al.* (2007).

The relationships between factors at the driving stage and those at the incident/accident stage have been intensively studied. Some research focused on analyzing individual driving behaviors with respect to behavioral or social sciences such as Wilde (2001), Sümer (2003)

and Fuller (2005); other research put the focus on measuring the effects of particular factors on accidents such as traffic flow, surface conditions, enforcement, etc. Although the involved factors are numerous, they can be roughly divided into three types: driver characteristics, vehicle characteristics, and environment factors (Kim *et al.*, 1995). To simplify our framework, all the factors at the driving and the incident/accident stages are represented in these three sets. Of which, the environment factors are further divided into local driving conditions, such as traffic, weather, light and enforcement, and the transportation infrastructure, such as the set up of speed limit, stop signs, surface condition, etc.

The last stage goes to the rescue stage. The factors at this stage are rarely discussed. The focus would be put on the response of emergency service. Detailed discussions will be given in the subsequent chapter.

All the factors and relationships are illustrated as in Figure 2-1. The proposed framework is constructed in two dimensions: the time dimension, and the factor interaction dimension. This framework represents that the occurrence of accidents is dynamic, and the factors are interacted at each stage. Moreover, four nodes (three dotted circles and one dotted star labeled “Crash”) are drawn to collect the effect of the interactions resulted from the aforementioned factors. The last dotted star represents the accident outcomes resulted from the accident chain. In addition, the dotted line connecting the five dotted nodes imply that the effects conducted at one stage would accumulate and affect the subsequent stages either immediately, intermediately, or in a long run. In addition to the age factor, another example regarding the vehicle’s characteristics at the prior-to-driving stage is the choice of cars. It is clear that the choice of cars would affect the driving behavior at the driving stage in terms of, for instance, whether the driver is familiar with the car, and also affects the accident severity at the accident and rescue stage in terms of, for instance, the compatibility of collided vehicles. Obviously, with the proposed framework, the accident generating processes can be more correctly identified and interpreted. Thus, research results based on the framework should be more convincing.

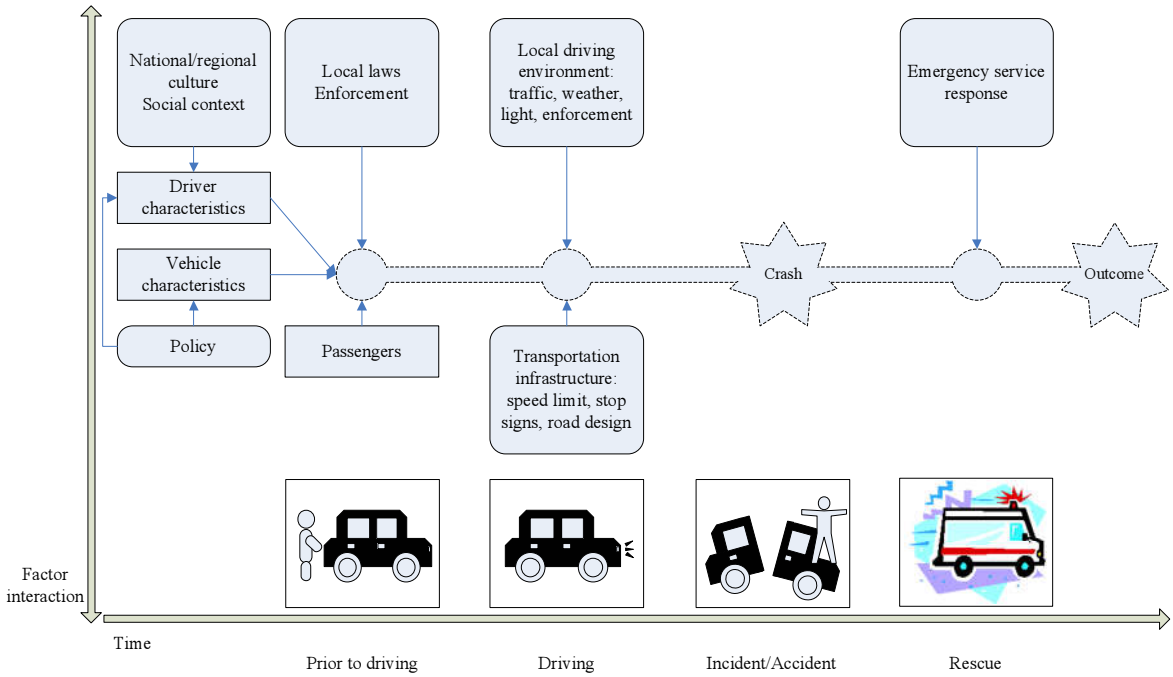


FIGURE 2-1 Conceptual framework of driving safety.

2.3 Risk Factors

Numerous factors have been studied for their relationship with accidents which are not possible to give a complete discussion in a study with limited length. Therefore, no attempt is made to provide a complete coverage of all possible risk factors. Instead, the aim is to introduce the representative factors and organize them at stages in the proposed framework.

2.3.1 Prior to Driving

Drivers are going to decide their driving plan at this stage including which route to go, which vehicle to use, what time to start the trip, and expected time to end the trip. These decisions are usually affected by the driver characteristics, vehicle characteristics, passenger characteristics, local laws, and enforcement.

Of the interactions between these factors, studies related to driver characteristics have grabbed most attentions. For example, different age groups would show different decision characteristics. Older drivers would like to develop strategies such as stopping night driving or finding co-pilots to compensate their declining ability to cope with complex traffic situations (Eby *et al.*, 2000; Simoes, 2003). Yet, young drivers were found relating to alcohol use and seat belt nonuse which would like to increase the accident risks (Ferguson, 1996).

Different enforcement schemes would also affect the driver's decisions on making trips. Mountain *et al.* (2005) claimed that speed management schemes can affect route choice and this can have a significant effect on accidents within the scheme.

For policy factors, the licensing procedures have particularly significant impacts on trip decisions. The licensing procedures for young drivers are concerned since their immature driving skills and they tend to seek risks. The effect of restrain the licensing procedures for young drivers, such as delaying privilege licensure, imposing night driving curfews, and extending periods of supervised practice of driving, have been found positive effect in many areas (Ulmer *et al.*, 2000). Therefore, different policy settings would affect the amount of traffic exposure and the way showing up on roads for different types of drivers.

2.3.2 Driving

1. Relationships between factors and accidents

The risk factors related to the driver characteristics, vehicle characteristics, and passenger characteristics are the first three classes of factors introduced, followed by the factors related to local laws, enforcement, and policy.

- Driver Characteristics

  Many factors related to driver characteristics have been considered as connecting to crashes. The socio-demographic factors have been the most intensively studied factors. Of which, age and gender are the two factors which have been particularly extensively studied. Younger drivers are argued to have high rates of crash involvement due to inexperience in assessing traffic situations. The over-representation in accidents for young drivers is partly due to the lack of driving experience (Williamson, 2003); another possibility attributes to young drivers' risk-taking behavior (Murray, 1997). Yagil (1998) surveyed 693 male drivers in the Israeli army with questionnaire; he found that young drivers are more likely to violate the law than older drivers either from instrumental motives (such as perceived danger of punishment from violations) or from normative motives (such as a sense of obligation to obey the law).

  As for the gender factor, some literatures found that male drivers have higher accident rates and result in severer accidents than female drivers do. Male drivers tend to be involved in fatal accidents since their risk-taking behaviors and attitudes; such behaviors include speeding and alcohol consumption. On the contrary, accidents related to female drivers are usually nonfatal due to their immature skills (Massie *et al.*,

1995, 1997; Laapotti *et al.*, 1998). While these observations for male drivers have been consistent in decades, those for female drivers are doubted because of a continuously increasing number of license holders and higher exposure on roads for female drivers than before (Kim, 1995; Forward *et al.*, 1998 and McKenna *et al.*, 1998). Laapotti *et al.* (2004) claimed that generally, male drivers are risk seeking while female drivers are risk aversion. Moreover, the immaturity in driving skills for female drivers directly relates to possibility and types of accidents although the skill differences between male and female drivers may have declined.

In addition to socio-demographic factors, factors such as psychological and situational factors would affect the occurrence and consequence of accidents as well. Psychological characteristics are very crucial for risk-taking preference and relate to traffic accidents. With observational studies, psychological characteristics are found significantly related to drivers' socio-demographic factors. Mizell (1997) found the majority of aggressive drivers are relatively young, poorly educated males who have criminal records, histories of violence, and alcohol problems. Shinar and Compton (2004) also found that men were more likely than women to commit aggressive actions. Furthermore, drivers' psychological characteristics are significant to accidents as well. Beirness *et al.* (1993) found that the crash-group display a low degree of self-confidence than the non-crash group; Gulian *et al.* (1989) found that poor self-esteem and high hostility formed a particularly lethal combination.

Situational factors include transient factors and personal habits. The former indicates the factors that may increase risk contributing to states of fatigue, distraction, irritability, and self-doubt (Norris *et al.*, 2000) while the latter refers to personal life habits affecting the occurrence of accidents such as drinking habits. Moskowitz and Fiorentino (2000) found that the impairment resulted from alcohol consumption include divided attention, drowsiness, decreasing vigilance, increasing reaction time, etc. Most of the studies found that male or younger drivers have significant relationship with alcohol-related accidents (Harrison, 1997; Abdel-Aty *et al.*, 2000; Keall *et al.*, 2005). As for the fatigue factor, it contributes accidents by deteriorating drivers' alertness, by impairing their judgment, and by slowing their reactions (Lyznicki *et al.*, 1998). As reviewed by Stutts *et al.* (2003), drivers' sleep habits and work pattern have been found significantly related to accidents. Night or rotating shift workers and commercial vehicle operators have significant relationships with accidents. Unlike the alcohol and fatigue factors, there is still uncertainty for the contribution of drugs and illness to accidents (Drummer *et al.*, 2004; Hansotia *et al.*, 1991).

- Vehicle characteristics

In the past, most concerns have been put on the relationships between accident severity and vehicle types as well as the protection equipments. Elvik *et al.* (1997)

found that the overall injury accident rate of heavy vehicles is nearly the same as for passenger cars, but accidents involving heavy vehicles more often result in fatalities or serious injuries than accidents involving passenger cars only. Three fundamental differences between heavy vehicles and passenger cars are found by Abdel-Aty (2004) including: mass incompatibility, stiffness incompatibility, and geometric incompatibility. In particular, the geometry incompatibility, i.e. the imbalance in ride height, would cause significant impact while the collision type is on the frontal (sight reduction) or side (intrusion into smaller vehicles).

Advanced safety vehicle (ASV) has recently become a popular way to avoid accidents. Of which, the installation of intelligent driving support systems aims to help drivers recognize the road environment correctly, warn drivers while errors occur, guide driver's maneuvering, or to proceed with automatic driving. The main feature of the system is to provide safety-related information to drivers to avoid incidences. Yet, only the right information provided at the right place and at the right time can bring positive effect on reducing accident risks. Inappropriate information style or too much information may cause information overloading or drive drivers to distraction (Yamada and Kuchar, 2006). Moreover, when drivers decrease their speed in response to the warning messages, they tend to raise their following speed to compensate the loss of time (Boyle and Mannering, 2004).

- Passenger characteristics

The presence of passengers may provide positive effects on accident prevention. Vollrath *et al.* (2002) found that the presence of passengers could provide a general protective effect; however this is not found for young drivers especially for driving during darkness, in slow traffic and at crossroads.

The seating position of passengers affects the passenger death and injury in traffic crashes. Glass *et al.* (2000) found that motor vehicle occupants are at a lower risk of death or non-fatal injury when riding in the rear seats of passenger vehicles as compared with riding in the front seat. Similar results are found by Smith and Cummings (2004).

- Environment characteristics

Numerous factors related to the environment factors may affect the occurrence of accidents and its consequences. These factors include road design and road furniture, road maintenance, traffic control, weather, and flow conditions. Interested readers can refer to the book by Elvik and Vaa (2004) which gives a very thorough discussion of these factors, except the last two, via systematic overview and meta-analysis.

The weather factors, in addition to their relationships between road factors, may affect drivers' cognition process. For example, one lagged effect of precipitation over

days was discovered in Eisenberg's research (2004); that is, if it rained a lot yesterday, then on average, today there are fewer crashes. This may come from the adaptive behaviors by drivers.

As for the flow factors, a significant relationship between crashes and mean speed and variation of speed has been found (Garber and Ehrhart, 2000; Golob *et al.*, 2004). The complexity of information perceived by drivers is higher and the predictability of traffic situation for drivers may be worse while the mean speed and variation of speed increases. Note that this relationship may not be linear since drivers may pay more attention on flow situation while it gets more complicated.

- Regulation and policy characteristics

  The stricter the rule enforcement the more drivers would comply with the rules. However, this is not necessary for all drivers. Yagil (1998) found that the young drivers' instrumental motives, which are a reaction initiated by a desire to avoid punishment or to receive positive rewards, are weaker than older drivers'.

2. Integration of factors

The driving behavior has been characterized by constantly solving problems that involve thinking, choosing and deciding between different alternatives (Vaa, 2001). Several models have been proposed; interested readers can refer to Fuller's study (2005) for a thorough review.

The Risk Homeostatic Model (RHM) proposed by Wilde (2001) was adopted as the basis in this research to connect other risk factors. Assumed all drivers would have a target level of risk which comes from the perceived costs and benefits of action alternatives. By comparing with the driver's perceived level of risk, the driver would tend to adjust his driving behaviors to achieve the target risk. The benefits and costs of action alternatives are obtained from either comparatively risky or safe behaviors. After the driving alternative is taken place, that would be lagged feedback to the driver that may increase or decrease his perceived level of risk or cause an accident. This simple and intuitive structure can accommodate these factors discussed above as illustrated in Figure 2-2 where the dotted box is RHM.

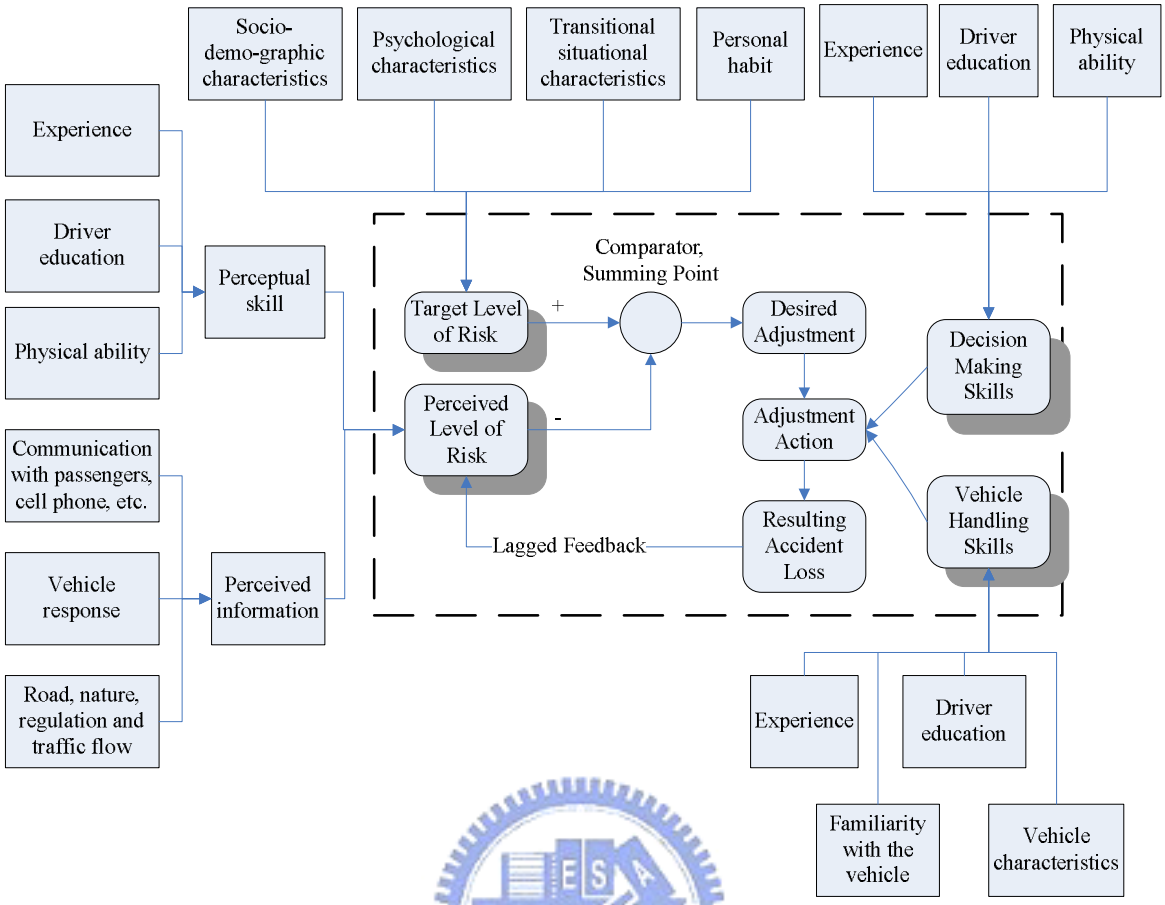


FIGURE 2-2 Interactions between drivers and risky factors.

The decision making skills are mainly based on the driver's experience, driver education and physical ability. With more driving experience, the driver is expected to make a decision closer to his desired adjustment more precisely and quickly. Moreover, when with good driver education, the driver is expected to have better sense to make a right decision and thus perform better in decision making skills. This skill is also affected by the driver's physical ability. For example, the reaction time for a drunk driver is longer.

The vehicle handling skills are affected by the driver's experience, familiarity with the vehicle, driver education and the vehicle characteristics. With more driving experience, the driver is expected to handle the vehicle better. Yet, this would be affected by his familiarity with the vehicle. The driver may not be able to handle the vehicle well if unfamiliar with the car. Moreover, when with good driver education, the driver is expected to perform better in vehicle handling. The vehicle handling skills are also affected by the vehicle functions. For example, driving a truck is more difficult than driving an automobile.

The perceived level of risk is based on the driver's perceptual skills and perceived information from the passengers, the vehicle and the environment. The perceptual skills are affected by the driver's experience, driver education, and physical ability. With more

driving experience, the driver is expected to be more sensitive to perceive the necessary information. For example, the experienced driver is expected to be able to perceive the necessary information from high speed flow than novice drivers. Moreover, when with good driver education, the driver is expected to be more sensitive to catch important information and thus perform better in perceptual skills. The driver's physical ability would also affect his perceptual skills such as spatial contrast sensitivity, color perception and visual field.

The driver's perceived information comes from communications, the vehicle and the environment. When the driver and passengers talk to each other, or the driver uses cell phones, the driver has to handle more complicated information. Moreover, the interaction between the driver and the car is another source of information. Some information is directly revealed from the vehicle equipments such as speedometer, thermometer, etc.; other information comes from the driver's control and the vehicle's response such as kinetic energy and friction. Furthermore, information also comes from the environment. The critical information generating from road environment includes horizontal and vertical alignment, degree of curvature, gradient, access control, speed limits, road markings and signs, etc (Proctor *et al.*, 2001). The information tells drivers the road condition and helps drivers adjust their behavior. The weather condition is also important since it would affect the driver's visual ability and vehicle movement. Therefore, natural light and rain condition is critical for drivers, and wind and snow information for some special areas.

Two types of flow information are critical to drivers: one is flow factors and the other is flow compatibility. As discussed in previous sections, mean speed and variation of speed are two major indexes to accidents. The driver needs to deal with much more information and response more quickly while the flow speed is high and fluctuate considerably. On the other hand, the flow with high mixed types of road users gives more information to drivers than the flow with low mixed types of road users.

The regulatory information reminds and warns the driver to obey the rules; different enforcement schemes provide different information to drivers. For example, the response for drivers may be to slow down the vehicle when seeing an automated photographic speed detector; however, they may also slow down their car to see what happened when seeing the police.

The target level of risk is affected by the driver's factors. The critical factors include: socio-demographic factors; psychological factors; transitional situational factors; and personal habit.

2.3.3 Incident/Accident

This stage describes a discontinuous situation within the road safety system such as the driver falls asleep or a sudden stop of the previous car. When a collision happens, a good response of the driver may be able to mitigate the severity. For example, a driver loses the control of the vehicle since the surface is iced; an experienced driver would brake the car gradually rather than immediately. Moreover, he can also take a suitable position to protect himself while a collision happens such as hold his head in the arms. The driver's action is affected by his experience, driver education and physical ability. With more experience, good driver education and physical ability (e.g. shorter reaction time), the driver is expected to mitigate the severity of the collision and protect himself well.

The protection equipments of a vehicle and the compatibility of collided vehicles would affect the severity of a collision. The equipments, such as seat belts, airbags and anti-lock brake systems, can protect the driver and occupants to some degree from a collision. On the other hand, the compatibility of the collided vehicles would affect the severity of a collision due to the mass incompatibility, stiffness incompatibility and geometric incompatibility as discussed on the previous section. Those incompatibilities depend on the vehicle types and bumping positions. The severity of a collision can be alleviated when the road design concerns about safety such as installations of safety fencing.

2.3.4 Rescue

An efficient emergency response provides better service to save the injuries. The efficiency of an emergency response depends on the distance between collision position and service providers, and the flow conditions.

2.4 Accident Data in Driving Safety Analysis

Comprehensive data provide solid bases to understand and model accident causality. Therefore, transportation engineers and professionals have devoted themselves to collect as much data as possible. As stated previously, numerous factors at each stage would affect the occurrence of accidents or their severity. Although it would be practically impossible to collect perfect data contents, with the improvement of technology and data collection methodologies, more accurate and comprehensive data have been trying to be gathered.

2.4.1 Crash-Centered Data

Crash-centered data include the data surrounding the occurrence of accidents, which usually include six types: crash information, roadway information, vehicle information, driver information, citation/adjudication information, and injury control information (Ogle, 2007).

1. Crash information

Crash data describe the information of events, vehicles, and persons involved in a crash. General characteristics include the date, time, location, drivers, occupants, and vehicles involved. Other categories are severity of the crash (whether the crash ended in property damage only, an injury, or a fatality) and the type of collision (single or multi-vehicle, pedestrian involved or not, etc.). The conditions of the roadway surface and of traffic control devices are also important aspects of crash data (NHTSA, 2003). Crash data in Taiwan are usually gathered by police departments. Hospitals also have the responsibility to report a death or injury to police departments as long as patients go into a hospital due to car accidents.

