MODELING TRUCK SAFETY CRITICAL EVENTS: EFFICIENT BAYESIAN HIERARCHICAL STATISTICAL AND RELIABILITY MODELS

Miao Cai, M.S.

Draft on May 4, 2020

Dissertation Presented to the Graduate Faculty of
Saint Louis University in Partial Fulfillment
of the Requirements for the Degree of
Public Health Studies, Ph.D.

2020

COMMITTEE IN CHARGE OF CANDIDACY:

Professor Steven E. Rigdon, Ph.D.
Chairperson and Advisor

Professor Hong Xian, Ph.D.

Assistant Professor Fadel Megahed, Ph.D.

DEDICATION

I dedicate this dissertation to my parents, Zhimin Cai (蔡致民) and Guizhen Xu (徐桂珍), who believe in the power of higher education, hard work, and always support me.

ACKNOWLEDGEMENT

I want to thank my mentor and committee chair Dr. Steven E. Rigdon, committee members Dr. Hong Xian and Dr. Fadel Megahed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Transportation safety

Traffic safety is a pressing public health issue that involves huge losses of lives and financial burden across the world. As reported by the World Health Organization (**?**), road injury was the eighth cause of death globally in 2016, killing approximately 1.4 million people, which consisted of about 2.5% of all deaths in the world. If no sustained action is taken, road injuries are predicted to be the seventh leading cause of death across the world by 2030 (**?**). Compared to the victims who were claimed lives by diseases, people killed in traffic are mostly early- or middle-aged, particularly those aged 4 to 44 years old (**??**). Without traffic accidents, these victims could have much longer lives with normal quality of life

Apart from fatal deaths, road traffic injuries were also reported to be the cause of 50 million non-fatal life injuries and approximately 75.5 million disability-adjusted life years globally (**?**). In high-income countries, most of non-death costs were attributable to non-fatal crashes, with 2% of non-fatal events leading to over 40% of life-time medical costs (**?**). Besides non-fatal injuries, traffic safety is a major economic burden. The global economic losses attributable to transportation safety were estimated to be 518 billion United States Dollars (USD), which accounted for 1% the gross domestic product (GDP) in low-income countries, 1.5% in middle-income countries, and 2% in high-income countries (**??**).

Specifically in the United States, transportation contributed to the highest number of fatal occupational injuries, leading to 2,077 deaths and accounting for over 40% of all fatal occupational injuries in 2017 (**?**). The National Safety Council reported that the number of deaths attributable to car crashes was at least 40,000 in 2018, which was the third straight year that this number was over 40,000 (**?**). Despite large amounts of investments in roads, improved vehicle protection and traffic policy implementation, and advanced emergency and trauma care, the reduction in traffic associated fatality rates is nominal (**?**). If the change of traffic fatality rates in the US match those in other unremarkable countries, 20,000 traffic deaths could have been prevented each year (**?**).

1.2 Truck safety

In the US, the large commercial truck industry is the backbone of the economy. Approximately 70% of freight is delivered via a truck at some point of their transportation, which account for 73.1% of value and 71.3% of volume of the domestic goods (**??**). However, large trucks are associated with more catastrophic accidents among all vehicles, so they are the primary concern of traffic safety. In 2016, the Federal Motor Carrier Safety Administration (FMCSA) reported that 27% fatal crashes in work zones involved large trucks (**?**). Among all 4,079 crashes involving large trucks or buses in 2016, 4,564 people (1.12 people per crash) were killed (**?**). Large truck crashes approximately claim 5,000 lives and cause 120,000 injuries each year, but only 15% of these fatalities occur in the trucks, with a predominate 78% occurred in the other vehicles (**?**). Besides, the economic losses associated with large truck crashes are also higher than those with passenger vehicles, with an estimated average cost of 91,000 USD per crash (**?**).

The high risk of large trucks is attributed to two aspects of reasons (**?**). First, large truck drivers generally need to drive alone for long routes, under on-time demands, and challenging weather and traffic conditions. Professional truck drivers usually need to work in shifts, and sometimes unavoidable late-night or early-morning shifts (**?**), which have been reported to

be associated with sleep deprivation and disorders (????). Besides, the long route, constant concentration, and overtime work, intertwine with sleep deprivation and disorder and induce the fatigue symptoms among truck drivers. It is estimated that fatigue among long distance truck drivers caused up to 31% of single vehicle fatal truck crashes (??).

On the other hand, trucks have huge weights, large physical dimensions, and potentially carry hazardous cargoes. Although these huge-size trucks boost the transportation efficiency by increasing cargo capacity and reducing fuel costs per trip, they also raise public safety concerns (?). Large trucks can weight up to 80,000 pounds by federal law, which are twenty times as heavy as a normal-weight passenger vehicle (?). If these trucks travel at the speed of 65 miles per hour on the highway, it will take them around 525 feet to stop, which is about two times the length of a football field (?). The large physical size also creates large blind spots on both sides of the truck, which poses more threat on smaller-sized vehicles. When a crash occurs between a large truck and a smaller vehicle, the sheer size and weight of the truck result in the tragedy that the victims are from the smaller vehicle instead of from the trucks in around 80% of the cases (?). In even worse case, commercial trucks crashes can cause massive casualties and regional public health emergency when the carried hazardous materials are leaked (such as gasoline and sulfuric acid). The importance of truck industry and the potential catastrophic consequences underscore the need to reduce crash risk and improve the safety of truck transportation.

1.3 Modern truck safety studies

To reduce the lives and economic losses associated with trucks, numerous studies attempted to accurately identify the risk factors for truck-related traffic crashes and make accurate prediction. However, there are several limitations of studies using crash data. First, traffic crashes are characterized by rare events (dozens to thousands of times fewer crashes than non-crashes) (??). To tackle this rare-event issue, the most common study design is a case-control study, which matches a crash with one to up to ten non-crashes, and then

use statistical models such as logistic regressions to explain the causes or predict the crashes (????). Unfortunately, a case-control study is limited in estimating incidence rates or overall average treatment effect. It may be contentious in selecting the ratio of controls to cases and how to select these controls (??). Second, due to the retrospective nature of crash data, it is unrealistic to trace back to the real-time traffic, weather, and other environmental factors that were associated with the crashes. Most of crash data reported by police and associated drivers were subject to recall and misinformation bias (?). Third, crashes are underreported, especially for those with no or minor injuries or economic losses (?). The National Highway Traffic Safety Administration estimated that 25% of minor-injury crashes and 50% of no-injury crashes were not reported, compared to nearly 100% reporting rate for fatal crashes (?).

Past truck safety literature almost exclusively focused on crashes, while ignoring precursors to crashes. A precursor to crashes, also known as safety critical events (SCEs), adverse events, or near-miss crashes, is an emerging pattern or signature associated with an increasing chance of crashes (??). Truck SCEs deserve more attention as they occur more frequently than crashes, potentially suggest fatigue, a lapse in performance, and potential catastrophic crashes (?).

1.4 Proposal

With the rapid development of modern technology, more real-time naturalistic driving and SCEs data have been collected by commercial truck companies (?). With advanced unobtrusive instrumentation, these data provide a unique opportunity to continuously study real-world driving performance and potential consequences (?). Naturalistic driving data can be further used to examine the risk factors associated with truck crashes, by fusing high-resolution data on the risk factors, such as driver behavior, vehicle condition, traffic, weather, and road geometry (?).

This prospectus proposal focuses on Bayesian hierarchical statistical and reliability mod-

els based on a large truck naturalistic driving data. The three specific research aims are:

1) quantify the association between truck crashes and SCEs,
2) construct scalable Bayesian hierarchical statistical and reliability models for truck critical events,
3) innovate the non-homogeneous Poisson process (NHPP) with power law process (PLP) intensity function to account for short rests between different trips.