2. Driver information

Driver information includes information about the licensed drivers. It may include: driver license number, type of license, license status, driver restrictions, convictions for traffic violations, crash history, and driver education data. This type of information is maintained by motor vehicle supervision offices, Directorate General of Highways, MOTC in Taiwan.

3. Citation/Adjudication information

Citation and adjudication information is also vital for describing driver characteristics. Information may include the identification of the type of violation, location, date and time, the enforcement agency, and so on. Motor vehicle incidents that would reflect enforcement activity are also useful for traffic safety purposes (NHTSA, 2003). This type of data is usually maintained by police agencies in Taiwan.

4. Vehicle information

Vehicle information includes information on the identification and ownership of vehicles registered in the country. This information should also be available for commercial vehicles and carriers. Data contents may include vehicle make, model, year of manufacture, body type, and miles traveled in order to produce the information needed to support analysis of vehicle-related factors. In Taiwan, motor vehicle supervision offices play the role to supervise such data. Insurance companies also own such information.

5. Roadway information

A system of roadway inventory is a collection of roadway characteristic data. It usually includes a list of the roads along with roadway location, identification, and classification. In addition, the inventory contains a physical description of the roadway components, such as alignment, number of lanes, lane width, presence of medians and shoulders, and type and presence of roadside barriers. Photograph/video-log data may also be a part of the roadway inventory (NHTSA, 2003). In Taiwan, this type of data is collected and maintained by different agencies. For example, highway information is primarily maintained by National Expressway Engineering Bureau, and the city road is mainly maintained by local governments.

Except roadway inventory information, traffic conditions on the roadways are also important and needed to be gathered. It may be collected manually or by means of automatic traffic recorders. In Taiwan, loop detectors have been installed in highways and expressways; recently, closed-circuit television and the system of electronic toll collection (ETC) become another useful ways to monitor traffic congestions and to detect possible incidents.

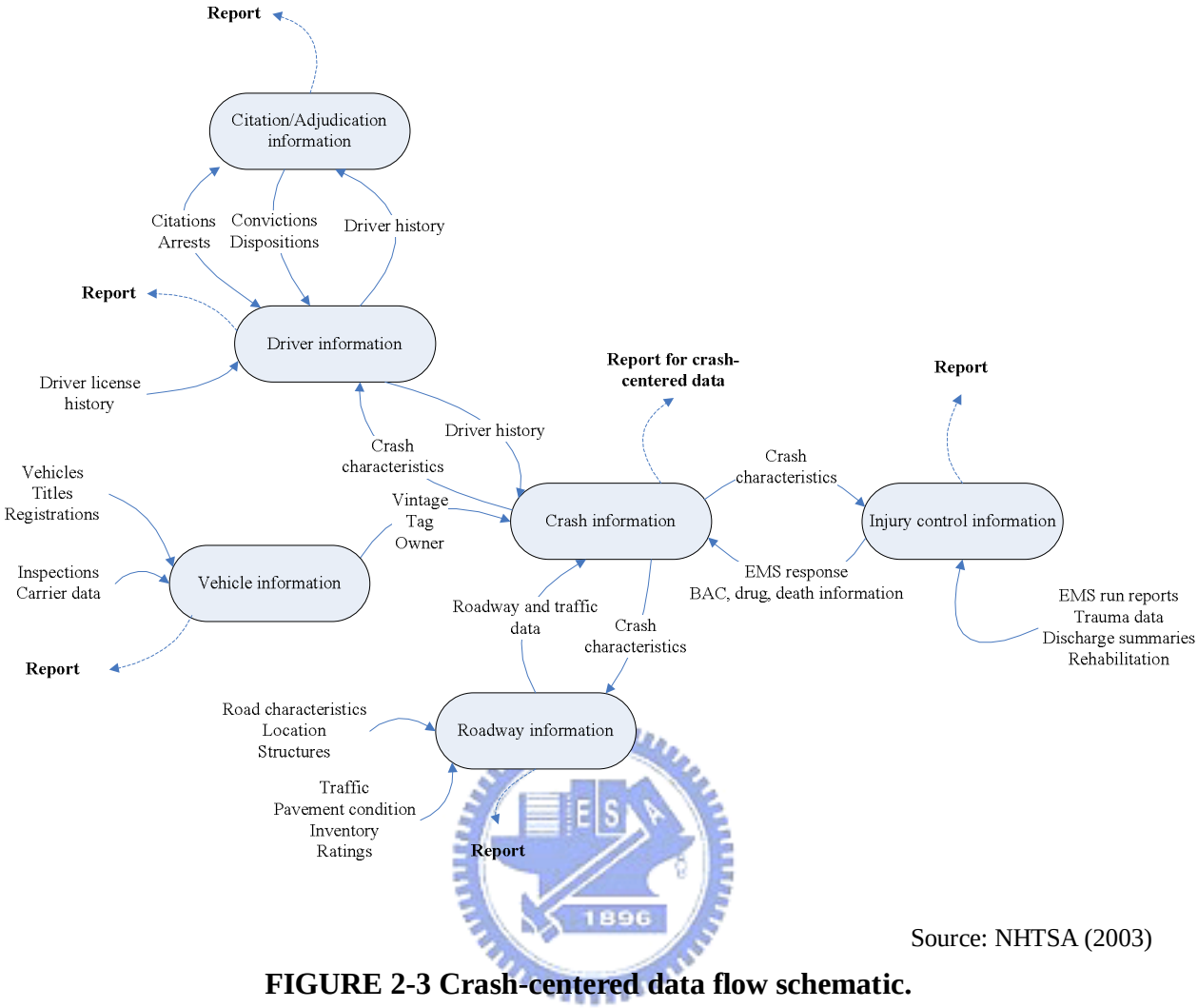
6. Injury control information

Injury control information refers to the information tracking injury causes, magnitude, costs, and outcomes. When the injury causes come from traffic incidents/accidents, such a case would be of interest. This type of information could be maintained by public health sectors such as Bureau of National Health Insurance or hospitals in Taiwan.

The input, output, and interrelationships among the aforementioned data types were depicted in Figure 2-3. This schematic was rearranged from the one proposed by NHTSA (2003) to map with the proposed conceptual framework of driving safety (Figure 2-1).

By comparing Figure 2-1 and Figure 2-3, it could be observed that except driving culture and government policies, other factors have been covered in Figure 2-3. Since the analysis of crash is the primarily interested outcome, crash information plays the central role in the schematic. Driver information, vehicle information, roadway information, and injury information have close connections with crash information while citation/adjudication information is connected to driver information.

The ellipse in the illustration implies an information processing center. For example, the input of vehicle information includes vehicles, titles, registrations, inspections, and carrier data. After data cleaning, integration, and analysis by the ellipse of vehicle information center, it outputs useful information such as vintage, tag, or owner information to crash information. Meanwhile, vehicle information itself also produces its own report.



Source: NHTSA (2003)

FIGURE 2-3 Crash-centered data flow schematic.

2.4.2 Behavior-Centered Data

Except crash-centered data, driving behavior data have been also concerned as an important source of data researching driving safety. While crash-centered data are usually collected by specific agencies or sectors, behavior-centered data are usually gathered by researchers or engineers according to their research or project of interest. This type of data includes all the data required to explain driving behaviors including unobservable factors such as driver's psychological factors.

One of the many ways to collect such data is using questionnaires; that is, collect information of interest based on self-rating methods. Different types of questionnaires have been developed. The very first driving behavior questionnaire was developed by Reason et al. (1990). They outlined 50 different abnormal driving behaviors and collected 520 samples in England. By using exploratory factor analysis, they concluded that driving behaviors could be roughly divided into three types: errors, deliberate violations and harmless

mistakes. This questionnaire had been followed by many researchers such as Blockey and Hartley (1995), Parker and Reason (1995), and Sullman et al. (2002). In addition to driving behavior questionnaire, researchers also explored driving behaviors from different perspectives. For example, Gulian et al. (1988) and Guilan et al. (1989) developed the so-called driving behavior inventory to measure the relationships among driving stress, aggregation and alertness. French et al. (1993) developed driving style questionnaire to examine the behaviors related to accident involvement and risky driving behaviors. One also can find other questionnaires scaling driver's other characteristics such as driver's attitudes to violations (West and Hall, 1997) or driver's vengeance intensions (Wiesenthal et al., 2000). Recently, Taubman-Ben-Ari et al. (2004) have tried to synthesize past studies into a multidimensional driving style inventory.

While the aforementioned questionnaires focused on driving behaviors, the other class of questionnaires concerns more about the psychological factors affecting behaviors such as attitude and perception. One of the most well known applications is the Theory of Planned Behavior, or TPB, developed by Ajzen (1985). The theory stated a structure that behavior is determined by driver's intention, and intention is determined by attitude, subjective, and perceived behavioral control. Researchers had applied this theory to evaluate the intentions committing some risky behaviors such as speeding or driving and drinking.

Except questionnaires, simulators are the other powerful tools for researchers to collect driving behaviors. In particular, driver's situation awareness would be the area which had adopted simulators to collect driving behavior information most frequently. Bolstad (2000) adopted simulators to explore whether driver's situation awareness would be significantly different with respect to age. Ma and Kaber (2005) evaluated the impact of using navigators and cell phones on situation awareness with simulators. Kass et al. (2007) developed different scenarios on simulator to measure the influences of using hand-free cell phones between novice and experienced drivers.

In addition to the above approaches, researchers also try to collect practical driving behavior information with in vehicle data recorders (IVDR). The first application of vehicle data recorders is event data recorder. Similar to the "black box" equipped in aircrafts, event data recorder began to be installed in vehicles in 1970's to record technical vehicle and occupant information for a brief period of time (seconds, not minutes) before, during and after a crash. For instance, EDRs may record (1) pre-crash vehicle dynamics and system status, (2) driver inputs, (3) vehicle crash signature, (4) restraint usage/deployment status, and (5) post-crash data such as the activation of an automatic collision notification (ACN) system (Ogle, 2007). More recently, recorders have also been used to study driver behaviors in non-crash situations. For example, IVDR has been adopted in the trucking industry to monitor and improve driving safety in the last twenty years (Toledo et al., 2007).

2.4.3 Other Data

In addition to the aforementioned two types of data, other data could also contain information regarding the occurrence of accidents and/or severity, and require being collected for a complete analysis of driving safety such as law enforcement or related regulations and laws.

2.5 DISCUSSION

This chapter contains a review on literature for potential factors and their possible connections to accidents. The review ended in the construction of a conceptual framework of driving safety, which consists of two dimensions: the interactions of potential factors and driving stages. It represented the belief that the occurrence of an accident results from a series of miserable or unfortunate events. This construction did not intend to reproduce all possible types of accident occurrences. Instead, the built framework was treated as a blueprint for the following analyses.

To adopt safety data into analysis, the issue of data quality should be born in mind. The first element of data quality is comprehensiveness. Collecting data has been a time consuming and high cost task; instead of full information, researchers could obtain only partial information. They should be aware of what they have and have not collected; more importantly, what the role of the collected information plays in the driving safety framework. The second element is timeliness of the data; that is, how quickly safety data are available and updated for use. The third element goes to the accuracy of accident data. This refers to how close the recorded accident characteristics to truth. For example, not all traffic accidents are reportable; not all reportable accidents are reported; not all reported accidents are correctly recorded. Researchers must be aware of the existence of measurement bias when applying accident data. The last element is the integration of accident data. Accident data come from many sources representing different levels of population. It would be a challenge to integrate all information and produce useful knowledge. Even though the information within the same level, how to correctly adopt them and explore reliable results is still a big challenge in safety research fields.

Safety data analyses are of many types such as before and after evaluations, cross-section evaluations, comparison group evaluations, analysis of collision trends, identification of hazardous locations, collision rate comparisons of locations with different features, cost-benefit analysis in development of countermeasures, risk estimation/analyses/ evaluations, and questionnaire-, simulator-, and video-based driver safety evaluation (Persaud, 2001). While some types of these analyses are specific to roadway safety, others

relate to driver or vehicle safety; and some analyses cover all three aspects.

Followed by accident data analyses, corresponding countermeasures could be designed and implemented. Based on analysis results at different levels, countermeasures focusing on different coverage of populations are developed. One should be careful to design countermeasures when applying analysis results. For example, when inferences about the nature of individual accidents are based solely upon aggregate statistics collected for the group to which those individuals belong, the so-called ecological fallacy would generate.

Due to the availability of data, this research adopted only crash-centered data, in particular, the traffic crash database maintained by National Police Agency, Ministry of the Interior. Accordingly, this research was a cross-sectional study. The data represent the whole population in Taiwan.

Chapter 3 METHODOLOGY

The purpose of this chapter was to introduce the methodologies applied in this research. Section 3.1 discussed the challenges and opportunities faced by today's traffic safety analysis; related literature was reviewed in Section 3.2. The methodologies designed specific to such data were developed in the subsequent sections, t. A two-stage approach for analyzing such data was introduced in Section 3.3. The primary method employed in this research was presented in Section 3.4. Moreover, an approach to analyzing the heterogeneity of accident data was proposed in Section 3.5, followed by an approach examining accident causality in Section 3.6. This chapter ended in a discussion in Section 3.7.

3.1 Challenges in Accident Analysis

The definition of causality is strict. In epidemiology, for example, the Surgeon General (1964) claimed that to diagnose cancer of smoking causes, the following ad hoc rules for judging causality could be adopted: 1) Strength of association (meaning some statistical measure of association is strong); 2) Dose-response effect (the more of the causal factor, the larger the effect); 3) No temporal ambiguity (disease follows exposure to risk factor); 4) Consistency of findings (several studies produce similar results); 5) Biological plausibility (the hypothesis makes sense in view of what is known in biology); 6) Coherence of evidence (some combination of 4 and 5); and 7) Specificity (causal factor causes this disease, and this disease is due to this causal factor). Some of these rules are deficient if being directly applied in traffic safety. Rule 2, for example, is not necessary true in traffic safety: empirical evidence shows that the relationship between expected accident frequency and traffic flow is usually not linear†. Yet, most of them are desirable (or just need a few modifications) in traffic safety including rules 1, 4 and 5. A more concise definition of causality is given by Pearl (2000) who asserted that causality has to meet three criteria: 1) Correlation: Cause and effect must vary together; 2) Time sequence: The cause must come before the effect; and 3) Non-spurious: The relationship between cause and effect cannot be explained by any third variable. These criteria can be viewed as the baseline for all kinds of causality including traffic safety.

Factual knowledge of causality is not easy to come by. The best way to obtain causality is via randomized experiments. Yet, it is technically impossible and immoral to do so in traffic safety research. Another two ways are observational before-after studies and cross-section studies. An observational before-after study is to randomly divide a set of

† Golob et al. (2004) gave a complete review on their published article, Freeway Safety as a Function of Traffic Flow, in *Accident Analysis and Prevention*, Vol.36, No.6, pp.933-946.

candidate entities into those to be treated and those not prior to the implementation of some effect. After a certain period of implementation, the differences between treated and untreated groups are compared. On the other hand, an observational cross-section study arises when the attributes and accident history of entities (such as road sections, intersections, drivers, etc.) are used in an attempt to estimate the safety effect of the difference in treatment (or attribute) in question. Observational before-after studies have been demonstrated being able to explore correct insights under a meticulous study design (Hauer, 1997) while the capability of observational cross-section studies still opens to question (Hauer, 2006).

Since observational studies, whether before-after or cross-section, are not as robust as randomized experiments in causal-effect interpretations, inconsistent or even controversial conclusions are sometimes found in reports or journal articles. For example, Davis (2004) mentioned that although many studies have used statistical methods to correlate accident experience with variations in traffic and road conditions, the transferability of such models have been found that the significance of accident predictors can differ for data collected in the same geographic region but at different times, as well as for data collected in different regions. In another example, Elvik and Greibe presented the result of a meta-analysis (2005) for the studies evaluating the road safety effects of porous asphalt. They concluded that "While some studies have evaluated these effects, not all of these studies can be trusted and their findings are highly inconsistent." These inconsistencies mainly result from four difficulties: the existence of confounding factors, the determination of scope of causality, the quality and availability of data, and the capability of methodologies.

The leading and the most important difficulty comes from confounding factors. A confounding factor is any exogenous (i.e. not influenced by the road safety measure itself) variable affecting the number of accidents or injuries whose effects, if not estimated, can be mixed up with effects of the measure being evaluated. The results of a study should never be trusted if confounding factors are not well controlled (Elvik, 2002). Factors that are commonly regarded as potential confounding factors in observational before-after studies include: long term trends affecting accident consequences; general changes of the number of accidents from before to after the road safety measure is introduced; any other treatments that have been implemented during the 'before' or 'after' periods; regression-to-the-mean‡; adjustments to the reportability limit; and traffic flow (Hauer, 1997). Confounding factors of accidents are abundant and various such that a well control over them becomes very difficult. This reflects in the following three difficulties.

‡ "The entities may have been chosen for treatment because they had unusually many or few accidents in the past… one can hardly hope that the 'unusual' is a good basis for predicting what would be expected in the future had treatment not been applied." Hauer (1997), pp.74.

Since confounding factors are numerous, an immediate issue raises: how to define the scope of the causality of an accident; i.e. which factors should be considered and which should not. In early days, the causes of an accident were usually attributed to the closest-to-accident factors. Researchers, however, have recently tended to analyze an accident more thoroughly – not only the accident itself but also the activities prior to and subsequent to the accident. For example, Eby et al. (2000) and Simoes (2003) found that elderly people tend to avoid night driving, reduce freeway driving, driving only in familiar areas, and driving with a co-pilot to compensate for their age-related decline and the corresponding difficulties in performing the driving task. An accident, therefore, may not occur if one or several undesirable activities in this accident chain were broken (Baker and Ross, 1961). An analysis of accident chains can be roughly divided into several stages, for example: the situation prior to driving, the driving situation, the accident or discontinuity situation, the emergency situation, and the collision situation (Fleury and Brenac, 2001). It is obvious that the driving situation, such as pavement material, illumination, traffic signals, etc., would affect accident occurrence, but the activities in other stages are difficult to recognize whether they have impacts on accident occurrence and/or severity.

The other concern on the selection of contributing factors is the use of statistical null hypothesis significance testing (NHST for short). Recall the first rule to define causality claimed by Surgeon General (1964): some statistical measure of association is strong. NHST has been regarded as a good measure to define the importance of factors. However, a 'not significant' factor in statistical sense is not equal to a 'not important' or 'useless' factor in traffic safety. A fair way to say about a non-significant factor is: "I cannot be sure that the safety effect is not zero". Since a 'non-reject' null hypothesis is of scarce help on dropping potential factors and it is expected that the farther a factor away from an accident (such as factors in the prior to driving stage), the more insignificant a factor would be, it becomes more and more difficult for researchers to choose factors via NHST in research.

The third difficulty goes to the availability and quality of data. Although most accident databases have been designed to contain as much information as possible, some attributes such as driver's psychological status are still difficult to discover except in some in-depth investigation projects. Thereafter, even though an accident case is fully described with all the recorded data, it is an incomplete description for the case. Furthermore, although accident databases are panel data, i.e. data of same targets are collected over some periods, the targets are usually defined by administrative areas such as city and county rather than specific intersections, road segments or specified populations. Moreover, not all traffic accidents are reportable; not all reportable accidents are reported; not all reported accidents are correctly recorded. With these deficiencies, the availability and quality of data is questionable. This problem exists in many countries including Taiwan (Lai et al., 2006).

Assume data has been screened where confounding factors are all considered; potential causal factors are determined; and the quality of data is assured. The last difficulty goes to the capability of analytical techniques. Statistical methodology has been the most frequently one to be adopted on analyzing accident data. Conventional statistical methods, such as logistic regression models, are great for analyzing relationships which are clear between dependent and independent variables. Moreover, few 'representative' variables are usually chosen to interpret dependent variables. The conventional statistical approach is great to explore relationships but would be inappropriate to examine causality since the complicated interrelationships among factors are difficult to be well controlled.

3.2 Literature Review of Crash-Centered Data Analysis

Some of the aforementioned challenges have been tackled. A very original technique is called *crash type analysis* developed by Snyder and Knoblaunch (1971) and applied to urban pedestrian crashes. A crash type analysis is mainly done manually. Trained analysts are asked to read crash reports and conclude crash types. For example, Preusser et al. (1995) asked one analyst who developed a preliminary set of crash type groups and preliminary definitions to review half of the computer generated crash reports. A second analyst then reviewed the preliminary group definitions. Cross-reviewing selected cases from each other, the two analysts together finalized the crash type definitions and made final crash type assignments for the total crash events. Ten simple crash types were defined including: ran off road, ran traffic control, oncoming, LT (left-turn) oncoming, motorcyclist down, run down, stop/stopping, and road obstacle. It was found that the five defined crash types accounted for 86% of all of the motorcycle crash events studies. This approach is easily implemented; however, it is labor intensive and time consuming. Moreover, since the capability of man brain, only single or few factors could be accounted simultaneously to determine crash types; i.e. only the most significant factors would be considered. Yet, this is counterintuitive to the contemporary theory of accident occurrence: the occurrence of accidents is a series of miserable or unfortunate events; as long as one or some those events are blocked, accidents would not occur (Baker and Ross, 1961; Davis and Swenson, 2006; Elvik, 2003; Fleury and Brenac, 2001; Heinrich, 1931; Reason, 1997).

In order to conquer those deficiencies, techniques which can consider multiple variables are developed and adopted. Two types of techniques have been applied to analyze the relationships between factors and accidents. One is traditional statistical techniques. In previous studies, logistic regression (Al-Ghamdi, 2002; Kim and Kim, 2003; Chandraratna et al., 2006), factorial analysis of correspondence (FAC) combined with hierarchical ascendant clustering (HAC) (Laflamme and Eilert-Petersson, 1997; Berg et al., 2004), and

entropy classification methods (Strnad et al., 1997; Vorko and Jović, 2000) are the most frequently applied techniques; yet, these techniques can contain only a limited number of variables. Unobserved heterogeneity was ignored and accident cases were treated as with complete information. Of these techniques, FAC combined with HAC is the only one appropriate to include abundant explanatory variables, while the other two approaches use a few "representative" explanatory variables in the analysis. As a consequence, some typical accidents are not well discovered and effects of the specified variables are improperly magnified.

The other types of techniques are artificial intelligence (AI) and data mining. The techniques of this category have become very popular recently due to the improvements in computer power. Some well-known techniques such as classification trees and neural networks have been adopted in accident research (Delen et al., 2006; Karlaftis and Golias, 2002; Sohn and Shin, 2001). Classification trees such as CART and C4.5 are top-down techniques which decompose accident data by loading explanatory variables sequentially. The top layer consists of input nodes (i.e. accident data). Decision nodes determine the order of progression through the graph. The leaves of the tree are all possible outcomes or classifications, while the root is the final outcome (for example, accident types). Neural networks are nonlinear techniques which mimic the operations of human brains and have been regarded as great techniques for prediction accuracy. In summary, traditional techniques are very efficient at solving problems with simpler relationships among explanatory variables with a continuous domain. However, most AI and data mining techniques can reflect the complicated relationships among numerous explanatory variables but they are usually of black-box type that is less helpful in interpreting accident causality.

To claim causality, one has to evaluate the relationships between factors and consequences with rigorous criteria, such as Pearl's three criteria: correlation, time sequence, and non-spurious relationships. Although classification methods are powerful to explore the complicated relationships between influential factors and consequences, they can not automatically determine the time sequence and non-spurious relationships. Accordingly, factors with significant classifying ability do not necessarily imply causality. For example, Clarke et al. (1998) presented a decision tree of onto accidents to classify injury levels. Season turned out to be the factor with the most classifying power. Yet, season might not be the closest-to-event factor, and some other factors might exist between season and injury level.

All in all, the methodologies analyzing crash-centered data have been evolved in the last 30 years. One could employ the state-of-the-art methodologies to explore the correlations of factors involved in accidents. However, to improve our understanding on accident causality, another approach is needed.

3.3 A Two-Stage Approach for Accident Chain Analysis

This study proposed a simple two-stage approach for exploring accident characteristics and uncovering accident causality based on crash-centered data. The proposed approach consisted of two steps: The first step was to classify accidents such that accidents belonging to same classifications are under the condition that most critical features are identical. The second step was to verify the causal relationships from classification results. The relevant methodologies were introduced in the following.

3.3.1 Classification

The classification step is expected to relieve the abundant heterogeneity existing among accidents. Heterogeneity represents the possible presence of unobserved or inattentively accounted driver-, trip-, area-, road-, and other-specific factors (Karlaftis and Tarko, 1998). Unless heterogeneity is appropriately controlled, the estimation results and causality interpretations can be trusted. The adoption of classification techniques can classify accidents into sets with relatively homogeneous attributes. Instead of a whole dataset, sub-datasets are analyzed and less heterogeneity effects are expected.

With the emergence of computational power, the applications of data mining techniques have become very popular including the traffic accident analysis and prevention field. The avoidance of pre-specified functional forms and the ability to simultaneously handle multiple factors may be the two most attractive features to adopt such methodologies (Chang and Wang, 2006). These advantages are particularly useful in adopting the proposed framework since the more the important risk factors are under control, the more homogeneous the results of classifications.

The primary two types of classification techniques in accident analysis are tree-based and rule-based classification techniques. The tree-based techniques are to sequentially break down a whole dataset into smaller and smaller sub-datasets such that the sub-datasets at the deepest nodes are of the least heterogeneity. The sequence of factor loading depends on the choice of classifiers. Common classifiers include entropy, Gini coefficient, Bayesian, etc. The differences of applying different classifiers on analyzing accidents are usually decided by their prediction accuracy while the entropy classification was popular in earlier research (Vorko and Jović, 2000). On the other hand, a relatively new technique, named classification and regression tree, becomes another popular choice. This technique can automatically search for the best predictors and the best threshold values for all predictors to classify the target variable, and has been shown a useful tool to effectively identify the risky factors affecting injury severity of traffic accidents (Chang and Wang, 2006).

The rule-based technique is another classification technique in traffic accident analysis and prevention. This type of techniques is to learn rules first from a given dataset; thereafter, the accidents in this dataset are classified based on the derived rules. Some common ways to learn rules from a given dataset include Apriori (Geurts, et al., 2003), neural networks (Tseng, et al., 2005), genetic algorithm (Clarke et al., 1998), etc. Recently, the use of the rough sets theory becomes another alternative to classify and analyze accidents. Its non-parametric and non-black-box type process enables the theory to become attractive in exploring the features of accident occurrences.

In short, the first step of traffic accident analysis and prevention from the chain perspective is to classify accidents into relatively homogeneous groups with multiple factors. Consequently, each group represents a specific type of accident conditions described by driver, vehicle, trip and environment characteristics, driver's behaviors, and accident consequences. However, the classification techniques can not identify the sequential relationships between factors which are required to interpret causality. To obtain more accurate causal relationships, another methodology is required.

3.3.2 Causal Inference

The causality between factors and accident consequences is not easy to verify since most accident analysis and prevention are observational studies rather than experimental studies. Studies adopted conventional statistical techniques such as logistic regression are difficult to verify all these elements except the correlations between factors. To overcome these problems, researchers have been proposing many causal inference models. Of which, the model proposed by Pearl (2000) was concerned as a particularly useful tool and has been applied in some traffic accident analysis (Davis and Swenson, 2006).

To construct a causal model, one needs to identify a set of exogenous variables, a set of endogenous variables, and for each endogenous variable a structural equation describing how that variable changes in response to changes in the exogenous and/or other endogenous variables. This possible causality is represented by a directed acyclic graph. Events are defined in terms of values taken on by the model's variables. Knowledge of these values will almost always be to some degree uncertain. To allow for uncertainty, Pearl (2000) defined a probabilistic causal model as a causal model augmented with a probability distribution over the values taken on by the model's exogenous variables, so that this probability distribution determines the probabilities to be assigned to the truth or falsity of counterfactual propositions. The probabilities attached to counterfactual statements can be computed by augmenting the model with nodes reflecting the counterfactual situation, and then applying algorithms for computing Bayesian updates on graphical models.

3.4 Rough sets Theory

Among the AI and machine learning techniques, rough sets theory was chosen to be implemented at the first stage rather than other techniques because of the following reasons: First, the algorithms for rough sets theory are explicit and easily understood which makes rough sets theory be preferred than some black-box type methods such as neural networks. Second, unlike classification trees which must consider all factors sequentially, rough sets theory can consider all factors either simultaneously or sequentially. It is convenient for researchers to deal with some factors where they are unsure about their occurrence order. Third, rough sets theory is non-parametric, so it avoids issues such as pre-specified function forms or multi-collinearity among independent variables as in traditional statistics or membership functions in fuzzy theory. Fourth, rough sets theory can effectively handle discrete variables with multilevel categories. Thus, it is believed that rough sets theory is a suitable method for analyzing relationships among factors and accidents under considerations of the process of accident occurrence.

Rough sets theory was proposed by Pawlak in 1982 and has been shown to be an excellent mathematical tool for the analysis of objects with incomplete information (Greco et al., 2001). Although accident databases have been designed to contain as much information as possible, they can not provide full information describing the occurrence of an accident.

Let U represent the universe, a finite set of objects and P denote a set of condition attributes, i.e. affecting factors for the occurrence of accidents. For example, five accident cases (U) described with four attributes (P) – driver's age, vehicle type, climate and accident type – are given as Table 3-1. For $x, y \in U$, x and y are indiscernible by the set of condition attributes P if $\rho(x,q) = \rho(y,q)$ for every $q \in P$ where $\rho(x,q)$ denotes the information function. A set that has objects within it that are indiscernible by the set of condition attributes P is called a P-elementary set. The family of all elementary sets is denoted by P^*. It represents the smallest partitions of objects by the specified condition attributes so that objects belonging to different elementary sets are discernible and those belonging to the same elementary sets are indiscernible. The P-lower approximation of a set of objects Y ($Y \subseteq U$), denoted by $\underline{P}Y$, and the P-upper approximation of Y, denoted by $\overline{P}Y$, are defined as:

$$\underline{P}Y = \bigcup X \quad \{X \in P^* \quad and \quad X \subseteq Y\}$$

$$\overline{P}Y = \bigcup X \quad \{X \in P^* \quad and \quad X \cap Y \neq \varnothing\}$$

The objects belonging to the set of lower approximation are those definitely definable by the elementary sets since objects in $\underline{P}Y$ can be fully identified by the elementary sets in P^*. On the other hand, those belonging to the set of upper approximation but not to the set of lower approximation can not be fully identified by the elementary sets in P^*.

TABLE 3-1 Example of Accident Cases with Describing Features

Case	Driver's age	Vehicle type	Climate	Accident type
1	Young	Motorcycle	Sunny	Off-road
2	Old	Automobile	Sunny	Off-road
3	Young	Motorcycle	Sunny	Rollover
4	Middle-aged	Motorcycle	Sunny	Rollover
5	Middle-aged	Automobile	Rainy	Rollover

Accident case 1, for instance, is characterized by the following statement:

The (off-road) accident is described by the following attributes:
(driver's age is young) and (vehicle type is motorcycle) and (climate is sunny).

The above statement is termed a *rule* in rough sets theory. The term in the first parenthesis is called a decision attribute which is the concept of concern, and the following attributes are all termed condition attributes which is the observed information. In this example, there are two concerned concepts, namely, off-road accident types and rollover accident types. Five cases are provided with three condition attributes characterizing them. The three condition attributes, *driver's age*, *vehicle type* and *climate* form four elementary sets – {1,3}, {2}, {4}, {5}. This means that case 1 and 3 are indiscernible while the other cases are characterized uniquely with all available information. Since case 1 and 3 are indiscernible and lead to different accident types, they are termed boundary-line cases representing those can not be properly classified with the available information. Therefore, the off-road accident type is described with the lower approximation set, {2}, and the upper approximation set, {1,2,3}. Similarly, the concept of the rollover accident type is characterized by its lower approximation set, {4,5} and upper approximation set, {1,3,4,5}.

Sometimes, some particular condition attributes can not distinguish objects; they are redundant. The condition attributes excluding redundant attributes are termed *reduct* in rough sets theory. One possibility for the redundancy could be that the condition attribute has the same value for all objects and is invariant; the other possibility is that its value can be substituted by values of other condition attributes or their combinations of Boolean relations.

The performance of the specified condition attributes can be described with two indicators: accuracy of approximation and quality of approximation. Accuracy of

approximation represents the percentage of the associated objects definable with the specified condition attributes. It can be defined as follows:

$$\pi_P(Y) = card(\underline{P}Y) \Big/ card(\overline{P}Y);$$

where *card* refers to cardinality. The accuracy value ranges from 0 to 1. The closer to 1 is the accuracy, the more discernible is the accident type; i.e. more accident cases of this accident type are discernible by the elementary sets generated by the specified condition attributes. It implies that the associated accident patterns do exist unambiguously.

On the other hand, quality of approximation represents the definable percentage of the whole universe. Let $X = \{Y_1, Y_2, ..., Y_n\}$ be a classification of U, i.e. $Y_i \cap Y_j = \varnothing$, $\forall i, j \le n$ $i \ne j$ and $\bigcup_{i=1}^{n} Y_i = U$. Y_i are called classes of X. The P-lower approximation and P-upper approximation of X are represented by sets $\underline{P}X = \{\underline{P}Y_1, \underline{P}Y_2, ..., \underline{P}Y_n\}$ and $\overline{P}X = \{\overline{P}Y_1, \overline{P}Y_2, ..., \overline{P}Y_n\}$, respectively. Quality of approximation of classification *X* by a set of attributes can be defined as follows:

$$\gamma_P(X) = \sum_{i=1}^{n} card(\underline{P}Y_i) \Big/ card(U);$$

The value of quality ranges from 0 to 1. The closer to 1 is the quality, the more objects of the universe clearly belong to a single class of X. This implies that the accident chains for all accident types can be clearly identified. Accidents thus can be more accurately recognized and the corresponding countermeasures be devised.

To recognize further the details of accident patterns, *rules* need to be extracted. A rule is a combination of values of condition attributes. Therefore, the theoretical maximum number of rules is the product of the categories of all condition attributes. However, some combinations may not show up since such accident patterns have never happened before. A rule exists if and only if at least one such accident exists. Many rule generation algorithms have been proposed in recent years (Greco et al., 2001), but it is beyond this research's scope to discuss those algorithms. This research simply applies the most frequently used algorithm – minimum covering – to generate rules. Its aim is to generate the minimum number as well as the shortest length of rules to cover all accidents.

Rough sets theory, as introduced, is a non-parametric approach which prevents the pre-specification of function forms or membership functions which are usually difficult to determine in accident research since the interactions and relationships among attributes are too complicated and uncertain. It allows researchers to adopt accident attributes as many as

possible; moreover, any redundant attributes will be discarded based on the definition of reduct. With all non-redundant attributes, the minimum covering principle is applied to generate rules. Those rules describe distinct accident scenarios for different accident types. It should be noted that although accidents belonging to the same rule are treated as being identical, accidents belonging to slightly different rules are not essentially different since some of the considered critical non-redundant attributes are overlapping.

3.5 Analyzing Heterogeneity of Accident Data

As stated previously, the heterogeneity of accident data plays a vital role in examining accident characteristics and designing countermeasures. The objective of this section was to propose an approach for analyzing the heterogeneity of accident data based on rules derived from rough sets.

For the purpose of accident analyses and prevention, people have been interested in causality and have tried to find the generating processes of accidents, especially for those that occur repeatedly. The occurring frequency of a rule is termed as rule strength in rough sets theory. A rule with high frequency of accident occurrence indicates that many accidents repeatedly occur under identical conditions for some critical factors. Consequently, strong causality between factors and outcomes may exist for such rules. On the other hand, a low-frequency rule refers to only a few accidents, occurring under the associated conditions. Accidents belonging to the same rule are treated as identical; however, it should be noted that accidents belonging to slightly different rules are not essentially different since some of the considered critical attributes could be partially overlapping in terms of the effect on accident occurrence. For example, trip time and illumination of roads both affect drivers' sight distance and consequently the occurrence of accidents. Therefore, to avoid over-strictly classifying accidents, instead of rules, the classification of accidents will be based on rule strength which stands for the occurring frequency of such accident conditions. Accidents associated with the rules with low-occurring frequency could be considered as by-chance accidents. On the other hand, accidents under rules with high-occurring frequency may imply that they did not occur by chance but for some reason or system error such as poor road design. These accidents should be paid more attention by both policy makers and traffic engineers. Therefore, the rule strength is considered as a helpful indicator to cluster accidents for further analyses.

As stated, the proposed approach consisted of two stages. In the first stage, accidents were grouped with respect to rule strength; accordingly, accident characteristics were extracted with multinomial logistic regression in the second stage. In the following, the proposed approach is explained step by step. The first four steps describe how to apply

rough sets theory and statistical tests to group accidents while the last step describes a way to use a multinomial logistic regression model in extracting accident characteristics. The whole process is depicted as in Figure 3-1.

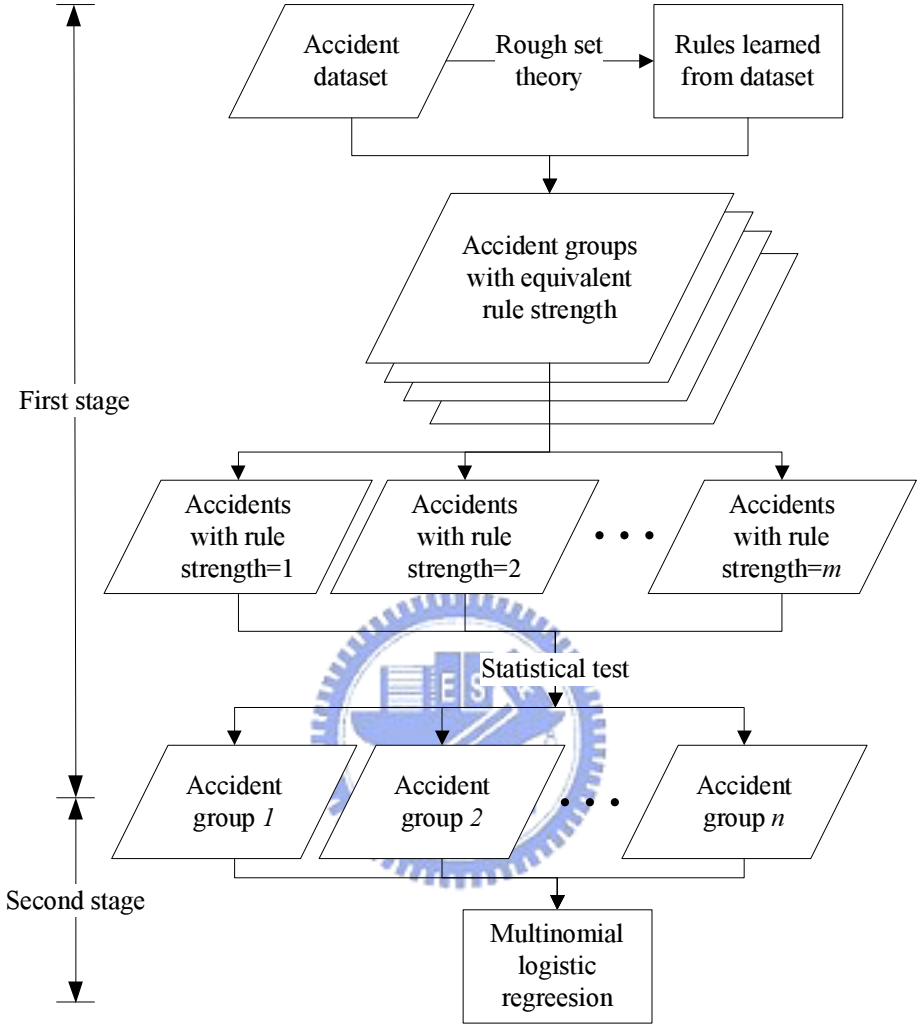


FIGURE 3-1 Framework of analyzing heterogeneous accident data.

- Step 1: Learning rules from accident datasets

  A whole accident dataset was first analyzed with rough sets theory. Condition attributes were filtered so that the attributes unable to distinguish accident cases were excluded. Thereafter, by learning from past accident cases, a minimum number of rules was generated to represent all distinct accident patterns. Each rule was represented by three elements: *variable combination of condition attributes*, *strength* and *belonging accidents*. *Combination of variables* describes the process of accident occurrence for a specific accident pattern. *Strength* represents the accident counts belonging to a rule, and *belonging accidents* refers to the accident cases under the rule.

- Step 2: Grouping accident cases based on rule strength

Accident cases were then grouped according to the associated rule strength. In other words, accident cases were grouped if their belonging rules were of equivalent strength. Consequently, two accident cases were put under the same group if and only if their belonging rule had equivalent strength. Accidents referring to distinct rules could belong to the same group as long as their strength was equivalent.

- Step 3: Ranking the aforementioned groups by the order of rule strength
  Rules and the corresponding accidents were then arranged in the order of strength.

- Step 4: Grouping the ordered accident groups
  The next step was to group the ordered accidents. For the convenience of interpretations, the number of the groups was set small. Meanwhile, the accident characteristics among groups were expected to be significantly different from one another where a χ^2 test was adopted for large sample sizes and a Fisher's exact test for small sample sizes in the significance test.

- Step 5: Exploring accident characteristics with multinomial logistic regression
  Finally, multinomial logistic regression was applied to explore the accident characteristics for the whole dataset as well as for each accident group. The characteristics of each accident group were then compared.

3.6 Examination of Accident Causality

The continuous expansion of accident databases and improvement of computing ability, however, provide the opportunity to explore causality. By controlling as many affecting factors as possible, accidents could be classified into subsets with very similar conditions. Therefore, comparing the features of these subsets would reveal the differences between what happened and what would have happened had the circumstances in question been different (Davis, 2004; Hauer, 1997), which might imply causal relationships. In addition, since an accident database can never contain sufficient factors for characterizing the occurrence of all types of accidents, it would be unreasonable to regard all the accidents as with complete information in a database. Therefore, for those accidents with insufficient information, instead of soft computing classification methods, other methods could be advantageous to analyze them.

3.6.1 Framework of Accident Causality Examination

The research framework consists of two stages as shown in Figure 3-2. The first stage is to identify the circumstances contained in an accident database. To fully describe the

circumstances, all available information should be considered such as driver characteristics, trip characteristics, vehicle information, behavioral information, and road and environmental factors. In order to accommodate the numerous factors, soft computing methods such as tree- or rule-based classification methods are preferred; in particular, rough sets theory was adopted in this research. Interested readers can refer to Pawlak (1982) and Pawlak and Skowron (2007) for a thorough introduction about rough sets theory. In addition, a nice tutorial about rough sets theory was presented by Walczak and Massart (1999).

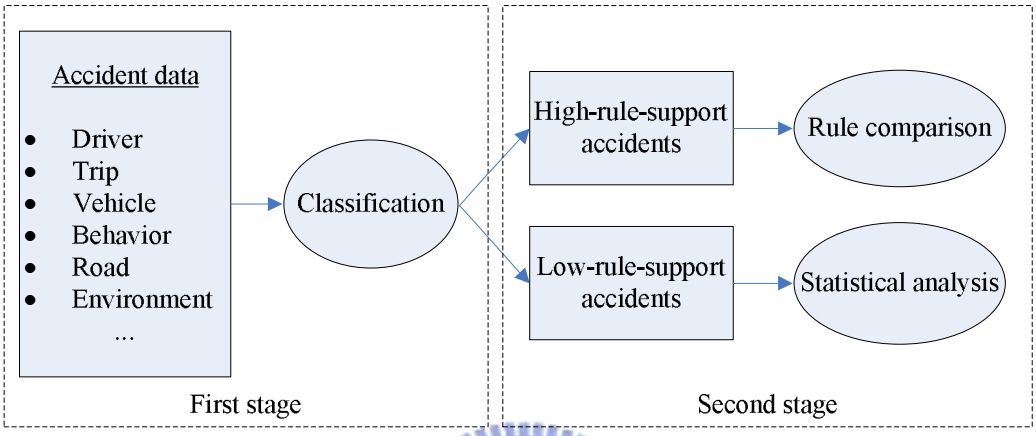


FIGURE 3-2 Framework of accident causality examination.

As a classification methodology, rough sets theory generates rules to identify the differences among accidents. Since each rule implied the indispensable circumstances under which accidents with specific injury levels occurred, the injury level would be different if one or several indispensable circumstances were different. Therefore, comparing the rules with high support offered the potential to understand the causes of accidents, and was the focus in this study.