I believe that this work will contribute to statistical theories in constructing scalable Markov chain Monte Carlo (MCMC) to perform Bayesian hierarchical statistical and reliability models. Realistically, these statistical models will provide insights into the relationship between crashes and SCEs, as well as SCEs and driver characteristics, traffic, weather, and other real-time driving environmental variables. These statistical models can be further used to provide data-driven evidence to optimize trucking routes, minimize unsafe driving behavior, and improve a safe driving environment.

CHAPTER 2

LITERATURE REVIEW

2.1 Naturalistic Driving Studies (NDS)

Traditional truck crash prediction studies almost exclusively use data that ultimately trace back to post hoc vehicle inspection, interviews with survived drivers and witnesses, and police reports (**??**). Despite these data can be in-depth and thorough, they have several inherent limitations. Firstly, truck crashes are extremely rare compared with non-crashes. According to the **?**, large truck and bus fatalities in 2017 were 0.156 per million traveled vehicle miles, which was a 6.8 percent increase from 2016. This rareness poses a challenge to infer unbiased estimates using traditional statistical models (**??**). Secondly, truck crash data almost exclusively rely on post hoc police reports. Although these data are generally accurate and detailed, they are limited in determining the information of the driver in a meaningful time period leading up to the crash (**?**). Some of the critical factors, such as distraction, are not reported or cannot be determined due to a variety of reasons (**?**), and these data were subject to recall bias even if they were reported. Thirdly, truck crashes are under-reported, particularly for no-injury and minor-injury crashes (**??**). It is estimated that 25% of minor-injury and 50% of non-injury crashes were not reported, while 100% of fatal crashes were reported (**?**).

Considering these limitations, a growing number of naturalistic driving studies have been

initiated worldwide to identify crash causation and improve traffic safety (**???**). NDS use unobtrusive devices, sensors, and cameras installed on vehicles to proactively collect frequent naturalistic driving behavior and performance data under real-world driving conditions (**??**). Compared with traditional post-hoc crash data that are road segment-based, NDS collect driver-based data that are more useful in comparing the rate of SCEs under different circumstances. In addition, NDS data provide high-resolution driver behavior and performance data, which enable researcher to access data shortly prior to the occurrence of crashes or SCE without information bias or selection bias (**?**). Third, collecting naturalistic data is considerably less costly and difficult per observation compared to traditional crash data that involve human resource, interviews, and witnesses, so NDS generally collect a large amount of data, which creates both an opportunity and a challenge to researchers. **?** provides an excellent review that compares empirical, naturalistic, and epidemiological data collection methods in traffic safety research.

The first large-scale NDS was the 100-Car Naturalistic Driving study conducted by the Virginia Tech Transportation Institute in the US (**??**). Other well-known NDS projects include the second Strategic Highway Research Program (**?**) and the UDRIVE NDS in Europe (**??**). There are also a few other NDS that target at specific populations, such as the 40-Teen NDS (**?**), the Older Driver Fitness-to-Drive NDS (**?**), and the Commercial Truck Driver NDS (**?**).

2.2 Safety-Critical Events (SCEs)

Instead of collecting extremely rare vehicle crash data, NDS focus on safety-critical events (SCEs) and near-crash events, defined as events that used last-second successful evasive maneuver that avoided crashes (**?**). Although near-crashes or SCEs were not real crashes, a few studies suggested that they were correlated with crashes among cars (**???**). The most commonly studied SCE is hard brakes (also knowns as hard-braking events or harsh braking), defined as a deceleration force higher than a pre-specified threshold, such as 0.3 g (**??**).

A more formal definition of near-crash, or the more general accident precursor, in safety analysis and accident prevention field was proposed and analyzed in **?**. An accident precursor was defined as a chain of adverse events following an initial off-nominal event, which can result in an accident if compounded with additional adverse conditions (**?**). A near-crash or near miss is a special case of accident precursor, with the feature of being close to a complete accident sequence. Accidence precursor has been widely studied in certain safety science in which the accidents are extremely rare, such as nuclear industry (**?**), chemistry (**?**), and aerospace industry (**?**).

The rationale for using near-crashes and SCEs as surrogates for crashes is Heinrich's Triangle. The Heinrich's Triangle assumes that less severe events are more frequent than severe events, and the frequency of severe events can diminish as that of less severe events decreases (**?**). The latter assumption can be quantitatively tested using crash and naturalistic driving data, but verifying the former assumption is challenging since the causal mechanism is complex and unknown (**?**). Applying SCEs in traffic safety studies to this Heinrich's Triangle can substantially increase the study sample size and may potentially enable the estimation of driving risk. However, a crucial question prior to the usage of SCEs in naturalistic studies is whether they are good surrogates of traffic crashes.

? proposed two critical principles for using near crashes as surrogates for crashes: 1) similar or the same causal mechanisms between crashes and surrogates, 2) a strong association between the frequency of surrogates and crashes. Based on the 100-car database, they investigated the two principles using a sequential factor analysis, a Poisson regression, and a sensitivity analysis. The study concluded that using near crashes as surrogates for crashes will lead to conservative risk estimates but significantly reduce the variance of estimation. They suggested that using near crashes as surrogates in small-scale studies will be informative for evaluating the risk of crashes.

2.3 Crashes and SCEs

The 100-Car NDS continuously followed 102 recruited drivers for 12 months, resulting in two million miles and over 40,000 hours of driving data. To maximize the number of SCEs, the research team intentionally chose more young drivers and high mileage drivers. Based on this data, **?** found that hard braking events were significantly associated with collisions and near-crashes. Since the number of near-crashes and incidents were significantly larger than crashes, they proposed to use near-crashes and incidents as surrogates of crashes.

? conducted a preliminary study to validate surrogates for road-departure crashes by spatially merging road geometry, average traffic, crashes, and naturalistic driving data. Bayesian seemingly unrelated Poisson models estimated with weighted least squares were used to examine if the same sets of predictor variables can have the same effects on crashes and surrogates respectively. They found that time to edge crossing and lane-departure warning were two useful surrogates for crashes on rural non-freeway roads, while lane deviation was a poor surrogate for lane-departure crashes.

? proposed a conceptual framework to estimate the crash-to-surrogate ratio π using the 100-Car study. The study found that the conditional probability of a crash was increased by 24 times with a lateral acceleration more than 0.7 g, but the probability was decreased by other factors such as the event occurring in daylight and dry pavement. A later study by **?** developed diagnostic procedures to screen crashes and near-misses under the NDS settings. The study applied their proposed framework to the 100-Car NDS and identified three conditions to define surrogate events: 1) maximum lateral acceleration difference of no smaller than 0.4 g, 2) non-intersection related, and 3) maximum lateral acceleration difference no smaller than 0.9 per event or between 0.8 and 0.9 g during night time.

? used linear referencing to link Global Positioning System (GPS) data with roadway features on 39 segments of Highway 101 in California. A negative binomial model and a random-effects negative binomial model that account for segment-specific variance were used

to investigate the relationship between historic crashes and hard braking. It was found that the freeway segments with high hard braking rates also had higher long-term crash rates, although the other three explanatory variables (average daily traffic, the presence of horizontal curvature, and auxiliary lanes) were not statistically significant.

? used in-vehicle data recorders (IVDR) data collected on 3,500 segments of inter-urban roads in Israel to examine the association between two types of safety-related events (braking and speed alert) and crashes on different road types. Negative binomial models were applied to account for the over-dispersion in the data, and they also included several road infrastructure characteristics as covariates. The number of braking events was found to be positively associated with injury crashes on single- and dual-carriageway roads while the association was not significant on freeways. However, they yield counterintuitive results that speed alert events (overspeed) were consistently and negatively associated with injury crashes on all road types. It was suggested that a speed alert event was not a good surrogate for crashes, possibly due to its rough definition.