Based on the classification results, it is possible to compare the rules and find potential causal factors, especially for those accidents that frequently appear. However, two difficulties should be noticed. First, the available information is unable to differentiate all accidents. Some accidents under identical circumstances may lead to different results. This mainly results from insufficient information. Second, even if accidents could be clearly distinguished, some rules may show extremely low frequency of occurrence (the frequency of occurrence is called support in rough sets theory). These low-support accidents may occur by chance (bad luck), and causal relationships between factors and accident consequences may not exist. Accordingly, these accidents and the corresponding rules would be inappropriate for rule comparisons. Instead, statistical analysis such as regression models would be more appropriate to catch the features of these low-support accidents. That is, using error terms to represent the insufficient information and the randomness. The problem now is how accidents can be distinguished between the accidents suitable for rule

comparisons and those suitable for statistical analysis. The choice of the threshold should result in a satisfactory performance on post-validity evaluations or predictions.

3.6.2 Procedure of Accident Causality Examination

The subset with accidents of high rule support was adopted for rule comparisons. The comparisons composed of two steps: the first was to find the most similar rules for each selected strong rule (i.e. a rule with support of at least six) from the remaining strong rules; the second was to check if the accident severities were different between the selected rule and its most similar rules. In the following, an example of rule comparison was provided.

Suppose a rule, denoted as the selected rule, was chosen from the rule set. This rule described a particular circumstance for SAV accident occurrence: A female driver with a valid driver license driving on a low-speed-limit road (less than 50 kph) with seat belt fastened but without specific trip purposes. The SAV accidents under such circumstances were of the type – injury only. If the specified attributes were changed (e.g. from female to male), the result was different (i.e. from injury only to death involved or to other). Other represents the accident severity of approximate rules, which can be injury only or death involved. It is noted that some condition attributes were specified, but others were not. The severity does not alter even though those unspecified attributes change. For instance, whether a driver was young, middle-aged, or old, the severity of the SAV accidents under the circumstance described by the selected rule would remain the same.

Based on the selected rule, its similar rules were searched. A similar rule is defined as the rule which has the greatest number of identical specified attributes to the selected rule. Two similar rules were found. Similar rule 1 described the condition that a middle-aged driver with a valid driver license, with seat belt fastened, cell phone not-used but without specific trip purposes driving on a low-speed-limit (less than 50 kph) road equipped with roadside marking and illumination. Similar rule 2 described the condition that a young male driver with a valid regular driver license, with seat belt fastened but without specific trip purposes driving on a low-speed-limit (less than 50 kph) straight road with dry surface and equipped with median marking but without signals at midnight.

Both similar rules had only one indispensable attribute value different from the selected one. This could be verified by expanding the unspecified attributes of the selected rule to match its similar rules. As shown in Figure 3-3, the attributes age, cell phone use, road shape, roadside, and illumination of the selected rule could be expanded to be identical to those of the similar rule 1 without affecting the accident severity of the selected rule. By comparing the expanded rule and similar rule 1 (the upper right table in Figure 3-3), it could

be observed that only the attribute gender was different where the expanded rule specified it as female but was unspecified in similar rule 1. Similarly, the same expansion could be done to compare the selected rule and similar rule 2: the attribute gender was also the only distinct one between these two rules (the lower right table in Figure 3-3).

Rule 1 pointed out that a male driver's accident severity was greatly reduced if he was mature (middle-aged and driving without using a cell phone) and driving on a friendly road environment (with roadside marker and illumination). Rule 2 pointed out that young male drivers' driving on an unfriendly environment (a not safety-oriented designed road at midnight) could be fatal. This result implied that the combined attributes (age + gender + road environment) might be critical factors diverting an injury only case to a death involved case under a circumstance described by the selected rule.

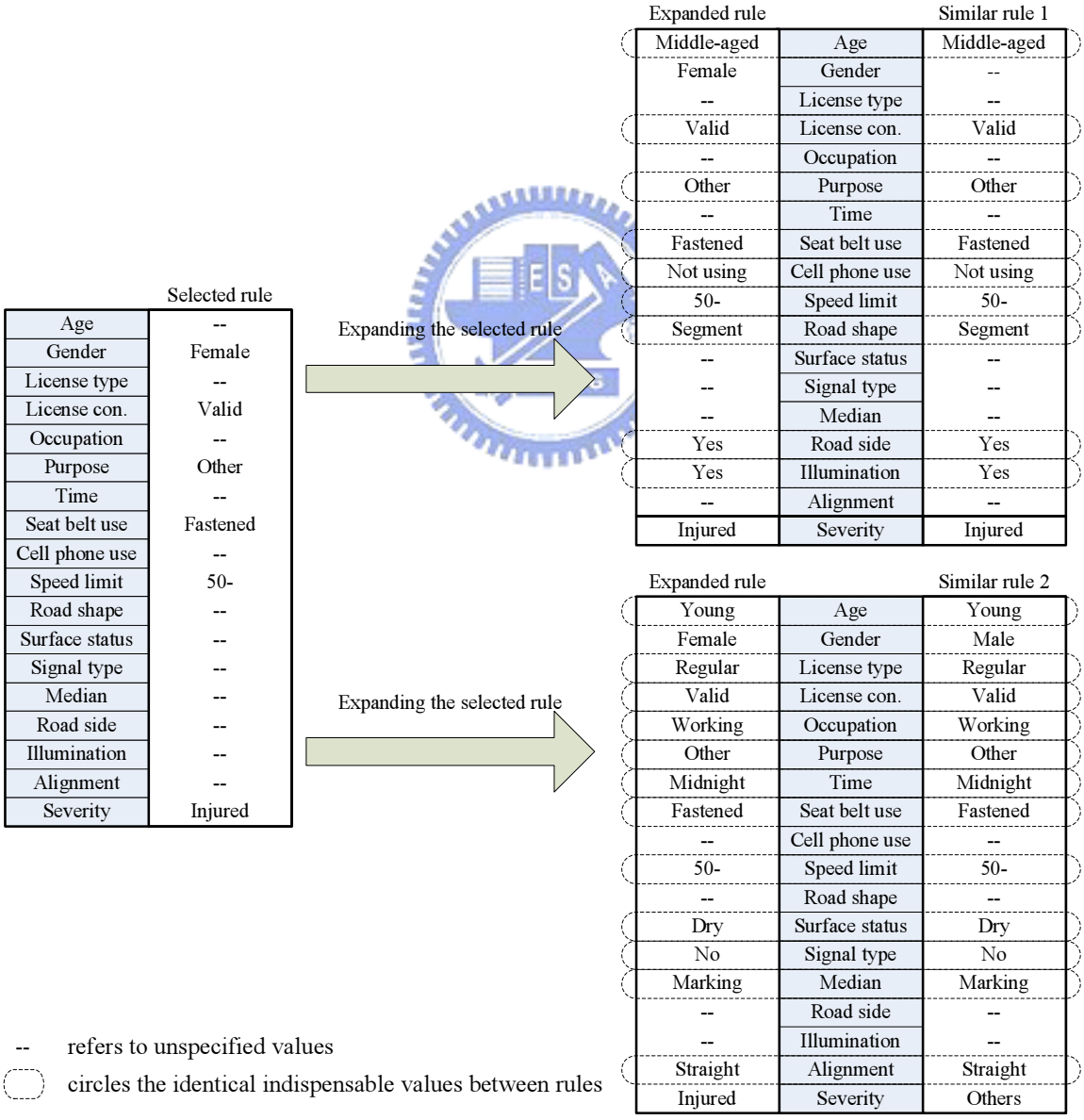


FIGURE 3-3 Framework of accident causality examination.

3.7 Discussion

Analyzing accidents from the chain perspective could capture the nature of the traffic accidents; the generation of an accident is coming from a series or a combination of activities. This study proposes to implement such an idea starting from classifying accidents from an accident database, and then infer the causality for each classification. Although methodologies and databases have been available and continuously improved, the factual knowledge of accident causality is still not easy to come by. Several issues are worthy of consideration.

The first issue is the robustness of the classification results. Each derived classification represents one type of causal chains. Accidents belonging to same causal chains suggest their accident occurrences are similar. In other words, provided that most important risk factors are considered for classifications, accidents coming from same causal chains should be bound together almost surely. Yet, some techniques, such as CART, have relatively unstable classification results; when different adoption strategies, such as stratified random sampling, are applied, the tree structure and the classification accuracy would alter significantly (Chang and Wang, 2006). Therefore, one should be very careful to choose an appropriate classification technique and the adopted strategies.

The subsequent difficulty lies on how to define which factors are important. Analyzing accidents from the chain perspective has the potential to overcome the confounding-factor effects; yet, the researchers should consider most important factors. However, defining the so-called numbers of important factors containing in an accident database is difficult; moreover, factors are not always important for all types of causal chains. A conventional way to select contributing factors is the use of statistical null hypothesis significance testing (NHST for short). NHST has been regarded as a good measure to define the importance of factors. However, a non-significant factor in statistical sense is not equivalent to an unimportant or useless factor in traffic safety (Hauer, 2004). Moreover, the relationships verified in one place may not hold in another place due to the differences of national or regional culture. Consequently, it would be extremely difficult to correctly specify the relationships between factors and accident consequences purely based on literature and professional knowledge. One possible way to relieve this problem is to examine the location of factors in the proposed framework. When a factor locates at earlier stages, such as driver characteristics, it has more potential to be adopted to interpret more types of causal chains. Another way goes to the use of well-behaved data mining techniques such as rough sets theory; however, the appropriateness should be further verified.

Different from the issue to select important factors from an existing database, the other issue is to collect vital information which is absent from on-hand databases, especially

information of some important indirectly observable or measurable factors. These types of information are not considered in the proposed approach. However, it is possible to collect this required information by experiments.

The next issue is the regression to mean (RTM) phenomenon. When classification is done, each causal chain contains one or numbers of accidents. One causal chain has more accidents than the other should not be immediately claimed that one is a more dangerous condition than the other. A causal chain with higher number of accidents may reveal by chance. To eliminate the RTM phenomenon, several approaches have been proposed such as statistical quality control or the adoption of empirical Bayes method (Elvik, 2006). However, how to integrate these methods into the proposed framework remains a problem.

The last issue is the determination of the structure for causal inference. In the study by Davis and Swenson (2006), the structure can be pre-determined since their target of interest is rear-end accidents which mainly follow physical laws. Not all causal chains have such explicit sequential relationships. Although some data mining techniques, such as Bayesian networks or EM-algorithms, could help find possible network structure, the plausibility of derived structure requires professional judgments. Therefore, the determination of the causal inference structure is still a problem.

Exploring accident causality is much more difficult than apprehending the correlations among its factors. This study proposes a framework to understand accident causality from the chain perspective. The becoming comprehensive accident databases, powerful computational capabilities, and mature methodologies provide the opportunities to learn causal chains by doing classifications and applying causal inferences. Moreover, the study presents a conceptual chain framework of accidents which hopes to become the basis for future implementation of such idea.

The derived causal chains have much potential in practical applications. Since the derived chains contain detailed information about accident occurrences, with such detailed information on-hand, one can estimate the risks faced by drivers by matching their current driver characteristics, trip characteristics, vehicle characteristics, and road and environment characteristics. The individualized, instead of general, safety warning messages, for example, can then be delivered to a certain driver at the matched time and environment.

There are still some data and methodological issues required to be resolved. However, studying accident causality from chain perspective provides an approach to be closer to accident causality.

Chapter 4 EMPIRICAL STUDY

The objective of this chapter is to demonstrate the methodologies presented in Chapter 3. Prior to these demonstrations, the database adopted in these studies is introduced in Section 4.1. Subsequently, the empirical study of the approach for exploring accident characteristics, of the approach for analyzing heterogeneity of accident data, and of the approach for examining accident causality are shown in Section 4.2, 4.3, and 4.4, respectively.

4.1 Taiwan Traffic Crash Database

The Taiwan traffic crash database has been maintained by National Police Agency, Ministry of the Interior. The collected crash types are twofold: A1 (death involved) and A2 (injury only)[§]. When the involved persons in a crash die within 24 hours due to the crash, this crash is classified as an A1-type crash. On the other hand, when the involved persons in a crash get injured only or died after 24 hours of the crash, this crash is classified as an A2-type crash. By law, hospitals are obligated to report to police departments if a patient is died of or serious injured from a car accident.

The investigation format of a crash consists of two sheets: one for recording crash characteristics and the other for recording personal characteristics. In the crash sheet, the collected items include the number of death and injured persons, natural environmental factors (e.g. weather and illumination), and road environmental factors (e.g. road type, road shape, median type, signal type, and so on). As to the personal sheet, the collected items include all the involved persons' socio-demographic characteristics (e.g. age, occupation or gender), behavior (e.g. protection equipment use or drinking condition), vehicle information (e.g. vehicle type or plate number), and crash related characteristics (e.g. crash type, police-judged causes).

Although it was understood that more data are more welcome in safety research, the accident database collected by police department was considered as the only data source in this research. However, it should be noted that the framework and approach proposed in this study could be extended when more data were available.

[§] Although the A3 crash type (property damage only) is also collected by police departments, it is not provided by National Police Agency. Moreover, since A3 is a less serious crash type, problems such as under reporting could damage the analysis. Therefore, A3 crashes were excluded in the study.

4.2 Patterns of Taiwan Single Auto-Vehicle Accidents

4.2.1 Data

Taiwan 2003 single auto-vehicle (SAV) accident data is chosen to demonstrate the feasibility and usefulness of rough sets theory and the proposed framework in accident chain analyses. Single auto-vehicle accidents are those in which only one vehicle is involved. Since no other vehicles or pedestrians, are involved, the problem can be more accurately defined. Meanwhile, far more information is required to explore the accident patterns of multi-vehicle accidents. Consequently, studying SAV accidents is a good start for the study.

The total number of SAV accidents, excluding invalid cases, was 2,316. The number of invalid cases was 20, which accounted for 0.86% of the total cases. These cases were invalid mainly due to the unknown attribute values of the driver's characteristics. They were directly ignored in the study based on their relatively small size. The collected attributes and their corresponding categories are summarized in Table 4-1. Accident type is chosen as the decision attribute while the other attributes are considered as condition attributes. The categories of the accident types herein were slightly different from the original data provided by the National Police Agency. While rollover crashes, off-road crashes, crashes with architectures, crashes with work zone and other crashes were directly adopted from the original database, the crashes with road facilities include crashes with guardrails, traffic signals, toll collection booths, median islands, trees and utility polls; and the crashes with non-fixed objects include those bumping into animals as well as other non-fixed objects.

A popular rough sets software, ROSE2 (Rough sets Data Explorer), was used in this study where LEM2 (Grzymala-Busse, 1992; Grzymala-Busse and Werbrouck, 1998) is embedded to generate a minimum rule set covering all objects. The results of rough sets analysis consist of five parts: rule generation, quality of approximation, rule validation, rule description and significance of condition attributes.

TABLE 4-1 Attribute and Category

Dimension	Attribute	Category
Driver characteristics (Condition attribute)	Age	Under(<18), Young(18-35), Middle-aged(36-55), Old(>55)
	Gender	Male, Female
	License type	Regular, Occupational, Military, Other
	License condition	Valid, Invalid, Unknown
	Occupation	Student, Working people, No job, Other, Unknown
Trip characteristics (Condition attribute)	Trip purpose	Work, School, Social, Shop, Sightseeing, Business, Other, Unknown
	Trip time	Morning peak (07:00-09:00 h), Day offpeak (09:00-16:00 h), Afternoon peak (16:00-19:00 h), Night offpeak (19:00-23:00 h), Midnight to daybreak (23:00-07:00 h)
Behavior and environment factors (Condition attribute)	Protect equipment use	Use, No use, Unknown
	Cell phone use	Use, No use, Unknown
	Drink condition	Drink, No drink, Other
	Road type	Highway, Other
	Speed limit	50-, 51-79, 80+
	Road shape	Intersection, Segment, Ramp, Other
	Pavement material	Asphalt, Other, No pavement
	Surface deficiency	Normal, Other (e.g. holes, soft, and so on)
	Surface status	Dry, Wet, Other
	Obstruction	Yes, No (within 15 meters)
	Sight distance	Good, Bad (based on road design speed)
	Signal type	Regular, Flash, No signal
	Signal condition	Normal, Abnormal, No signal
	Direction divided facility	Island, Marking, None
	Roadside marking	Yes, No
	Climate	Sunny or cloudy, Rainy, Other
	Light condition	With light, No light
Accident (Decision attribute)	Accident type	Bump into bridge or architecture (198)[a]
		Bump into road facility (1 564)
		Bump into non-fixed object (17)
		Bump into work zone (21)
		Off-road (297)
		Rollover (93)
		Other (126)

[a] sample size of the accident type

4.2.2 Rule Generation

As shown in Table 4-2, the number of rules generated increases with the completeness of the specified condition attributes. Since all the condition attributes are categorical variables, the incorporation of any additional condition attribute with *n* categories would expand the possible classifications *n* times. However, while the quality of approximation is much enhanced, the number of rules does not increase proportionally but only with limited growth. This implies that the condition attributes included are valid enough to classify the accident types and that some patterns do exist for the SAV accidents in Taiwan rather than all SAV accidents being regarded as unique.

TABLE 4-2 Rough Sets Results

Approach		Accident type	Generated rules	Accuracy	Quality of classification	Hit rate	Overall hit rate
		bridge		0.19%		4.55%	
		facility		2.26%		5.05%	
	D[a]	non-fixed obj.		0.00%		11.76%	
1	↓	work	104	0.00%	3.02%	23.81%	6.30%
	A	off-road		0.51%		4.04%	
		rollover		0.11%		12.90%	
		other		0.50%		21.43%	
		bridge		0.00%		0.00%	
		facility		0.26%		1.73%	
	T	non-fixed obj.		0.00%		23.53%	
2	↓	work	38	0.00%	0.26%	23.81%	4.62%
	A	off-road		0.00%		11.11%	
		rollover		0.00%		27.96%	
		other		0.00%		9.52%	
		bridge		7.52%		21.21%	
		facility		31.59%		27.88%	
	B	non-fixed obj.		1.59%		29.41%	
3	↓	work	508	9.66%	38.69%	23.81%	25.60%
	A	off-road		7.88%		21.21%	
		rollover		2.34%		24.73%	
		other		4.96%		15.08%	
		bridge		1.68%		20.20%	
	D	facility		16.47%		21.93%	
	↓	non-fixed obj.		0.97%		0.00%	
4	T	work	474	0.76%	20.16%	23.81%	20.60%
	↓	off-road		3.38%		19.53%	
	A	rollover		1.77%		18.28%	
		other		1.31%		11.11%	
		bridge		39.89%		16.67%	
	D	facility		67.78%		53.45%	
	↓	non-fixed obj.		19.64%		0.00%	
5	B	work	766	79.17%	74.65%	9.52%	42.01%
	↓	off-road		41.39%		22.90%	
	A	rollover		17.39%		21.51%	
		other		30.22%		11.11%	
		bridge		31.29%		19.19%	
	T	facility		64.05%		49.36%	
	↓	non-fixed obj.		8.43%		0.00%	
6	B	work	787	45.95%	70.68%	14.29%	39.21%
	↓	off-road		33.96%		21.55%	
	A	rollover		18.38%		18.28%	
		other		21.11%		11.11%	
	D	bridge		74.68%		12.63%	
	↓	facility		90.57%		69.69%	
	T	non-fixed obj.		41.94%		5.88%	
7	↓	work	808	100.00%	92.88%	23.81%	51.38%
	B	off-road		80.65%		17.51%	
	↓	rollover		66.39%		9.68%	
	A	other		69.81%		6.35%	

[a] *D*: Driver characteristics; *T*: Trip characteristics; *B*: Behavior and environment factors; *A*: Accidents

4.2.3 Quality of Approximation

The accuracy of approximation for rollover and bump-into-non-fixed object accidents is extremely low, except when all condition attributes are included. However, the accuracy of approximation for the bump-into-bridge accidents, off-road accidents, and other accident types can be increased to 30%~40% if B&E factors are combined with either driver characteristics or trip characteristics. This can be raised to 70% or even 80% if all condition attributes are included. Roughly speaking, bump-into-facility and work zone are the most definable accident types, while bump-into-bridge, off-road, and other accident types are moderately definable accident types, and rollover and bump-into-non-fixed object are the least definable accident types.

The quality of classification is proportional to the completeness of selected attributes. Approach 7 shows the highest quality, while Approach 2 shows the lowest. B&E factors show the most important attributes for the quality of classification partly due to their wide coverage of affecting factors, which are also proximal factors. Each dimension alone (Approaches 1, 2, 3) does not yield a good quality of classification. If at least two dimensions are combined, the quality of classification is much enhanced. For example, the quality of classification for B&E alone is 38.69%. However, it is raised to 70.68% by merely combining it with trip characteristics in which only two more attributes are included.

These results suggest that accidents should not be resolved by single factor, but by a chain of factors. Previous countermeasures focused mostly on B&E proximal factors. It is effective; however, to further improve road safety, all factors associated in the factor chain may need to be taken into serious consideration. Furthermore, neglecting factors in a chain may result in rather different stories and blur the interactions among accident features.

4.2.4 Rule Validation

The *10*-fold cross-validation technique is used to conduct validation test of classification results. The hit rate, i.e. the percentage of correct prediction, for the bump-into-facility accidents can be improved by up to 70 percent when all condition attributes are considered. On the other hand, the hit rates for the remaining accident types all range from 0 to 20 or 30 percent. This suggests that the occurrence of a bump-into-facility accident may follow similar paths and is more predictable. But for other accident types, the rules generated from their training cases may not be representative since their occurrences are mostly random.

The higher the quality of approximation, the higher the overall hit rate and the hit rate for the bump-into-facility accidents. Yet, the bump-into-bridge and bump-into-non-fixed

object accidents show the highest hit rate in Approach 3, which consists of B&E proximal factors only and reveals the unexpected and random characteristics of these kinds of accidents. Its hit rate becomes lower if other condition attributes are included. These results suggest that except for the bump-into-facility accidents where more information is useful, different accident types have their corresponding useful condition attributes. For example, the condition attributes of driver characteristics are useful for the bump-into-work zone and the other accident types, and those of trip characteristics are useful for rollover accidents. All these results are helpful for devising adequate countermeasures.