Some researchers have been skeptical of NDS. **?** challenged the validity of using naturalistic driving data and SCEs by stating that the purpose of traffic safety studies is to identify causes of crash harm, which is defined as property damage, injury, income lost, and all other consequences of different severities (**?**). In contrast, NDS often use SCEs as surrogates of crashes, but very few or no crashes, let alone human harm. Therefore, **?** argues that SCEs are not an appropriate part of the Heinrich's Triangle, so researchers generally cannot derive valid quantitative conclusions on causations of harm based on NDS datasets.

Another study by **?** specifically targeted Hour-of-Service rule research and relevant policy revisions among commercial truck drivers. He argued that HOS studies with a quasi-experiment design were subject to confounding variables, so these studies are limited in demonstrating a causal relationship between HOS and safety outcomes. The paper also argued that NDS lacked external validity since no large truck NDS had examined the causal link between crashed and SCEs. Lastly, the construct validity was doubted since the relationship

between driver fatigue, HOS, and SCEs had not been validated.

2.4 Risk factors for traffic safety

2.4.1 Fatigue

Fatigue has been the most pressing risk factor for truck crashes and SCEs. It is estimated that approximately 32% of drivers drive with fatigue over twice a month (**?**). The American Automobile Association Foundation for Traffic Safety claimed that 16.5% of fatal traffic accidents and 12.5% of injuries-related collisions were associated with driving with fatigue in 2010 (**?**). The National Highway Traffic Safety Administration (NHTSA) estimated that 60% of fatal truck crashes were attributable to the driver falling asleep while driving (**??**). The FMCSA estimated that the causal role of fatigue is around five times higher in fatal than in property damage truck crashes (**?**).

Fatigue is often defined as a multidimensional process that leads to diminished worker performance, which may be a result of prolonged work, psychological, socioeconomic, and environment factors (**??**). However, this definition has low specificity since there are other factors associated with a decreased worker performance, such as cell phone use, which does not result in driver fatigue. There is no uniform and succinct definition on fatigue since it involves interactions between biological, behavior, and psychological process. A comprehensive review on fatigue definition and measurement is provided by **?**.

The mechanism of fatigue leading to SCEs is that the driver's capability to stay alert to ambient traffic and pedestrians will be largely impaired, and the driver's reaction time is subsequently prolonged in that situation (**?**). It is estimated that 17 hours of continuous working lead to a deterioration of driving performance equivalent to a blood alcohol level of 0.05% (**?**). What makes the outcomes worse is that fatigue driving is more likely to happen on expressways and major highways where the speed limit is over 55 miles per hour (**?**). This is especially concerning because SCEs on these segments are more likely to result in serious

injuries and fatalities, compared with non-fatigue driving safety critical events.

Although fatigue has been recognized as the primary reason for traffic safety, preventing drowsy driving has not been effective since there is no simple way to objectively measure fatigue driving (**?**). In view of the difficulty of measuring fatigue, researchers attempted to use different proxies of fatigue associated with truck drivers, such as cumulative driving time, ocular and physiological metrics, sleep patterns, and night driving, but none of them has shown significant superiority.

Cumulative driving time has also been a measure of driver's fatigue level, especially among NDSs. For example, **?** found that there was a significant increase in the fatigue of drivers after 12 hours of continuous driving. **?** used cumulative hours of driving, time of the day, driving patterns over multiple days, rests after driving, and the 34-hour recovery policy as measures of driver fatigue. They found that more driving time was associated with increased odds of crashes among 686 less-than-truck-load drivers, with the highest odds in the 11-th hour. From the fifth hour to the 11th hour, the odds of crashes were consistently increasing (**?**). In contrast, **?** found no significant difference in safety outcomes between the 11-th driving hour and 8-, 9-, or 10-th driving hours using the Naturalistic Truck Driving Study data, but they suggested a working day that starts with several hours of non-driving work and then followed by 14 hours of driving was significantly associated with risk of SCEs. Another study by **?** used a Poisson regression to quantify the association between driver performance and predicted fatigue level based on naturalistic driving data from 106 commercial truck drivers. The number of hard-braking events, defined as deceleration force greater than 0.3 g, were considered the outcome variable in that study. The fatigue level was predicted using a complex biomathematical model provided by **?**. After accounting for time of the day, they reported a significant association between predicted fatigue and the rate of hard-braking events (relative risk 1.078, 95% CI: 1.013-1.146).

Lack of sleep or specific sleep patterns have also been used as proxies of fatigue. For example, **?** used negative binomial regression to identify the association between four sleep

patterns and driving performance based on the Naturalistic Truck Driving Study data. They revealed that shorter sleep, early-stage sleep in a non-work period, and insufficient sleep between 1 am and 5 am were associated with increased safety-critical event rates. **?** reported that truck drivers with a restart break of only one nighttime period (defined as 1 am to 5 am) experienced more lapses of attention, elevated lane deviation at night, and higher sleepiness measured by subjective questionnaires. A naturalistic driving study by **?** found that more sleeping hours was found to be reducing near crashes.

A significant amount of research emphasizes the association between time of the day (such as night driving) and fatigue development (**?**). Night driving is often accompanied by changes in shift scheduling, inadequate sleep, sleep apnea and disorder. For example, **?** reported that the crashes caused by drivers falling asleep occurred primarily from mid-night to 7 am and from 2 pm to 4 pm. **?** monitored 80 truck drivers in North America using 24-hour electrophysiological measures. They found that drivers had an average of 5.18 hours of sleep in bed and 4.78 hours of electrophysiologically validated sleep per day, which were significantly less than needed to stay alert on job. It was also suggested that late-night or early-morning work were detrimental to the drivers' sleep. **?** investigated the risk factors associated with crashes on Texas urban freeways between 2006 and 2010. They ran separate random-effects logistic regressions on five time periods: early morning (12 am to 4 am), morning (5 am to 9 am), mid-day (10 am to 3 pm), afternoon (4 pm to 8 pm), and evening (9 pm to 11 pm). The results suggested that different time periods contributed differently to the severity of injuries, which highlighted the importance of time of day.

2.4.2 Driver characteristics

Young and older drivers have been reported to have higher risk of crashes or SCEs. The reasons for young drivers having higher risk of driving are not fully explained, but could largely be attributed to inexperience and reckless driving. In contrast, older drivers may find it difficult to adjust for the sleep-wake cycle to keep pace with intense schedule required

by their employer company, which may increase the likelihood to be sleepy or fatigued. **?** reviewed published literature on age-related safety issues among professional heavy vehicle drivers. The review suggested a U-shaped relationship: the chance of driving safety issues declines before 27 years old, plateau until the age of 63, and starts to grow up again after 63.

Young drivers are much better in the sense of physical health and resistance to fatigue compared with aged drivers. However, they are generally inexperience in driving compared with aged drivers. **?** suggested that young drivers (17–19 years old), especially males, have significantly more accidents than other drivers during the hours of darkness, on rural curves, and rear-end shunts compared to male drivers aged 20 -25 years. **?** found that truck drivers under the age of 19 were over-involved in fatal accidents by a factor of 4, and those aged between 19 and 20 were over-involved by a factor of 6. **?** revealed that the drivers under the age of 25 accounted for 55% of the 4,333 crashes in which the drivers were judged to be asleep whiling driving. **?** tested the sleepiness of 36 profession drivers in simulated driving sessions using electroencephalogram, the Karolinska sleepiness scale, and visual analog scales. They found that young driver experienced a significant decrease in alertness and a strong tendency to sleep compared to middle-aged drivers. Based on the 100-Car Naturalistic Driving Study dataset, **?** found that drivers under the age of 25 were more likely to have crashes and SCEs.