The classification results show that most of the bump-into-bridge, bump-into-facility, off-road and rollover accidents are assigned to the bump-into-facility accident type and least into the bump-into-non-fixed and bump-into-work zone accident types. This suggests that, while most accidents are associated with some critical condition attributes which lead to the similar classification pattern, bump-into-non-fixed and bump-into-work zone accidents are related to very distinctive characteristics. This also implies that some similarities may exist in the occurrence of the bump-into-bridge, bump-into-facility, off-road and rollover types since they are all related to road geometry and driving environments. These similarities are the reasons for the low hit rates for the bump-into-bridge and off-road accident types, since they can be easily assigned to the bump-into-facility accidents due to the fact that the sample size for the bump-into-facility accident type outweighs theirs. As a consequence, more rules associated with the occurrence of the bump-into-facility accident type are generated and dominate the classification pattern. On the other hand, the remaining accident types, such as the bump-into-non-fixed object, are more closely related to driver characteristics and are relatively unique.

4.2.5 Description of Significant Rules

Rules are generated from the accident database by rough sets theory, and the significant rules for each accident type are shown in Table 4-3. The rule strength – the number of accident cases matching the rule – for most accident types is small except for the bump-into-facility type. The highest strength for most types is about 3 or 4. This shows the uniqueness of those accident types, especially, the infrequent and stochastic occurrences of the bump-into-non-fixed objects. Interestingly, the derived factor chain shows that a drinking driver without regular license exhibits a relatively high possibility of being involved in bump-into-non-fixed object accidents on a secondary road without roadside marking and light.

TABLE 4-3 Description of Significant Rules

Accident type	Rule description[a]
Bump into facility (35)[b]	● Driver: Working people; ● Behavior: Not drinking; ● Environment: Road segment ; Median island ; Wet surface ; No obstruction within 15 meters;
Off-road (7)	● Driver: Regular license; Student; ● Environment: Speed limit 50-79 ; Median marking ; With roadside marking ; With light;
Bump into bridge or architecture (4)	● Driver: Middle-aged; Working people; ● Behavior: Drinking; ● Environment: Speed limit less than 50; Collision position rather than intersection, segment and ramp; With roadside marking;
	● Behavior: Drinking; Cell phone use unknown; ● Environment: Flash signal ; No roadside marking ; Dry surface;
	● Driver: Young; Working people; ● Trip: Other trip purpose; Between midnight and daybreak; ● Behavior: Not drinking; ● Environment: No signal ; Median marking ; With roadside marking ; With light ; Poor sight distance;
	● Behavior: Not drinking; ● Environment: Collision position rather than intersection, segment and ramp ; Pavement rather than asphalt ; No directional-divided facility ; No roadside marking ; No obstruction within 15 meters;
Bump into work zone (4)	● Driver: Male; Regular license type; Unknown occupation; ● Trip: During midnight to daybreak; ● Environment: Speed limit less than 50; Asphalt pavement; No signal ; Obstruction within 15 meters;
Rollover (3)	● Driver: Young; Working people; ● Trip: Social trip; Night offpeak; ● Behavior: Not drinking; ● Environment: Median marking
	● Driver: Young; Male; Regular license type; Working people; ● Trip: Day offpeak; ● Environment: Speed limit less than 50 ; Regular signal;
Bump into non-fixed object (2)	● Driver: Other license type; ● Behavior: Drinking; Cell phone use unknown; ● Environment: Speed limit less than 50 ; No roadside marking ; No light;

[a] please refer to Table 4-1 for the details of condition attributes
[b] the value represents the rule strength

The most significant rule for the bump-into-work zone suggests that there is a relatively high risk when a driver approaches work zone on a road with speed limit less than 50 (kph) around midnight. This information suggests that more effective and sufficient work zone traffic controls should be installed, particularly in the dark work zone on those secondary roads. The rule reflects the fact that, to save cost, it is often the case that safety measures are not properly implemented, especially on rural secondary roads.

For rollover accidents, two significant rules describe young working people who are driving during off-peak period as being more likely involved in the rollover accidents, probably due to the low traffic and high speed.

Four significant rules for the bump-into-bridge accidents describe two conditions: drinking driving under normal road environment and sober drivers under abnormal road

environment. Specific deficiencies exist on both conditions for this accident type. This shows the necessity for the government to prevent this type of accident by improving the road environment or raising the penalties for drinking driving.

The derived factor chain for off-road accidents shows that student drivers who are young and less experienced exhibit a relatively high possibility of being involved in off-road accidents. This result echoes the graduated licensing scheme currently existing in many countries (Simpson, 2003). Moreover, the factor chain shows that the corresponding driving environment is normal, i.e. no particularly unfavorable factors such as drinking driving or poor sight distance appear on the chain. Since other driving groups such as working people do not show similar accident patterns as off-road accident type, the government should seriously consider educating student drivers to enhance their situational awareness of driving environment and reduce their risk-driving behavior on roads.

The rule with the highest strength goes to bump-into-facility accidents. It describes 35 employed sober drivers rather than students driving on an island-divided road segment where the surface was wet and there were no obstructions within 15 meters. The wet surface denotes lower friction on road surfaces that increase the difficulty of handling vehicles. Meanwhile, drivers generally might slow down their driving speed to maintain vehicles at an "acceptable" speed. Therefore, the extremely high supporting evidence may imply that those drivers overestimated their driving skills and underestimated the risk of the decrease in surface friction.

4.2.6 Significance of Condition Attributes

The significance of condition attributes is measured by their presence on the derived rules. When a condition attribute shows up more frequently in the rules, it is more likely being used to describe the occurrence of accidents and hence is more significant in distinguishing accident types. The presence of a condition attribute is represented with presence percentage which is calculated by summing up its presence in each rule weighted with cases of the associated rule divided by total cases. Here, only the rules derived from Approach 7 are adopted in the calculation since Approach 7 shows the most satisfactory performance. Moreover, since condition attributes with more categories tend to distinguish accident types more effectively, comparisons are made on those with same number of categories. As shown in Figure 4-1, gender, roadside marking and light condition; speed limit, road shape and directional divided facility; age, occupation, trip time and drinking condition are those attributes with a relatively higher presence percentage among all condition attributes with two, three and four or more categories, respectively.

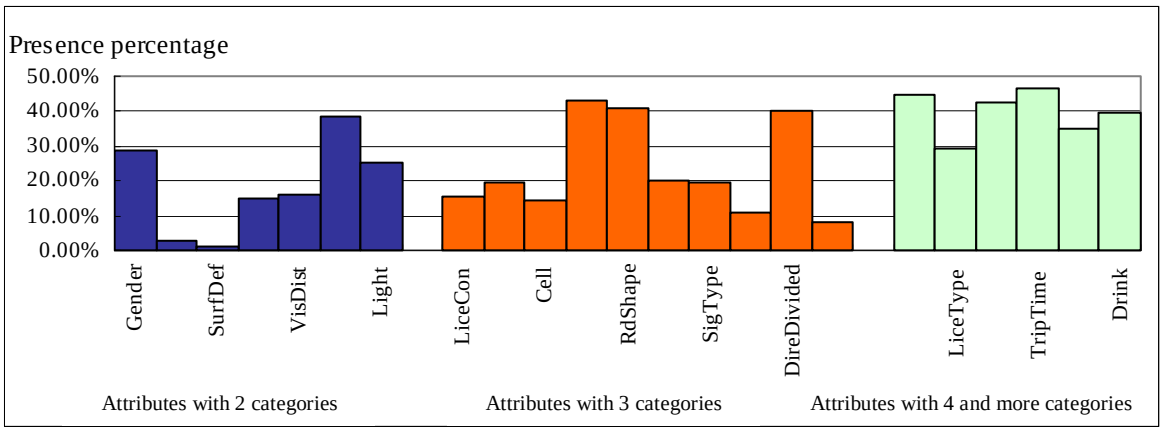


FIGURE 4-1 Presence percentage of condition attributes.

4.3 Heterogeneity of Taiwan Single Auto-Vehicle Accidents

The data and software used in the previous section were adopted to demonstrate the feasibility and usefulness of the proposed framework on analyzing heterogeneity of accident data.

4.3.1 Strength of Accident Pattern

With 23 condition attributes (*pavement material* is redundant and excluded), 808 rules were generated as the minimum requirement to cover 2,316 accident cases; i.e., one rule stood on average for three accident cases. As shown in Table 4-4, the frequencies of some rules were high while some were low. The maximum strength was 35 for one rule while the minimum strength was 1 for 285 rules. More than half of the rules were of strength equivalent to 1 or 2. This demonstrates the uniqueness of most accident patterns for Taiwan's SAV accidents in 2003; that is, most accidents occurred with different driver characteristics, different trip characteristics and/or different behavior and environmental factors. Nevertheless, for those rules with high strength, they represent a large portion of accidents occurring repeatedly with identical patterns.

TABLE 4-4 Strength and the Corresponding Number of Rules

Strength	1	2	3	4	5	6	7	8	9	10
No. of rules	285	167	76	47	28	23	31	24	20	13
Rule percentage (%)	35.27	20.67	9.41	5.82	3.47	2.85	3.84	2.97	2.48	1.61
Strength	11	12	13	14	15	16	17	18	19	20
No. of rules	11	19	7	10	4	8	6	2	1	7
Rule percentage (%)	1.36	2.35	0.87	1.24	0.50	0.99	0.74	0.25	0.12	0.87
Strength	21	22	23	25	26	27	29	35		
No. of rules	5	2	3	4	1	1	2	1		
Rule percentage (%)	0.62	0.25	0.37	0.50	0.12	0.12	0.25	0.12		

The differences of accident characteristics between rules with high frequencies and those of low frequencies are the primary concerns in this research. This study adopted 23 condition attributes to describe the occurrence of accidents, which made the analysis at a very microscopic level. As a consequence, each accident may follow its exclusive pattern rather than identical patterns. Nevertheless, in addition to the rules with low frequencies, the rules with high frequencies were also derived. This shows that stereotype accidents do exist.

4.3.2 Accident Grouping

For the convenience of interpretations, two to six groups were preferred, in which the more significantly different condition attributes existed among groups, the more desired they were. In this research, a bottom-up procedure was implemented to determine the boundaries of accident groups. Statistical tests were employed to determine the appropriateness of cluster boundaries. The χ^2 test was adopted for large sample sizes while the Fisher's exact test for small sample sizes. The significance level was set at 0.10, and three clusters were then determined after thorough analysis. The corresponding rule strength intervals for the clustered groups were 1-2, 3-23 and 25-35 with the number of accidents being 619, 1451 and 246, respectively. Seen in Table 4-5, the *license type* and *roadside marking* attributes were the only two non-significant condition attributes among clusters. All other condition attributes were significantly different among groups.

TABLE 4-5 Test Results of Condition Attributes for the Final Partition

Driver characteristics		Trip characteristics		Behavior and environmental factors	
Condition attribute	*P*-value	Condition attribute	*P*-value	Condition attribute	*P*-value
Ages	0.0047**	Trip purposes	0.0000**	Protect equipment use	0.0044**
Genders	0.0001**	Trip time	0.0000**	Cell phone use	0.0074**
License types	0.6558			Drinking condition	0.0000**
License conditions	0.0009**			Road type	0.0073**
Occupations	0.0000**			Speed limit	0.0000**
				Road shape	0.0000**
				Pavement material	0.0118**
				Surface deficiency	0.0022**
				Surface status	0.0034**
				Obstruction	0.0307**
				Sight distance	0.0000**
				Signal type	0.0000**
				Signal condition	0.0000**
				Median	0.0000**
				Roadside marking	0.2621
				Weather	0.0704*
				Light condition	0.1000*

*: 0.10 significance level; **: 0.05 significance level

The characteristics of the accident groups as well as the whole data are shown in Table 4-6. This shows that the accident characteristics of the whole dataset were relatively close to

those clustered in the weak and medium rule strength. However, the accident characteristics of the high rule strength group appeared substantially different from the others and showed relatively high percentages of the following attributes: drivers were *male* and *young*; drivers' licenses were *invalid*; trips occurred between *midnight* and *dawn*; trip purposes were *not* specified; speed limit was *medium* (51-79 KPH); median was an *island*; crash positions were at *intersections*, signals were under *flash* operation*;* road surfaces were *wet*; roads had *no obstructions*; sight distances were *good*; and drivers were under the conditions of *wearing seatbelts, not talking on their cell phones* and *not drinking*.

These results may suggest that the accidents with strong patterns, i.e. high rule strength, are most likely related to high-risk drivers. Young and male drivers, compared with elderly and female drivers, respectively, have been identified as high-risk drivers in many studies (Massie et al., 1995; Massie et al., 1997; Murray, 1997; Kim et al., 1998; Laapotti and Keskinen, 1998; Shinar and Compton, 2004). Drivers on road without a valid driver license have explicitly exhibited risky behavior. The road environment between midnight and dawn has been associated with a more risky driving environment compared with driving during daytime (Lin and Fearn, 2003; Keall et al., 2005). Although drivers who drive between midnight and dawn can not be automatically considered as high-risk drivers, there is a high possibility that more high-risk drivers are among them since a relatively high percentage of these drivers are driving for no specific purpose. In other words, they are probably enjoying the night lifestyle and not driving for school, business or other necessary purposes.

In addition, accidents associated with strong patterns occur under conditions that may not appear for average or conservative drivers. No obvious causes from the road or natural environment were found in these patterns – neither obstructions on the road nor poor sight distance. Interestingly, these drivers were not using cell phones, had not drunk alcohol and were wearing seatbelts. This shows that they were rational drivers and were following the law. In particular, it might reflect the culture differences in drinking – drinking is probably not as common for the young males in Taiwan as those in Western countries. As to the accident location, the findings met our expectations: single vehicle accidents occur more likely on road segments than at intersections. This may result from the fact that traffic flows at intersections are more complicated and subject to more conflicts; consequently, multi-vehicle accidents are more likely to happen at intersections. However, since most SAV accidents with strong patterns at intersections turned out to be collisions with road facilities, this implies that facilities near intersections may be the critical contributing factor for high-risk drivers, especially during the night when traffic is low, which encourages fast driving for some. Moreover, a wet road surface increases the difficulty of maneuvering a vehicle. The relatively high percentage of wet surfaces as a factor in the occurrence of accidents with strong accident patterns may imply that the drivers have immature skills or

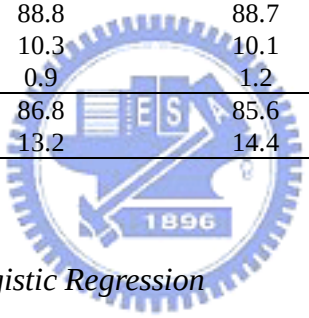
that they are overconfident.

TABLE 4-6 Accident Characteristics for Whole and Partitioned Accident Groups

Condition attribute	Category	Whole dataset (%)	Weak pattern (Strength = 1-2) (%)	Medium pattern (Strength = 3-23) (%)	Strong pattern (Strength = 25-35) (%)
Age	Under	0.3	0.5	0.2	0.0
	Young	60.3	59.9	60.9	67.4
	Middle-aged	32.2	29.7	33.1	26.5
	Elderly	6.5	8.8	5.4	5.7
	Other	0.7	1.1	0.4	0.4
Gender	Male	86.0	84.3	86.3	95.1
	Female	14.0	15.7	13.7	4.9
License type	Regular	81.6	80.8	80.6	80.8
	Occupational	6.9	7.2	7.4	4.9
	Military	0.4	0.7	0.3	0.4
	Other	11.1	11.3	11.7	13.9
License condition	Valid	86.7	87.3	85.4	84.5
	Invalid	8.0	8.0	8.7	14.3
	Unknown	5.3	4.7	5.9	1.2
Occupation	Student	4.0	6.5	2.8	3.3
	Working people	69.1	51.1	55.7	67.3
	No job	8.1	8.2	7.9	5.7
	Unknown	18.8	34.2	33.6	23.7
Trip purpose	Work	7.3	6.0	8.7	6.1
	School	0.4	1.3	0.0	0.0
	Social	9.0	9.1	8.9	8.2
	Shop	1.9	2.8	1.5	1.6
	Sightseeing	4.8	4.7	4.8	2.4
	Business	2.1	2.1	2.0	2.4
	Other	52.5	50.8	51.6	67.9
	Unknown	22.0	23.2	22.5	11.4
Trip Time	Morning peak	5.8	5.2	6.8	1.6
	Day offpeak	21.5	22.3	19.0	18.0
	Afternoon peak	10.7	13.7	9.0	13.1
	Night offpeak	15.8	15.9	16.0	12.2
	Midnight to daybreak	46.2	42.9	49.2	55.1
Protect equipment use	Use	83.8	85.8	82.0	90.2
	No use	3.8	4.1	4.3	2.9
	Unknown	12.4	10.1	13.7	6.9
Cell phone use	Use	0.9	0.5	1.1	0.8
	No use	87.1	88.7	86.1	93.5
	Unknown	12.0	10.8	12.8	5.7
Drinking condition	Drinking	28.2	26.8	27.3	26.1
	Not drinking	61.5	62.6	60.5	72.2
	Unknown	10.3	10.6	12.2	1.7
Road type	Highway	7.7	5.5	9.5	9.0
	Other	92.3	94.5	90.5	91.0
Speed limit	50-	55.4	59.9	55.2	29.8
	51-79	37.0	34.1	35.8	60.4
	80+	7.6	6.0	9.0	9.8
Road shape	Intersection	20.6	19.6	22.4	31.0
	Segment	79.0	79.4	77.5	69.0
	Ramp or other	0.4	1.0	0.1	0.0
Surface deficiency	Normal	98.7	97.4	99.1	99.2
	Other	1.3	2.6	0.9	0.8

Table 4-6 Accident Characteristics for Whole and Partitioned Accident Groups (Contd.)

Condition attribute	Category	Whole dataset (%)	Weak pattern (Strength = 1-2) (%)	Medium pattern (Strength = 3-23) (%)	Strong pattern (Strength = 25-35) (%)
Surface status	Dry	86.6	87.3	84.3	77.6
	Wet	13.0	12.2	15.5	22.4
	Other	0.4	0.5	0.2	0.0
Obstruction	Yes	94.6	93.0	94.8	97.6
	No	5.4	7.0	5.2	2.4
Sight distance	Good	89.6	87.9	89.7	93.9
	Bad	8.3	7.7	9.1	4.1
	Unknown	2.1	4.4	1.2	2.0
Signal type	Regular	9.8	6.9	12.0	14.3
	Flash	7.0	6.9	7.5	25.3
	No signal	83.2	86.2	80.5	60.4
Signal condition	Normal	15.9	12.1	18.6	39.6
	Abnormal	0.2	0.3	0.2	0.0
	No signal	83.9	87.6	81.2	60.4
Median	Island	34.0	27.2	38.0	59.2
	Marking	45.9	49.9	42.2	15.1
	None	20.1	22.9	19.8	25.7
Roadside marking	Yes	57.3	54.6	58.1	56.3
	No	42.7	45.4	41.9	43.7
Weather	Sunny or cloudy	88.8	88.7	86.9	84.1
	Rainy	10.3	10.1	11.9	15.9
	Other	0.9	1.2	1.2	0.0
Light condition	With light	86.8	85.6	86.5	81.6
	No light	13.2	14.4	13.5	18.4



4.3.3 Results of Multinomial Logistic Regression

To further explore the characteristics for each sub-dataset, multinomial logistic regressions are conducted for a variety of clustered accidents. Five models were devised and tested, including base model (whole dataset, 2316 cases), weak strength model (619 cases), medium strength model (1451 cases), weak plus medium model (2070 cases) and medium plus strong model (1697 cases). For fair comparisons, all models were estimated with an identical specification which was developed based on the whole dataset. Based on concerns about sample size and the limitation of logistic regression, only those attributes showing up in over 35% of the rules were considered, which included age, trip time, drinking condition, speed limit, road shape, median and roadside marking. Moreover, to avoid empty cells, some small categories which represented unclear conditions, such as unknown or other, were excluded (413 cases were excluded). The likelihood ratio test at the significance level of 0.10 was adopted to select the variables. This resulted in five variables being included in the final specification. They were, age (young, middle-aged, elderly), trip time (peak, off-peak, midnight), drinking (not drinking, drinking), road shape (intersection, segment) and median (island, marking, none). The estimation results for the proposed models are shown in Table 4-7, where the reference accident type was set to the *collision with road*

facility. All models were shown to be well fitted based on the χ^2 goodness of fit tests at the significance level of 0.10. Overall, some significant differences were observed among the models.

From the results of the base model in Table 4-7, several factors contributing significantly to a variety of accident types could be clearly identified. They were interpreted, based on the comparison to collisions with road facilities, in detail as follows:

1. Young drivers, compared to collisions with road facilities, were more likely to be involved in rollover accidents. The odds of a middle-aged driver involved in rollover crashes was 0.547 times that of a young driver. This is consistent with past studies that young drivers exhibit higher percentages of rollover accidents (Farmer and Lund, 2002) than other age groups.

2. The odds ratios show that midnight accidents were more likely to be related to collisions with structures, and daytime accidents were more likely to be off-road and rollover accidents. These findings can be related to visibility of structures which are not as easily identified during the night time compared to regular road safety facilities. On the other hand, since fixed facilities can be better spotted and avoided during daytime, both off-road and rollover accidents are more likely to occur than collisions with road facilities. This may suggest that during daytime, drivers themselves, not road facilities, play a key role in the occurrence of single auto-vehicle accidents.

3. Drunk drivers tend to lose situational awareness and are much likely to lose control of their vehicles and hit structures or generate off-road accidents compared with crashing into road facilities. The odds of a drinking driver involved in collisions with structures and in off-road crashes compared to collisions with road facilities were 1.785 and 1.395 times respectively the odds a not-drinking driver would.

4. Intersections, where more road facilities (such as traffic lights) are expected and where vehicles tend to slow down, are more likely to have collisions with road facilities. On the other hand, off-road and rollover accidents are more likely to occur on road segments. These results were clearly shown in odds ratio values.

5. Referring to collisions with road facilities, the low odds ratios (0.295, 0.177 and 0.259) clearly suggest that roads with median islands could significantly reduce collisions with structures, work zones and off-road accidents. This result reflects the fact that higher road standards with better safety facilities help reduce some accidents, but will also create pitfalls if the safety facilities are not properly provided.

TABLE 4-7 Estimating Results of Multinomial Logistic Regression Models

Accident type		Whole (Base)		Weak strength		Medium strength		W+M		M+S	
		Coeff.	Odds ratio	Coeff.	Odds ratio	Coeff.	Odds ratio	Coeff.	Odds ratio	Coeff.	Odds ratio
Structure[1]	Intercept	-2.456**[2]		0.952*		-3.433**		-2.304**		-3.574**	
Age	Middle-aged	0.060	1.061	0.292	1.339	0.376	1.457	0.021	1.021	0.420	1.521
	Elderly	-0.236	0.790	-0.840	0.432	-0.159	0.853	-0.263	0.769	-0.140	0.869
Trip time	Peak period	-0.430*	0.651	-0.429	0.651	-1.449**	0.235	-0.454*	0.635	-1.443**	0.236
	Off-peak period	-0.063	0.939	-0.031	0.969	-0.431	0.650	-0.096	0.909	-0.408	0.665
Drinking	Drinking	0.579**	1.785	0.292	1.340	0.766**	2.151	0.549**	1.732	0.801**	2.228
Road shape	Intersection	-0.204	0.815	0.889**	2.433	-0.758*	0.469	-0.151	0.860	-0.825*	0.438
Median	Island	-1.222**	0.295	-1.120**	0.326	-2.005**	0.135	-1.194**	0.303	-2.061**	0.127
	Marking	0.316	1.372	-0.769*	0.463	0.559	1.749	0.238	1.269	0.630*	1.877
Non-fixed object	Intercept	-4.724**		-1.654*		-22.207**		-4.614**		-22.256**	
Age	Middle-aged	0.168	1.182	0.935	2.548	--[3]	--	0.151	1.163	--	--
	Elderly	0.925	2.523	0.770	2.160	--	--	0.949	2.583	--	--
Trip time	Peak period	-0.951	0.387	-0.802	0.448	--	--	-0.974	0.378	--	--
	Off-peak period	0.391	1.479	0.545	1.724	-0.411	0.663	0.374	1.454	-0.408	0.665
Drinking	Drinking	-0.452	0.636	-0.254	0.776	--	--	-0.477	0.621	--	--
Road shape	Intersection	-0.858	0.424	0.231	1.260	--	--	-0.827	0.437	--	--
Median	Island	-0.804	0.448	-0.848	0.428	16.728**	1.8E+07	-0.726	0.484	16.650**	1.7E+07
	Marking	0.031	1.032	-0.922	0.398	17.507**	4.0E+07	-0.021	0.979	17.490	3.9E+07
Work zone	Intercept	-4.091**		-0.913		-4.984**		-3.941**		-5.156**	
Age	Middle-aged	0.784	2.191	1.139*	3.123	0.892	2.441	0.723	2.061	0.977	2.657
	Elderly	0.973	2.646	0.761	2.140	--	--	0.927	2.526	--	--
Trip time	Peak period	-0.327	0.721	-0.376	0.686	-0.802	0.449	-0.379	0.685	-0.771	0.463
	Off-peak period	-0.209	0.811	-0.036	0.964	-1.382	0.251	-0.250	0.779	-1.358	0.257
Drinking	Drinking	-0.889	0.411	-0.427	0.653	--	--	-0.913	0.401	--	--
Road shape	Intersection	-0.174	0.840	0.713	2.041	0.225	1.253	-0.125	0.882	0.157	1.170
Median	Island	-1.729**	0.177	-2.537**	0.079	-0.485	0.616	-1.670**	0.188	-0.482	0.617
	Marking	-0.619	0.538	-1.550**	0.212	-0.416	0.660	-0.673	0.510	-0.340	0.712
Off-road	Intercept	-1.770**		1.468**		-2.596**		-1.654**		-2.703**	
Age	Middle-aged	-0.121	0.886	0.144	1.155	0.071	1.074	-0.147	0.863	0.095	1.099
	Elderly	0.214	1.239	-0.336	0.714	0.233	1.263	0.203	1.225	0.237	1.267
Trip time	Peak period	0.391**	1.479	0.005	1.005	0.360	1.433	0.377**	1.458	0.363	1.438
	Off-peak period	0.464**	1.590	0.105	1.110	0.587**	1.799	0.443**	1.558	0.603**	1.827
Drinking	Drinking	0.333**	1.395	-0.293	0.746	0.655**	1.926	0.303**	1.354	0.687**	1.988
Road shape	Intersection	-1.124**	0.325	0.166	1.180	-1.587**	0.204	-1.075**	0.341	-1.634**	0.195
Median	Island	-1.350**	0.259	-1.148**	0.317	-1.471**	0.230	-1.284**	0.277	-1.561**	0.210
	Marking	-0.167	0.847	-1.125**	0.325	0.011	1.011	-0.216	0.806	0.054	1.056

[1] The reference category for accident type is collision with road facility, for age is young, for trip time is midnight, for drinking is not drinking, for road shape is segment, and for median is no median.

[2] * significance level for Wald χ^2 statistic at 0.10; ** significance level for Wald χ^2 statistic at 0.05

[3] -- zero accident count for that accident type and condition attribute category

Table 4-7 Estimating Results of Multinomial Logistic Regression Models (Contd.)

Accident type		Whole (Base)		Weak strength		Medium strength		W+M		M+S	
		Coeff.	Odds ratio	Coeff.	Odds ratio	Coeff.	Odds ratio	Coeff.	Odds ratio	Coeff.	Odds ratio
Rollover	Intercept	-3.198**		0.337		-6.386**		-3.088**		-6.529**	
Age	Middle-aged	-0.604**	0.547	-0.077	0.926	-0.751	0.472	-0.615**	0.541	-0.719	0.487
	Elderly	-0.235	0.791	-0.758	0.469	-0.400	0.670	-0.198	0.821	-0.468	0.626
Trip time	Peak period	0.352	1.422	-0.038	0.963	0.872	2.393	0.355	1.427	0.851	2.343
	Off-peak period	1.078**	2.937	0.874**	2.396	1.853**	6.378	1.067**	2.908	1.870**	6.490
Drinking	Drinking	-0.574*	0.563	-0.508	0.602	-19.701	0.000	-0.606**	0.546	-19.670	0.000
Road shape	Intersection	-0.889**	0.411	-0.131	0.878	-0.213	0.808	-0.875**	0.417	-0.155	0.856
Median	Island	-0.289	0.749	-0.293	0.746	0.832	2.298	-0.228	0.796	0.779	2.180
	Marking	-0.121	0.886	-0.948*	0.387	1.158	3.185	-0.177	0.837	1.237	3.447
Other	Intercept	-3.373**		0.253		-5.367**		-3.218**		-5.523**	
Age	Middle-aged	0.237	1.267	0.602*	1.825	0.941*	2.563	0.205	1.227	0.991*	2.693
	Elderly	0.924**	2.520	0.562	1.754	--	--	0.902**	2.464	--	--
Trip time	Peak period	0.272	1.313	0.036	1.036	-0.508	0.602	0.236	1.266	-0.484	0.617
	Off-peak period	-0.122	0.886	-0.285	0.752	-0.105	0.900	-0.147	0.863	-0.077	0.926
Drinking	Drinking	0.162	1.176	-0.037	0.963	-0.066	0.936	0.134	1.143	-0.043	0.958
Road shape	Intersection	0.170	1.185	1.104**	3.016	0.170	1.186	0.188	1.207	0.144	1.155
Median	Island	-0.233	0.792	-0.216	0.806	-0.920	0.398	-0.202	0.817	-0.968	0.380
	Marking	0.363	1.437	-0.655	0.519	0.870	2.386	0.286	1.331	0.945	2.574

Additionally, results from models with different rule strengths show some very interesting characteristics of accidents and were also observed and are worth noting.

1. The results from the weak strength model showed many differences. This may imply that the characteristics of accidents occurring uniquely are highly different from accidents with medium or strong rule strength. The age, trip time and drinking attributes played insignificant roles in differentiating the accident types, except work zone accidents, under the weak strength model. On the other hand, road-facility-related attributes (including road shape, median island and median marking) contributed significantly in differentiating the accident types under weak strength accidents. This is consistent with the fact that the occurrence of weak rule strength accidents is rather stochastic on poorly constructed roads.

2. In comparing the medium plus strong model with the medium strength one, the differences were slight. It may be because of the fact that the sample size of accidents with strong strength was relatively small (7.86% of the total accidents). The only difference was the occurrence of collisions with structures on the roads with median marking. The significantly high possibility of drivers associated with the strong rule strength being involved in collisions with structures suggests that there is a small portion of high-risk drivers who may easily ignore the unfavorable road attributes.

3. The median island attribute showed very consistent estimation results among all models. Almost all coefficients under this category were negative and significant. This may suggest that the relatively higher safety standards of roads with median islands reduce the occurrence of facility-irrelevant accidents.

4. Except for the weak strength model, the intersection area which is equipped with more facilities than road segments is consistently prone to the occurrence of facility-related accidents.

5. Except for the weak strength model, the drinking attribute showed positive signs towards the structure and off-road types under all models. This may result from the fact that drunk drivers usually drive faster, have lower capability of handling their vehicles and are in lower awareness of traffic and road conditions.

6. As for the trip time attributes, the coefficients of off-road and rollover types were consistently and positively significant among most models during off-peak periods. This may suggest that drivers themselves, rather than the road environment (structure, work zone, facility, etc.), play the key role in the occurrence of single auto-vehicle accidents.

In summary, the findings from multinomial logistic regression analyses indicate that drivers involved in accidents with strong rule strength are at somewhat high-risk, although the sample size compared to general drivers is limited and only part of their associated attributes can be specifically identified. Therefore, corresponding countermeasures may be focused on enhancing drivers' awareness of potential threats on roads and on their dangerous driving behaviors. On the other hand, it was found that rather than the driver and trip characteristics, road facilities – such as median and roadside marking – play the key role in accidents associated with weak rules. Thus, improvement in the quality of road maintenance may prevent such accidents. It is clear that countermeasures designed to target accidents with strong and with weak rules should focus on different preventive aspects.

4.4 Causality of Taiwan Single Auto-Vehicle Accidents

4.4.1 Data

The 2005 Taiwan single auto-vehicle (SAV) accident data was adopted to demonstrate the feasibility of the proposed approach for accident causality analysis. In particular, accident severity was considered as the target variable for this study. The primary reason of replacing the dataset used in the previous two sections with another dataset is that the rule

support is extremely low except the bump-into-facility crash type. It demonstrates the uniqueness of those accident types and might result in the void of rules with relationships.

The 2005 Taiwan single auto-vehicle (SAV) accident data was also collected by police departments including all the death involved and injury only accidents. The total number of SAV accidents, excluding invalid cases, was 3,138. The number of invalid cases was 27, which accounted for 0.86% of the total cases. These cases were invalid mainly due to the unknown attribute values of the driver's characteristics. They were directly ignored in the study based on their relatively small size. The collected attributes and their corresponding categories are summarized in Table 4-8.

TABLE 4-8 Attribute and Category

Attribute	Category
Age	Under (<18), Young (18-35), Middle-aged (36-55), Elderly (>55)
Gender	Male, Female
License type	Regular, Occupational, Other
License condition	Valid, Invalid, Unknown
Occupations	Student, Working people, No job, Unknown
Trip purpose	Necessary (Working, school, business), Other
Trip time	MP (07-09), DOP (09-16), AP (16-19), NOP (19-23), Midnight (23-07)
Seat belt use	Fastening, Not fastening, Unknown
Cell phone use	Using, Not using, Unknown
Drinking condition	Drinking, Not drinking, Other
Road type	Highway, Urban, Rural
Speed limit	50-, 51-79, 80+
Road shape	Intersection, Segment, Ramp or other
Pavement material	Asphalt, Other, No pavement
Surface deficiency	Normal, Other (e.g. holes, soft, and so on)
Surface condition	Dry, Wet or other
Obstruction	Yes, No (within 15 meters)
Sight distance	Good, Poor (based on road design speed)
Signal type	Regular, Flash, No signal
Signal condition	Normal, Abnormal, No signal
Median	Island, Marker, Marking, None
Roadside marking	Yes, No
Weather	Sunny or cloudy, Rainy, Other
Illumination	With light, No light
Alignment	Straight, Curved, Other
Accident severity	Death involved, Injury only

4.4.2 Classification with Rough sets

The Taiwan 2005 SAV accident data was first analyzed with rough sets theory to generate a minimum rule set covering all objects. This analysis consisted of two steps: variable selection and rule induction. The former step was to identify the variables that were unable to differentiate the accident severity. In the analysis, four out of 25 variables were

identified as redundant, including *pavement material*, *surface deficiency*, *signal condition*, and *weather condition*, which may arise from the following two reasons. First, their effects could be replaced by other variables. For example, the effect of the *weather* variable could be substituted by that of the *surface condition* variable since raining would result in wet surface. It is understood that the weather condition would affect not merely surface conditions; for example, strong wind or large snow fall would raise the difficulty on drivers' control of their vehicles. However, these weather conditions rarely appear in Taiwan. The second reason was that these redundant variables had no significant impact on accident severity. For example, 98.6% and 98.5% of the accidents were reported on roads with an asphalt pavement and on roads without surface deficiency respectively. Therefore, the *pavement material* and *surface deficiency* variables were reported as redundant. After excluding the four redundant variables, the remaining 21 variables were considered in generating rules.

With 21 non-redundant explanatory variables, 315 rules were generated with rough sets theory to represent the 3,138 accident cases. This study applied the most frequently used algorithm – minimum covering – to generate rules. Its aim was to generate the minimum number as well as the shortest length of rules to cover all accidents. Of which, 295 rules were exact rules and 20 were approximate rules. An exact rule refers to a situation that the severity of an accident could be identified under a particular circumstance. On the other hand, an approximate rule represents a certain circumstance under which the accident severity could not be uniquely determined.

The rule support histogram was shown in Figure 4-2, where the number of rules in the vertical axis is shown against the number of support, the horizontal axis. The right-skewed shape showed that most rules were of low support. It suggests that most SAV accidents hold relatively unique patterns. On the other hand, some rules showed high support even though 21 factors were considered.

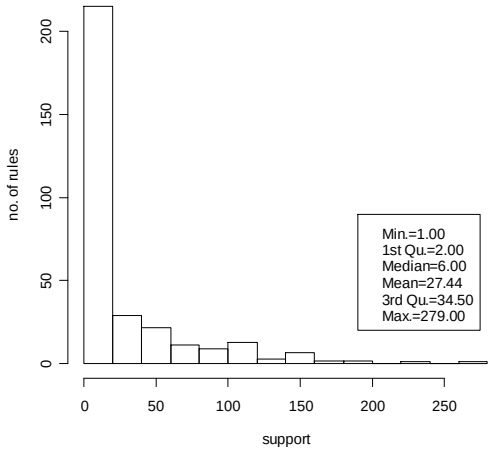


FIGURE 4-2 Rule support.

4.4.3 Determination of Rule Support Threshold for Differentiating Accidents

For the purpose of analysis, the accident cases were separated into two subsets: one subset includes accidents of support high enough such that their relationship could be claimed; the other subset consists of the remaining accidents whose relationship may not exist. The choice of threshold of rule support was determined by examining the average hit rate of accidents related to different levels of support. The whole data were first tested. Second, accidents related to rule support of one were excluded, and the remaining accidents were tested. Then, accidents related to rule support less than or equal to two were excluded, and the remaining accidents were tested. The test continued until the accidents related to rule support less than or equal to nine were excluded. In each test, decision trees were employed to obtain the average hit rate with Monte Carlo simulations of 2000 times; 75% of cases were selected for training and the remaining 25% of cases were adopted for testing for each simulation[**]. Moreover, a reference average hit rate was created for comparing the improvement. A reference hit rate was obtained by testing data randomly selected from the original dataset with specified sample size and injury/death case ratio. The sample size and injury/death ratio was determined by the aforementioned dataset selected by rough sets rules as shown in Table 4-9.

TABLE 4-9 Dissimilar Strong Rules Leading to Death or Other

Data	Included cases	Sample size			Injury/Death ratio
		Total	Injury	Death	
Whole	Whole	3138	2834	304	9.32
G1	Support > 1	3010	2776	234	11.86
G2	Support > 2	2940	2772	168	16.50
G3	Support > 3	2907	2771	136	20.38
G4	Support > 4	2867	2767	100	27.67
G5	Support > 5	2837	2755	82	33.60
G6	Support > 6	2800	2741	59	46.46
G7	Support > 7	2773	2736	37	73.95
G8	Support > 8	2757	2720	37	73.51
G9	Support > 9	2725	2715	10	271.50

The average hit rate was shown in Figure 4-3. The hit rate for data selected by rough sets rules was illustrated with solid lines; the reference hit rate was drawn with dotted lines. It could be observed that the average hit rates were increasing with the exclusion of accidents related to low support rules, especially for the minority class – fatal accidents. Especially, when accidents related to rules greater than five (G5) or seven (G7), the average

[**] Stratified random sampling was employed to partition data into training and testing groups. That is, 75% of injury only cases were randomly chosen for training, and so for 75% of death only cases.

hit rate of death involved cases significantly increased as labeled with solid circles in the graph. Although the G7 point showed relatively significant increase, the G7 data consisted only 37 death involved cases. On the other hand, the G5 data contained 82 death involved cases and raised the hit rate from 0.2 to around 0.5. Therefore, the support of six was considered as the threshold to differentiate between rules. That is, accidents related to rules with support greater or equal to six were considered as high-support-rule accidents.

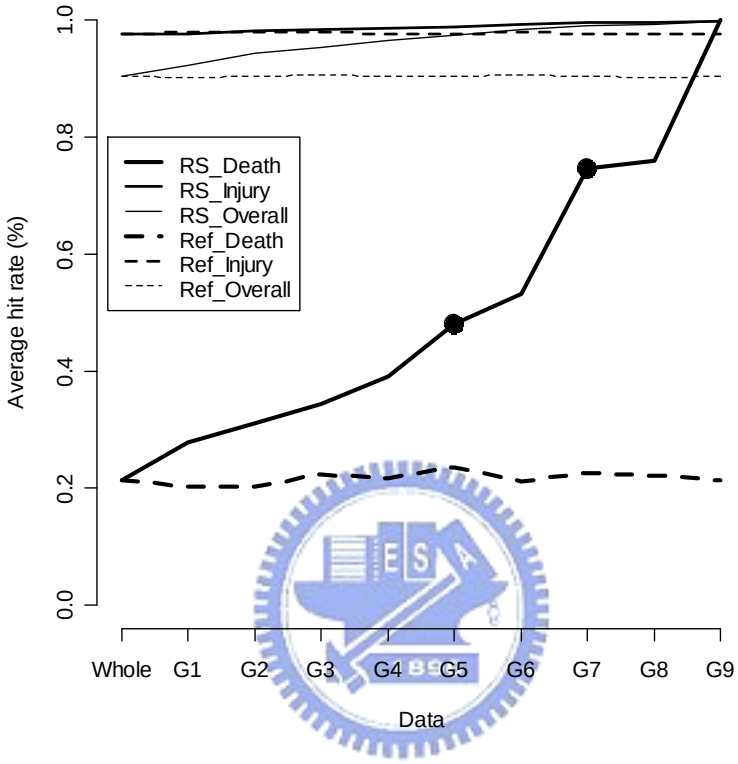


FIGURE 4-3 Average hit rate with respect to accidents related to rules with different support.

* RS_Death, RS_Injury, and RS_Overall refer to the average hit rate for death involved, injury only, and overall cases selected by rough sets rules, respectively. Ref_Death, Ref_Injury, and Ref_Overall refer to the average hit rate of reference for death involved, injury only, and overall cases, respectively.

4.4.4 Rule Comparison for High-Rule-Support Accidents

Among all the 315 rules, 164 of them were strong rules; 19 of those strong rules led to *death involved* or *other* accidents, and the remaining 145 strong rules led to *injury only* accidents. The following comparisons focused on the differences between *death involved* or *other* accidents and *injury only* accidents. In other words, the possible causal factors diverting an *injury only* accident to a *death involved* or *other* accident were examined

The rules having no similarity to *injury only* rules and the remaining 16 strong rules were demonstrated in the following two paragraphs, respectively.

1. Dissimilar death involved or other rules

There were three *death involved* or *other* rules having no similarity to *injury only* rules as listed in Table 4-10. The first dissimilar rule, D1, describes the young working drivers who were drinking and might be using cell phones driving on a curved road with poor sight distance but with lighting. While normal drivers would lower their speeds to safely pass a curve, the leading-to-death rule suggests that the corresponding driving speeds would not be low. Moreover, the curved road with poor sight distance raised the difficulty of driving. Although there were another 10 strong rules relating to curved roads and leading to *injury only* cases, none of them were specified as young drinking drivers. This might suggest that these drivers can easily misjudge the safe driving speed and can not properly maneuver the vehicle while passing a curve with a poor sight distance.

Seen in Table 4-10, the D2 and D3 rules describe the corresponding *death involved* accidents occurring under the condition that the drivers were not wearing seatbelts and were possibly drinking driving. Fastening the seatbelt and drinking driving have long been critical policy issues for the government of Taiwan; violating either one, especially the latter, leads to a substantial fine. Therefore, it is expected that these two unlawful behaviors occurring at the same time, as described in D2 and D3, will be rare. However, committing both these violations, whether combined with an unfriendly road environment or not, a *death involved* case would likely occur.

TABLE 4-10 Dissimilar Strong Rules Leading to Death or Other

Rule / Attribute[1]	D1	D2	D3
Age	Young	--	--
Occupation	Working	--	--
Seat belt use	--	Not using	Not using
Cell	Unknown	Unknown	--
Drink	Drinking	Unknown	Unknown
Road type	--	--	Rural
Sight distance	Poor	--	--
Illumination	Yes	--	Yes
Alignment	Curved	--	--
Severity	Death	Death	Death

[1] The attributes where all the three rules were unspecified were not represented to reduce the space.

2. Similar death involved or other rules

There were 16 *death involved* or *other* rules similar to *injury only* rules as listed in Table 4-11. The S1 and S2 rules were the rules most similar to *injury only* rules; these two rules had been cited as similar rules by injury only rules for 47 and 46 times, respectively. The rule S1 illustrated the condition that regular-valid-licensed young male working drivers driving with unspecified purposes and wearing seatbelts had been drinking alcohol and were driving around midnight on straight rural roads at low speed limits, dry surface, median

marking, and no signals. Although this describes drinking and driving behaviors, drinking itself can not fully represent the cause shifting the accident to a fatal one. By looking into the strong rules, some of them also related to drinking and driving behavior; however, as long as the drivers were not young people, it was not midnight, the quality of the corresponding road environment was not poor (i.e. it was a urban road, a road with a median island, or a road at a higher speed limit), or the surface was not dry, the accident severity was shown to be *injury only*. When the driver is young, the corresponding behavior could be somewhat risky and a more risky driving environment is usually associated with midnight driving (Lin and Fearn, 2003). Moreover, a road with poor quality could not mitigate the bumping impact of an accident; and when the surface is dry, it might encourage fast driving especially under low traffic (midnight on rural roads). Therefore, the combined unfavorable factors led to *death involved* accidents.

As stated, the rule S2 illustrated a condition very similar to S1. These two rules were almost identical except that the rule S2 did not specify the drinking behavior, but specified that the corresponding road environment may encourage fast driving – low traffic and good sight distance (around midnight driving along a straight rural road with illumination and roadside marking). Though the corresponding driver was not specified as drinking, the possibly more speedy driving behavior also led to *death involved* accidents.

In contrast to the first two rules, the rules S3 and S4 illustrate the accidents occurring on high-quality roads (highways or urban roads with median islands). The driving speeds on these roads are usually high especially on highways with a minimum speed of 80 kph. The high driving speeds combined with the impaired maneuvering skills, as well as lower situational awareness due to drinking, once an accident occurs, a *death involved* case is expected. When compared to their similar rules, these *death involved* cases could be merely *injury only* if the driver was not a young male (middle-aged, elderly or female), if the road was narrower (an urban road without roadside marking), or if the road did not mislead drivers to drive at an inappropriately high speed. Having either one of the factors could reduce the driving speeds or make the drivers drive more carefully.

The rules S5, S6 and S7 describe the conditions that the accidents occurred on low-speed-limit rural roads or in a low traffic environment (midnight) except that the trip purposes were unspecified, the drinking conditions were unknown, and the seatbelt usages were unknown. By looking into their similar rules, all else equal, the S5, S6 and S7 accidents became *injury only* if the driver did wear a seatbelt or if the driver was certainly not drinking. This addresses the effect of injury prevention by wearing a seatbelt and avoiding the deteriorated maneuvering skills as well as lower situational awareness due to drinking.

TABLE 4-11 Strong Rules Leading to Death or Other

Rule / Attribute	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
Age	Young	Young	Young	Young	Young	Young	Young	Young	--	--	--	--	Young	Middle	Young	--
Gender	Male	Male	--	Male	Male	Male	--	--	Male	Male	Male	--	Male	--	--	--
License type	Regular	Regular	Regular	Regular	Regular	--	Regular	--	--	Regular	Regular	--	--	--	--	--
License con.	Valid	--	--	--	--	--	--	--	--	--	--	Valid	--	--	Valid	--
Occupation	Working	Working	--	Working	--	Working	Working	Working	Working	--	Working	Working	Working	--	--	--
Purpose	Other	Other	Other	--	--	--	--	--	--	Other	--	--	--	--	--	--
Time	Midnight	Midnight	Midnight	Midnight	--	Midnight	--	DOP	Midnight	--	--	NOP	--	--	Midnight	Midnight
Protection	Using	Using	--	Using	Unknown	Unknown	Unknown	Using	--	--	Unknown	--	--	--	--	Unknown
Cell	--	--	Not using	Not using	--	Unknown	--	--	Unknown	Unknown	--	Unknown	Unknown	Unknown	Unknown	--
Drink	Drinking	--	Drinking	Drinking	Unknown	--	Unknown	--	--	--	Unknown	Drinking	Unknown	Unknown	Unknown	--
Road type	Rural	Rural	Urban	Highway	Rural	--	Rural	Highway	Highway	--	--	--	--	Rural	--	--
Speed	-50	-50	--	--	-50	--	--	80+	--	--	51-79	-50	--	-50	51-79	51-79
Road shape	--	Segment	Segment	Segment	Segment	Segment	Segment	Segment	--	--	Segment	--	--	Segment	--	Other
Surf. status	Dry	Dry	--	Dry	--	--	--	Dry	--	--	--	--	--	--	--	--
Obstruction	--	--	No	--	--	No	--	--	--	No	--	--	--	--	--	--
Sight dist.	--	--	--	--	Good	Good	Good	--	--	Poor	Good	--	Good	--	--	--
Signal type	No	No	--	--	No	--	No	--	--	--	--	--	--	--	--	--
Median	Marking	Marking	Island	--	--	Marking	--	Island	Island	--	--	--	--	--	Island	Island
Rd. side	--	Yes	Yes	--	--	Yes	Yes	--	--	--	--	--	No	--	--	--
Illumination	--	Yes	Yes	--	Yes	--	--	--	--	No	--	--	Yes	--	--	--
Alignment	Straight	Straight	Straight	--	--	--	--	--	--	--	--	--	--	--	--	--
Severity	Other	Other	Other	Other	Death	Death	Death	Other	Death	Death	Death	Death	Death	Death	Death	Death
Similarity[1]	47	46	18	16	11	7	7	7	3	3	3	2	1	1	1	1

[1]Similarity referred to the number of rules which were similar to this rule but led to injury only crashes.

The rule S8 describes young working people driving on highway segments with a dry surface during day off-peak periods and wearing seatbelts. When compared to the similar rules, all else equal, the accidents became *injury only* cases if the driver was certainly not drinking, if the driver owned an occupational or military driving license, or if the trip time was during the afternoon peak hours. Only soldiers in charge of driving can obtain a military driving license, therefore, under a high-speed-driving environment, drivers with occupational or military licenses are expected to be more capable to avoid fatal accidents than normal drivers once an accident occurs. Moreover, the traffic flow during peak hours is denser than that during off-peak hours; consequently, the corresponding driving speed is expected to be lower. Once an accident occurs, the severity should be less severe.The rule S9, similar to S8, describes the accidents that occurred on highways, but the drivers were specified as male drivers instead of young drivers; moreover, the trip time was around midnight rather than off-peak periods during the day. When compared to its similar rules, the S9 accidents could become less severe if the trip time was during afternoon peak periods. The denser traffic during peak hours might restrict the driving speed. Even though the drivers could be of high risk (young or male drivers), the environment might limit their driving speeds and the corresponding accidents might not be fatal.

The rule S10 describes the regularly-licensed male drivers driving on poorly-sighted roads without any obstructions on the roads. When compared to its similar rules, all else equal, the accidents could be less severe if there were obstructions on the roads. According to the definition, obstructions are defined as any obstacles within 15 meters of the crash. This distance is much shorter than the defined safe sight distance which is 45 meters under a normal 40-kph driving speed, and a driver might spot the obstacles and lower his/her driving speed. On the other hand, the male drivers driving at relatively high speeds, even though the road has poor sight distance, result in a fatal accident.

The rule S11 describes regularly-licensed working people driving on a medium-speed-limit road with good sight distance. Its similar rules suggest that these accidents could be less severe if the drivers were certainly not drinking. Similarly, the accidents under the rules S12 and S13 would be less severe if the drivers were certainly not using cell phones or not drinking driving. The accidents under the same driving environment described by S15 were less severe if the drivers were the elderly, who are usually considered to be of lower risk than young drivers. Even under a road encouraging fast driving (medium speed limit with median island), the elderly drivers might drive carefully and maintain a reasonable driving speed while the young drivers might not.

The information provided by the remaining rules, S14 and S16, is relatively vague since most attributes were unspecified and all the behavioral attributes were either unspecified or unknown. Moreover, the associated similar rules were different in behavioral

attributes. Therefore, it is relatively difficult to tell the differences between the selected rules and their associated similar rules.

4.4.5 Logistic Regression Analysis for the Remaining Accidents

Different from the accident cases with strong causal relationships, the 363 accidents associated with the weak support rules or the approximate rules were analyzed with regression methods to investigate the possible associations between factors and extract the variations due to insufficient information. In particular, binary logistic regression models were adopted. The model structure was revised from the one proposed by Kim *et al*. (1995) where the accident severity was affected by driver characteristics, trip characteristics, behavioral factors, environmental factors, and interactions between driver and behavioral factors. Backward elimination was applied to select variables.

The reference severity was *injury only* and the estimation results were summarized in Table 4-12. The estimated Hosmer-Lemeshow *p*-value was 0.293 (> 0.100) which indicated the goodness of fit was acceptable. The final variables included age, trip time, signal type, surface status, median, roadside marking, and the interaction between age and drinking. The results showed that accidents with rarely occurring patterns and those with frequently occurring patterns were different. Young drivers were less likely to be involved in a *death involved* case provided that they were not drinking. Yet, under the condition that the young drivers were drinking, they would be more likely to be involved in a *death involved* case. Moreover, the accidents occurred around midnight (compared to other time periods) were less likely to be involved in a *death involved* accident. These two results contradicted the results of the previous section that young drivers and midnight accidents were death-prone, which may imply distinct features between these two types of drivers.

Furthermore, accidents occurring on roads having a dry surface (compared to wet or other surface conditions) and with roadside marking (compared to roads without roadside marking) were less likely to be *death involved* accidents. On the other hand, those accidents that occurred on roads with warning flash signals (compared to no signals) and with median markers (compared to no medians) were more likely to be *death involved* accidents. A road with warning flash signals indicates possible traffic conflicts within the area and the signals warn the drivers to pay attention. In addition, a road with median markers implies that this section of the road is rather dangerous, and the markers warn the drivers not to drive across the centerline. These results suggested that a better road environment seems to help prevent such death involved accidents.

TABLE 4-12 Logistic Regression Estimation Result[1]

Parameter	Estimate	P-value	Odds		
			Odds ratio	95% Wald confidence interval	
Intercept	2.841	<.0001**[2]	17.124	6.426	49.844
Age (Young vs. Middle or Old)[3]	-1.099	0.002**	0.333	0.164	0.662
Trip time (Midnight vs. Other)	-0.786	0.004**	0.456	0.267	0.777
Signal type (Regular vs. None)	-0.548	0.216	0.578	0.243	1.377
Signal type (Flash vs. None)	1.583	0.040**	4.871	1.072	22.137
Surface status (Dry vs. Other)	-0.942	0.009**	0.390	0.193	0.787
Median (Island vs. None)	0.448	0.336	1.565	0.628	3.899
Median (Marker vs. None)	1.452	0.015**	4.271	1.320	13.821
Median (Marking vs. None)	0.186	0.690	1.204	0.484	2.997
Roadside marking (Yes vs. No)	-1.191	0.000**	0.304	0.157	0.589
Age*Drink (Drinking vs. Not drinking)	0.716	0.036**	2.047	1.047	4.002
Age*Drink (Unknown vs. Not drinking)	1.196	0.015**	3.308	1.267	8.635

[1] Goodness-of-fit test: Hosmer-Lemeshow *p*-value = 0.2933

[2] ** 0.05 significance level

[3] The latter term in brackets refers to the reference;

Chapter 5 ISSUES

The purpose of this chapter is to discuss the issues related to the methodologies presented in Chapter 3 and the empirical findings demonstrated in Chapter 4. The connection between rough sets rules and accident chains are discussed in Section 5.1. The heterogeneity of accident data are shown in Section 5.2; and the issue of aggregation bias is presented in Section 5.3. Finally, the confounding effects are discussed in Section 5.4.

5.1 Connection between Rough Sets Rules and Accident Chains

Taking advantages of rough sets, this research implemented the idea that the occurrence of an accident is a series of errors or mishandling. The illustrated case shows that it is feasible to apply rough sets theory to analyze the links among affecting factors and accident types. The proposed factor structure can be easily transformed and extended based on an analyst's knowledge and his/her on-hand accident databases. Any factor structures can be tested by similar steps proposed in this research. In addition, a large number of condition attributes were included without any prior judgments except when being grouped with respect to the temporal and logical sequence of the occurrence of an accident. A condition attribute was dropped only when the removal did not have any impact on defining accident types. In our empirical study, only one redundant condition attribute (pavement material) was found when all the attributes were included. This procedure differs from conventional statistical approaches where non-significant attributes are usually immediately dropped and are sometimes claimed to have no impact on the occurrence of an accident.

Rules generated from rough sets provide fruitful information describing conditions under which certain type of accidents may occur. For example, as mentioned in the previous section, the most significant rule for the bump-into-work zone suggests that there is a relatively high risk when a driver approaches work zone on a road with speed limit less than 50 (kph) around midnight. When it comes to employment of the modern ITS technologies (FHWA, 2006), specific warning messages could be devised and sent to the drivers conforming to this particular scenario; consequently, the potential accidents could be prevented. In short, the derived rules have the potential to distribute the right information to the right drivers at the right time for them to be able to act properly.

On the other hand, hundreds of rules were generated in the end, which makes it difficult for analysts to conclude which rules or accident patterns are the most significant. This result may partly come from the fact that some accident types, such as the bump-into-non-fixed object accidents or rollover accidents, are so stochastic and unique, and partly from the lack of detailed information about drivers' characteristics in the database

that hinder the possibility of more effectively recognizing accident characteristics. Despite the fact that these accident types are the least definable and the least classifiable, some protective measures still can be implemented to reduce the accident possibility and severity such as preventing animals crossing roads or increasing the strength of the vehicle roof. On the other hand, the most definable and recognizable accident type – the bump-into-facility accidents – is regarded as being preventable. In addition, the bump-into-bridge and off-road accidents showing similar classification patterns as the bump-into-facility accidents, are also expected to be preventable.

In order to find representative rules for occurrence of those avoidable accident types, more advanced rough sets models, such as the hybrid approach combining rough sets with genetic programming (Mckee and Lensberg, 2002), can be adopted in future research. However, for the low-performing (unpredictable) accident types which are highly related to driver characteristics and unpredictable environment conditions (i.e. non-fixed objects), more related data need to be collected for further study. Meanwhile, instead of preventing accidents, measures for reducing the negative effects of those unpredictable accidents may be more effective and are worth investigating.

The estimation results showed that the accuracy of approximation, the quality of approximation and the hit rates could be dramatically enhanced by considering at least two sets of condition attributes while the inclusion of overall condition attributes generally gave the most satisfactory quality of classification. This suggests that collecting more detailed data on some specialties rather than aimlessly increasing survey items is more effective. Nonetheless, additional attributes are welcomed and could be collected and examined by testing their redundancy and their effect on the accuracy of approximation, quality of approximation as well as hit rates to determine whether they are worthwhile.

5.2 Heterogeneity of Accident Data

The heterogeneity discussed in this manuscript is different from past studies. It is based neither on driver characteristics (such as age or gender) nor on environmental characteristics (such as urban or rural roads). Instead, the heterogeneity in the study originates from a hypothesis in which the features for frequently repeated processes of accident occurrence and for sparsely unique processes of accident occurrence may be essentially different. The distinct features of accident groups uncovered in this empirical study did show the possible existence of such heterogeneity. The accidents associated with weak rules occur rather uniquely. Since they occur by chance and tend not to lead to similar consequences under similar processes and conditions, it is intuitively expected that it would be relatively inefficient to devise the corresponding countermeasures for them. Surprisingly, it is

observed that those accidents are heavily related to road environment and could be possibly improved by carefully providing adequate road facilities.

Countermeasures for traffic accidents have been previously either focused on drivers who break laws such as drunk driving or speedy behaviors or are concentrated on road design to build a smooth road. Although these measures are generally known and effective, less attention is put on identifying the risky but rational drivers associated with the strong pattern accidents. That means more research and information from studies is required to identify this type of drivers and specific measures devised for them to prevent accidents. It is noted that preventing accidents associated with weak patterns is as crucial as preventing those with strong patterns. However, the efficiency of specifically designed countermeasures to prevent accidents related to the strong patterns will be prominent since accidents associated with the weak patterns are highly diverse. Thus, when detailed heterogeneous accident information is taken into account, countermeasures, such as on-board warning messages and smart roadside safety facilities which try to provide right safety information to right drivers at right statuses, are expected to be effective for the occurrence of strong pattern accidents and are worth being studied.

5.3 Aggregation Bias

The issue of aggregation biases has been noticed and studied by many studies (Davis, 2004; Hewson, 2005; Vlahogianni et al., 2004; Walker and Catrambone, 1993), of which Davis (2004) presented a thorough discussion using simulated data. He argued that since accident data have no independent status, the statistical regularities are simply the result of aggregating particular types and frequencies of mechanisms. The aggregation step implemented in this study could raise similar issues. Despite of the difficulty, aggregation does lay a concrete basis for understanding accident scenarios and further studying those associated with strong pattern with detailed design experiments.

Analyzing each rule instead of accident groups provides a possible way to alleviate such problems. Each rule is herein treated as an individual mechanism since rules are derived under the condition that many critical factors have been controlled. By examining the characteristics of each rule classified as strong patterns, most rules are found to support the findings from crosstab analysis and multinomial logistic regression models where accidents with strong patterns indicate that the drivers involved are somewhat high-risk. This suggests that the proposed approach can be effective in processing the heterogeneous accident data, although the aggregation bias issue must be faced.

It is unfortunately far more difficult to interpret individual rules with weak and

medium strength since the number of rules runs into the hundreds. An alternative way is to loosen up a little on the pattern requirements after the most (and least) important attributes have been identified. This can be achieved by using an index called significance of attributes (Pawlak, 1991). This index evaluates the number of objects which can not be distinguished with the elementary sets while one condition attribute is dropped from the model. In doing so, the number of rules is expected to decrease. However, the thoroughness of the process of accident occurrence described by the rules will also decrease at the same time. The issue of overwhelming number of rules derived from rough sets theory has also been noticed by researchers (Løken and Komorowski, 2001) and requires further studies.

5.4 Confounding Effects in Causality Analysis

Finding causal factors on safety in observational studies, especially in cross-section studies, is an unresolved issue (Hauer, 2006). The main difficulty lies in the numerous confounding effects while doing comparisons. Consequently, if the majority of these attributes is not well controlled, the analysis results would be biased.

As an attempt to resolve this issue, this research identified the possible causal factors by comparing the differences between entire accident patterns instead of estimating the marginal effects of each attribute. Based on rough sets analysis, the accident data was separated into two subsets: one contained the accidents which could be fully described by the on-hand information and consisted of a certain number of accidents representing the possible existence of causality; the other contained the remaining accidents. The rules, derived from the rough sets analysis, were then compared with each other. The comparison design was used to find the most similar rules for each rule and to examine the differences. This allowed the control of many confounding factors as possible, and partially revealed the differences between what happened and what would have happened had the circumstances in question been different.

Since the causal factors were found by comparing the complete rules, it is obvious that the comprehensiveness of on-hand data determines to what extent the confounding effects are controlled. In our empirical study, 23 attributes were considered. These attributes were presumed to have impact on accident occurrence and examined with rough sets theory to determine whether some of them were redundant. Basically, more information is welcome in such research provided that it is relevant to the decision attribute. Moreover, there is theoretically no limitation in the attributes that rough sets theory can adopt as long as the computational time is tolerant. Yet, it should be noted that including attributes with similar meanings could produce unnecessary rules and impede the interpretations. For example, two rules with all other things are equal except that one rule specifies the road type as a

freeway and the other rule specifies a high speed limit which could only show up on freeways. There is no difference between these two conditions in the real world. A careful selection of the entry attributes could avoid such redundancy.

Chapter 6 CONCLUSION AND RECOMMENDATION

The objectives of this research were to propose an approach for identifying accident patterns and exploring their characteristics and to propose an approach for examining accident causality. The summary of the work performed in this research was described in Section 6.1. Recommendations for further research were drawn in Section 6.2.

6.1 Conclusion

In this study, accident characteristics and causality were examined by analyzing accident chains derived from cross-sectional databases. The contributions and findings related to methodologies in this study were summarized in the following points:

1. Taking advantage of rough sets theory, this study proposed a research framework which could effectively examine the characteristics of cross-sectional accident databases from chain perspective. In particular, the variations of rough set indicator values with respect to different sets of condition attributes indicate the similarities and differences of the underlying accident generating process among accident types or severities. They also provide the information about the usefulness of considered attributes in identifying accident chains as well as the randomness of accident chains embedded in a database. These indicators include lower and upper approximation, accuracy of approximation, quality of approximation, number of generated rules, and hit rates.

2. Rules generated from rough sets theory provide fruitful information describing conditions under which certain type of accidents may occur. The illustrated case shows that it is feasible to analyze the links among affecting factors and accident consequences by interpreting the derived rules. However, it should be noted that the quality of derived rules depends on the comprehensiveness of on-hand data, and the accuracy of rule interpretations depends on analysts' professional knowledge.

3. It is a fundamental belief in all statistics that non-significant factors, do not bias estimation results. However, some studies have confused "non-significant factors in statistical sense with unimportant in common sense." The rough sets theory provides an alternative way to account for the importance of factors.

4. From the perspective of accident investigation, the entire causal chain for each accident is the primary focus. But from the perspective of applications, some effective measures to reduce accident occurrences are eager to be devised. Although the features of individual rules do not completely agree with the results from multinomial logistic

regression models, most rules did support the findings from the cross-tab analysis and the multinomial logistic regression models. Understanding contributing factors for those large member rules, therefore, can be advantageous.

5. Comparing the features of rules would reveal the differences between what happened and what would have happened had the circumstances in question been different. These differences might imply causal relationships. The proposed approach provides an alternative to examine causality from cross-sectional databases, which have been considered an unresolved issue in pass studies.

This study mainly examined the characteristics of Taiwan's single auto-vehicle crashes. The findings were summed up in the following points:

1. The occurrence of crashes with facility may follow similar paths and is more predictable. But for other accident types, the rules generated from their training cases may not be representative since their occurrence are mostly random. Moreover, except for the crashes with facility where more information is useful, different accident types have their corresponding useful condition attributes. In addition, some similarities may exist in the occurrence of the crashes with bridge, with facility, off-road and rollover types since they are all related to road geometry and driving environments.

2. Student drivers who are young and less experienced exhibit a relatively high possibility of being involved in off-road accidents under normal driving environment, i.e. no particularly unfavorable factors such as drinking driving or poor sight distance show on the chain. Since other driving groups such as working people do not show similar accident patterns, the government should seriously consider educating student drivers to enhance their situational awareness of driving environment and reduce their risk-driving behavior on roads. The result echoes the graduated licensing scheme currently implemented in many countries.

3. Most Taiwan's single auto-vehicle accidents occurred with different driver, trip characteristics and/or different behavior and environmental factors. Nevertheless, there is still a large portion of accidents occurring repeatedly with identical patterns. The large member rules justify considerable efforts at intervention and that behavioral interventions could be applied to a large number of collision types with similar causal patterns.

4. Accidents should not be resolved by single factor, but by a chain of factors. Previous countermeasures focused mostly on behavioral and environmental proximal factors. It is effective; however, to further improve road safety, all factors associated in the factor chain may need to be taken into serious consideration. Furthermore, neglecting factors

in a chain may result in rather different stories and blur the interactions among accident features.

5. Significantly different features were shown between frequently repeated and unique rules for Taiwan's single auto-vehicle accidents. The drivers involved in accidents with frequently repeated rules reflected the characteristics of high-risk drivers shown in past studies. These characteristics were not limited to driver characteristics and included all critical factors related to accident occurrences. On the other hand, it is road conditions that played the key role in accidents associated with unique rules. That is to say, certain road conditions are safe under average circumstances. However, when combined with other risk factors, though it rarely happened, the safe road conditions may still become dangerous. This suggests that road design, road furniture, road maintenance, traffic control and work zone setup should be considered in a more comprehensive perspective; and as a consequence, there may be fewer accidents corresponding to unexpected circumstances.

6. Although not shown significantly in causal patterns, highway interventions are suggested via the rules of low frequencies. It is not saying that these interventions are unimportant. Instead, the improvement of road maintenance quality may prevent such accidents. Highway interventions should be considered in a more comprehensive perspective; and as a consequence, accidents corresponding to unexpected circumstances could be reduced.

7. There are some culture differences in drinking between Eastern and Western countries. The young Taiwanese do not drink as much as their counterparts in Western countries. Due to business and social activities, however, middle-aged Taiwanese are more likely to drink. Therefore, although young drivers involved in single auto-vehicle accidents numbered twice the middle-aged drivers, only 30.1% of the young drivers were drinking driving compared to 37.1% of the middle-aged. The majority of male drinking driving accidents appeared in the medium strength rules, not in strong strength rules. This suggests that the accident patterns related to drinking driving are associated with various diverse circumstances. A possible explanation could be situational awareness and maneuvering skills deterioration due to drinking.

8. Instead of one single factor, the combinations of unfavorable factors would be the causes leading to fatal accidents including young, male or less experienced, their behaviors of drinking, wandering on roads around midnight, and overestimating their own driving capabilities and underestimating the possible dangers hidden in the environment. Furthermore, the distinct features were shown between the accidents related to rules with high support and those with low support. A better road

environment would be helpful to prevent fatal accidents for the latter kind of drivers, but not necessarily for the former kind of drivers.

6.2 Recommendation

Although this study has taken a step forward in the direction of examining accident characteristics and causality from chain perspective, some limitations should be noticed and some findings are worth further studies.

1. This study is a new attempt to apply rough sets as a complementary tool for accident analyses. A lot of information is still embedded in the derived rules that might provide useful knowledge for researchers and analysts and requires further exploration. Advanced models, however, should be considered in the future to improve and to address the issues related to performance of rule extraction and case validation. For example, besides the *AND* operator, one could consider other logic operators such as *OR* or *NOT* into rule generations.

2. The proposed approaches can be adopted in other datasets or be used to analyze different accident outcomes. These approaches were analyzed in analyzing single auto-vehicle accidents. Such accidents involve only a single vehicle and thus the underlying process of accident occurrence would be much simpler than other accident types such as multi-vehicle accidents. One should carefully examine her or his on-hand data to determine which subjects to examine.

3. Comparisons between rough sets and other methodologies would be very interesting. However, it is necessary to have a very careful design to conduct these comparisons; particularly, the nature of rough sets theory is quite different from other methodologies.

4. Possible aggregation biases and the overwhelming numbers of rules have limited this research. The derived rules could help reveal the aggregation bias in the process of retrieving contributing factors. To resolve the issue of aggregation bias and shed light on the whole features of accidents by using the rule based approach, however, needs further research.

5. Although this approach allows the control of all relevant factors, it does not mean that the findings under this approach must be the true causal factors. The primary reason is the limited information provided by accident databases. Accidents are observable only after they have occurred. Some information is thus difficult to obtain especially for the fatal accidents. For example, vehicle features are critical to accident severity, and exposure data are critical to claim the relatively high frequency of a rule. But it is a

pity that such information is not provided in the database. Consequently, the uncontrolled confounding factors should be carefully taken into account in ascertaining the findings and require further studies.

6. Experimental designs for exploring driving behaviors would be helpful to complement the aforementioned shortcoming. In particular, these designs could be based on the interested rules; for example, the most significant rule leads to fatal accidents. Since a rule contains rich information, the corresponding experimental design would be specific and effective.

REFERENCES

Abdel-Aty, M.A., Abdelwahab, H.T., 2000. Exploring the relationship between alcohol and the driver characteristics in motor vehicle accidents. Accid. Anal. Prev. 32 (4), 473-482.

Abdel-Aty, M.A., Abdelwahab, H.T., 2004. Analysis and prediction of traffic fatalities resulting from angle collisions including the effect of vehicles' configuration and compatibility. Accid. Anal. Prev. 36 (3), 457-469.

Ajzen, I., 1985. From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), Action control: From cognition to behavior (pp. 11-39). Heidelberg: Springer.

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accid. Anal. Prev. 34 (6), 729-741.

Baker, J.S., Ross, H.L., 1961. Concepts and classification of traffic accident causes (Part 1). International Road Safety and Traffic Review 9 (31), 11-18.

Beirness, D.J., Simpson, H.M., Mayhew, D.R., 1993. Predicting crash involvement among young drivers. In Utzelmann, H.D., Berghaus G., Kroj G., Alcohol, Drugs and Traffic Safety, Proceedings of the International Conference on alcohol, Drugs and Traffic Safety- T'92, Verlag TUV Rhineland, Cologne, 885-890.

Berg, H.Y., Gregersen, N.P., Laflamme, L., 2004. Typical patterns in road-traffic accidents during training – An explorative Swedish national study. Accid. Anal. Prev. 36 (4), 603-608.

Blockey, P.N., Hartley, L.R., 1995. Aberrant driving behaviors – errors and violations. Ergonomics 38 (9), 1759-1771.

Bolstad C.A., 2001. Situation awareness: Does it change with age? Presented in the Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting.

Boyle, L.N., Mannering, F., 2004. Impact of traveler advisory systems on driving speed: Some new evidence. Transp. Res. Part C 12 (1), 57-72.

Chandraratna, S., Stamatiadis, N., Stromberg, A., 2006. Crash involvement of drivers with multiple crashes. Accid. Anal. Prev. 38 (3), 532-541.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accid. Anal. Prev. 38 (5), 1019-1027.

Davis, G.A., 2004. Possible aggregation biases in road safety research and a mechanism approach to accident modeling. Accid. Anal. Prev. 36 (6), 1119-1127.

Davis, G.A., Swenson, T., 2006. Collective responsibility for freeway rear-ending accidents?: An application of probabilistic causal models. Accid. Anal. Prev. 38 (4), 728-736.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 (3), 434-444.

Drummer, O.H., Gerostamoulos, J., Batziris, H., Chu, M., Caplehorn, J., Robertson, M.D., Swann, P., 2004. The involvement of drugs in drivers of motor vehicles killed in Australian road crashes. Accid. Anal. Prev. 36 (2), 239-248.

Eby, D., Shope, J.T., Molnar, L.J., Vivoda, J.M., Fordyce, T.A., 2000. Improvement of older driver safety through self evaluation: The development of a self-evaluation instrument. UMTRI Report-2000-04, Transportation Research Institute, University of Michigan.

Eisenberg, D. 2004. The mixed effects of precipitation on traffic crashes. Accid. Anal. Prev. 36 (4), 637-647.

Elvik, R. 2002. The importance of confounding in observational before-and-after studies of road safety measures. Accid. Anal. Prev. 34 (5), 631-635.

Elvik, R. 2003. Assessing the validity of road safety evaluation studies by analyzing causal chains. Accid. Anal. Prev. 35 (5), 741-748.

Elvik, R. 2004. To what extent can theory account for the findings of the road safety evaluation studies?. Accid. Anal. Prev. 36 (5), 841-849.

Elvik, R., Greibe, P., 2005. Road safety effects of porous asphalt: a systematic review of evaluation studies. Accid. Anal. Prev. 37 (3), 515-522.

Elvik, R., Mysen, A.B., Vaa, T., 1997. Trafikksikkerhetshandbok. Tredje utgave, Transportøkonomisk Institutt, Oslo.

Elvik, R., Vaa., 2004. The handbook of road safety measures. Elsevier, Amsterdam.

Ferguson, S.A., Leaf, W.A., Williams, A.F., Preusser, D.F., 1996. Differences in young driver crash involvement in states with varying licensure practices. Accid. Anal. Prev. 28 (2), 171-180.

Fleury, D., Brenac, T., 2001. Accident prototypical scenarios: A tool for road safety research and diagnostic studies. Accid. Anal. Prev. 33 (2), 267-276.

Forward, S., Linderholm, I., Jarmark, S., 1998. Women and traffic accidents, causes, consequences and considerations. In Proceedings of the 24th International congress of Applied Psychology, 9-14 August 1998, San Francisco.

French, D.J., West, R.J., Elander, J., Wilding, J.M., 1993. Decision-making style, driving style, and self-reported involvement in road traffic accidents. Ergonomics 36, 627–644.

Fuller, R. 2005. Towards a general theory of driving behavior. Accident Anal. Prev. 37 (3), 461-472.

Garber, N.J., Ehrhart, A.A., 2000. Effects of speed, flow, and geometric characteristics on crash frequency for two-lane highways. Transp. Res. Rec. 1717, 76-83.

Glass, R.J., Segui-Gomez, M., Graham, J.D., 2000. Child passenger safety: Decisions about seating location, airbag exposure, and restraint Use. Risk Analysis, 20 (4), 521-527.

Golob, T.F., Recker, W.W., Alvarez V.M., 2004. Freeway safety as a function of traffic flow. Accid. Anal. Prev. 36 (6), 933-946.

Grzymala-Busse, J.W., 1992. LERS- a System for learning from examples based on rough sets. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publisher, Dordrecht.

Grzymala-Busse, J. W., Werbrouck, P., 1998. On the best search method in the LEM1 and LEM2 algorithms. In E. Orlowska (ed.), Incomplete Information: Rough Set Analysis, pp. 75-91. Heidelberg-New York: Physica-Verlag.

Gulian, E., Glendon, I., Matthews, G., Davies, R., Debney, L., 1988. Exploration of driver stress using self-reported data. In: Rothengatter, T., de Bruin, R. (Eds.), Road User Behaviour: Theory and Research. Van Gorcum & Co B.V., Assen, The Netherlands, pp. 342–347.

Gulian, E., Matthews, G., Glendon, A.I. Davies, D.R., Debney, L.M., 1989. Dimensions of driver stress, Ergonomics 32 (6), 585-602.

Hansotia, P., Broste, S.K., 1991. The effect of epilepsy or diabetes mellitus on the risk of automobile accidents. New England Journal of Medicine 324 (1), 22-26.

Harrison, W.A., 1998. The occupations of drink drivers: Using occupational information to identity targetable characteristics of offenders, Accid. Anal. Prev. 30 (1), 119-132.

Hauer, E., 1997. Observational before-after studies in road safety. Pergamon.

Hauer, E., 2004. The harm done by tests of significance. Accid. Anal. Prev. 36 (3), 495-500.

Hauer, E., 2006. Cause and effect in observational cross-section studies on road safety. Presented at 2006 TRB 85th Annual Meeting.

Heinrich, H., 1931. Industrial accident prevention. New York: McGraw-Hill.

Hewson, P., 2005. Epidemiology of child pedestrian casualty rates: Can we assume spatial dependence? Accid. Anal. Prev. 37 (4), 651-659.

Juarez, P., Schlundt, D.G., Goldzweig, I., Stinson, N.Jr., 2007. A conceptual framework for reducing risky driving behaviors among minority youth. Injury Prevention 12, 49-55.

Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accid. Anal. Prev. 34 (3), 357-365.

Karlaftis, M.G., Tarko, A.P., 1997. Heterogeneity considerations in accident modeling", Accid. Anal. Prev. 30 (4), 425-433.

Kass, S.J., Cole, K.S., Stanny, C.J., 2007. Effects of distraction and experience on situation awareness and simulated driving. Transp. Res. Part F 10 (4), 321-329.

Keall, M.D., Frith, W.J., Patterson, T.L., 2004. The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in New Zealand, Accid. Anal. Prev. 36 (1), 49-61.

Keall, M.D., Frith, W.J., Patterson, T.L., 2005. The contribution of alcohol to night time crash risk and other risks of night driving. Accid. Anal. Prev. 37 (5), 816-824.

Kim, S., Kim, K., 2003. Personal, temporal and spatial characteristics of seriously injured crash-involved seat belt non-users in Hawaii. Accid. Anal. Prev. 35 (1), 121-130.

Kim, K., Li, L., Richardson, J., Nitz, L., 1998. Drivers at fault: Influences of age, sex, and vehicle type. Journal of Safety Res. 29 (3), 171-179.

Kim, K., Nitz L., Richardson, J., Li, L. 1995. Personal and behavioral predictors of automobile crash and injury severity. Accid. Anal. Prev. 27 (4), 469-481.

Laapotti, S., Keskinen, E., 1998. Differences in fatal loss-of-control accidents between young male and female drivers. Accid. Anal. Prev. 30 (4), 435-442.

Laapotti, S., Keskinen, E., 2004. Has the difference in accident patterns between male and female drivers changed between 1984 and 2000? Accid. Anal. Prev. 36 (4), 577-584.

Laflamme, L., Eilert-Petersson, E., 1997. School-injury patterns: A tool for safety planning at the school and community levels. Accid. Anal. Prev. 30 (2), 277-283.

Lai, C.H., Huang, W.S., Chang, K.K., Jeng, M.C., Doong, J.L., 2006. Using data linkage to generate 30-day crash-fatality adjustment factors for Taiwan. Accid. Anal. Prev. 38 (4), 696-702.

Lin, M., Fearn, K.T., 2003. The provisional license: Nighttime and passenger restrictions – a literature review. Journal of Safety Res. 34 (1), 51-61.

Løken, T., Komorowski, J., 2001. Rough modeling – a bottom-up approach to model construction. Int. J. Appl. Math. Comput. Sci. 11 (3), 675-690.

Lyznicki, J.M., Doege, T.C., Davis, R.M., Williams, W.A., 1998. Sleepiness, driving, and motor vehicle crashes. The Journal of the American Medical Association 279 (23), 1908-1913.

Ma, R., Kaber, D.B., 2005. Situation awareness and workload in driving while using adaptive cruise control and a cell phone. Industrial Ergonomics 35, 939-953.

Massie, D.L., Campbell, K.L., Williams, A.F., 1995. Traffic accidents involvement rates by driver age and gender. Accid. Anal. Prev. 27 (1), 73-87.

Massie, D.L., Green, P.E., Campbell, K.L., 1997. Crash involvement rates by driver gender and the role of average annual mileage. Accid. Anal. Prev. 29 (5), 675-685.

McKee, T.E., Lensberg, T., 2002. Genetic programming and rough sets: A hybrid approach to bankruptcy classification. European Journal of Operational Res., 138 (2), 436-451.

McKenna, F.P., Waylen, A.E., Burkes, M.E., 1998. Male and female divers: How different are they? AA Foundation for Road Safety Research, The University of Reading, Hampshire, UK.

Mizell, L., 1997. Aggressive driving. Prepared for AAA Foundation for Traffic Safety, Washington, D.C.

Moskowitz, H., Fiorentino, D., 2000. A review of the literature on the effects of low doses of alcohol on driving-related skills. NHTSA Report No. DOT HS 809028, US Department of Transportation, Springfield, VA, USA.

Mountain, L.J., Hirst, W.M., Maher, M.J., 2005. Are speed enforcement cameras more efficient then other speed management measures? The impact of speed management schemes on 30 mph roads. Accid. Anal. Prev. 37 (4), 742-754.

Murray, A., 1997. The home and school background of young drivers involved in traffic accidents. Accid. Anal. Prev. 30 (2), 169-182.

Norris, F.H., Matthews, B.A., Riad J.K., 2000. Characterological, situational, and behavioral risk factors for motor vehicle accidents: A prospective examination. Accid. Anal. Prev. 32 (4), 505-515.

Ogle, J.H., 2007. Technologies for improving safety data. NCHRP Synthesis 367,

Sponsored by the American Association of State Highway and Transportation Officials in Cooperation with the Federal Highway Administration

Parker, D., Reason J.T., 1995. Driving errors, driving violations and accident involvement. Ergonomics 38(5), 1036-1048.

Pawlak, Z., 1982. Rough sets. International Journal of Computer and Information Science 11 (5), 341-356.

Pawlak, Z., 1991. Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers, The Netherlands.

Pawlak, Z., Skowron, A., 2007. Rudiments of rough sets. Information Sciences 177 (1), 3-27.

Pearl, J., 2000. Causality: Models, reasoning, and inference. Cambridge University Press, New York.

Preusser, D.F., Williams, A.F., Ulmer, R.G., 1995. Analysis of fatal motorcycle crashes: Crash typing. Accid. Anal. Prev. 27 (6), 845-851.

Proctor, S., Belcher, M., Cook, P., 2001. Practical road safety auditing. TMS consultancy and Thomas Telford Limited.

Reason, J., 1997. Managing the risks of organizational accidents. Ashgate, Aldershot.

Reason, J., Manstead, A., Stradling, S., Baxter, J., Campbell, K., 1990. Errors and violations on the roads: a real distinction? Ergonomics 33 (10), 1315-1332.

Shinar, D., Compton, R., 2004. Aggressive driving: An observational study of driver, vehicle and situational variables. Accid. Anal. Prev. 36 (3), 429-437.

Simoes, A., 2003. The cognitive training needs of older drivers. Recherche Transports Securite 79, 145-155.

Smith, K.M., Cummings, P., 2004. Passenger seating position and the risk of passenger death or injury in traffic crashes. Accid. Anal. Prev. 36 (2), 257-260.

Snyder, M.B., Knoblauch, R.L., 1971. Pedestrian safety: The identification of precipitating factors and possible countermeasures. DOT-FH-11-7312. Washington D.C. U.S. Department of Transport.

Sohn, S.Y., Shin, H., 2001. Pattern recognition for road traffic accident severity in Korea. Ergonomics 44 (1), 107-117.

Strnad, M., Jović, F., Vorko, A., Kovacic, L., Toth, D., 1997. Young child injury analysis by

the classification entropy method. Accid. Anal. Prev. 30 (5), 689-695.

Stutts, J.C., Wilkins, J.W., Osberg, J.S., Vaughn, B.V., 2003. Driver risk factors for sleep-related crashes. Accid. Anal. Prev. 35 (3), 321-331.

Sullman, M.J.M., Meadows, M.L., Pajo, K.B., 2002. Aberrant driving behaviours amongst New Zealand truck drivers. Transp. Res. Part F 5 (3), 217-232.

Sümer, N., 2003. Personality and behavioral predictors of traffic accidents: Testing a contextual mediated model. Accid. Anal. Prev. 35 (6), 949-964.

Surgeon General of the United States, 1964. Smoking and health. U.S. Government Printing Office.

Taubman - Ben-Ari, O., Mikulincer, M., Gillath, O., 2004. The multidimensional driving style inventory—scale construct and validation. Accid. Anal. Prev. 36 (3), 323-332.

Toledo, G., Shiftan, Y., Hakkert, S., 2007. Framework for analysis and modeling of driving behavior incorporating in-vehicle data recorders. In newsletter of Traffic Safety Center, U.C. Berkeley, 4 (2).

"Traffic Records: A Highway Safety Program Advisory," National Highway Traffic Safety Administration, Washington, D.C., June 1, 2003 [Online]. Available: http://www.nhtsa.dot.gov/people/perform/pdfs/Advisory (Retrieved in 2007)

Tseng, W.S., Nguyen, H., Liebowitz, J., Agresti, W., 2005. Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files. Industrial Management & Data Systems 105 (9), 1185-1205.

Ulmer, R.G., Preusser, D.F., Williams, A.F., Ferguson, S.A., Farmer, C.M., 2000. Effect of Florida graduated licensing program on the crash rate of teenage drivers. Accid. Anal. Prev. 32 (4), 527-532.

Vaa, T., 2001. Cognition and emotion in driver behavior models: Some critical viewpoints. In Proceedings of the 14th ICTCT Workshop, Caserta, Italy.

Vlahogianni, E., Golias, J., Karlaftis, M., 2004. Short-term traffic forecasting: Overview of objectives and methods. Transport Reviews, 24 (5), 533-557.

Vollrath, M., Meilinger, T., Kruger H., 2002. How the presence of passengers influences the risk of a collision with another vehicle. Accid. Anal. Prev. 34 (5), 649-654.

Vorko, A., Jović, F., 2000. Multiple attribute entropy classification of school-age injuries. Accid. Anal. Prev. 32 (3), 445-454.

Walker, N. and Catrambone, R., 1993. Aggregation bias and the use of regression in evaluating models of human performance. Human Factors 35 (3), 397-411.

West, R., Hall, J., 1997. The role of personality and attitudes in traffic accident risk. Appl. Psychol. Int. Rev. 46, 253-264.

Wiesenthal, D.L., Hennessy, D., Gibson, M.P., 2000. The Driving Vengeance Questionnaire (DVQ): the development of a scale to measure deviant drivers attitudes. Violence Victims 15, 115-136.

Wilde, G.J., 2001. Target risk. 2nd Edition, PDE Publication.

Williamson, A., 2003. Why are young drivers over represented in crashes? Reports for Motor Accident Authority, NSW Injury Risk Management Research Centre, University of New South Wales.

Yagil, D., 1998. Instrumental and normative motives for compliance with traffic laws among young and older drivers. Accid. Anal. Prev. 30 (4), 417-424.

Yamada, K., Kuchar, J., 2006. Preliminary study of behavioral and safety effects of driver dependence on a warning system in a driving simulator. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 36 (3), 602-610.