To meet the huge demand services and supply chain management, it is common to extend retirement age or reemploy retired workers, especially in developing countries (**?**). Aged drivers have an increased chance of driving safety issues for three reasons: impaired eyesight, prolonged reaction time to exogenous stimuli, and vulnerability to fatigue (**?**). Aged drivers often have eyesight diseases or functionality impairment, such as cataracts, narrowed peripheral vision and decreasing visual acuity (**?**). In addition, working for truck companies often means irregular shifts and taking night schedules, which disrupt the circadian time-keeping systems, especially for aged workers (**?**). It is indicated that the "critical age" of shiftwork intolerance is about 45 to 50 years, at which sleep disorder, persisting fatigue and digestive problems become the most prominent (**?**).

2.4.3 Traffic

Traffic characteristics are also viewed as an important risk factor for traffic safety issues. For the sake of availability and low cost, most prior studies used aggregated traffic data as proxies of traffic, such as Annual Average Daily Traffic (AADT). More recently, an increasing number of studies start to use real-time traffic data as a high-resolution proxy of traffic characteristics. Three published papers reviewed the impact of traffic variables on traffic safety issues (**???**).

Traffic variables include flow (traffic volume), occupancy/density, and speed (**??**). Traffic flow is defined as the number of vehicles passing through a specific road segment in a given unit time. Traffic occupancy or density is defined as the number of vehicles in a unit area of road at a moment. Speed can be computed from the road perspective as the mean speed of vehicles passing that road segment (such as the AADT), or from the vehicle perspective as the speed of the vehicle (such as real-time speed). Compared to traffic flow and speed, traffic density is relatively less investigated due to a lack of relevant data.

For example, based on multinomial logit and negative binomial models, **?** used vehicle and driver characteristics, traffic, environment, and road geometry to predict the frequency and severity of large truck-involved crashes. They found that the percent of large trucks, AADT, driver condition, and weather characteristics were significantly associated with both crash frequency and severity. **?** used hourly aggregated traffic variables, including flow, occupancy, mean time speed, and percentage of trucks to predict crash occurrence with a bias-correction logistic regression. This study found that the main risk factor, average speed had a negative effect on crashes. Instead of studying risk factors of crashes, **?** examined the association between Hard Braking Incidents (HBIs) and geometric and traffic variables among large trucks at roundabouts. They found that HBIs were influenced by traffic and geometric variables in a similar fashion as crashes.

2.4.4 Weather

Weather variables, including precipitation, visibility, wind speed, and temperature, have reported to have both direct and indirect effects on traffic safety events. **?** provides a review on weather characteristics and road safety.

Real-time extreme weather conditions such as heavy rain, fog, storm, and snow can either impair the driver's visual capability or reduce the safety of driving on the road (**???**). A positive linear relationship between precipitation and traffic accidents can be observed in both driver accidents and pedestrian accidents (**??**). **?** used ordinal and multinomial regression models with random-effects to investigate crash severity under various weather conditions. They found that wind speed, rain, humidity, and temperature were associated with single-vehicle truck crashes. **?** used detector and sensor data to successfully predict more than 70% of accidents with low visibility conditions.

In addition, the increase of ambient temperature places risks on occupational safety, and possibly leads to cognition loss, heat stroke, and impairment of wakefulness. Previous evidence showed that the risk of mistakes and SCEs are elevated in hot weather (**??**). **?** found that for a day with temperature above 80 °F, there is a 9.5% increase in fatality rates compared with a day at 50-60 °F. A review by **?** found that 11 out of the 13 included studies indicated an increase in unintentional injuries associated with high temperatures. In contrast, when low temperature is present, truck drivers are likely to be faced with snowy or icy road conditions, as well as the presence of fog, which substantially increase the risk of driving (**?**). For example, **?** reported that truck-involved crashes were 19% more likely to occur than no truck-involved crashes when snow or strong wind were present.

2.4.5 Road characteristics

Based on engineering theory, road characteristics are potential risk factors of road safety (**?**). Commonly used road characteristics in traffic safety studies include the number of lanes,

lane with, speed limits, horizontal curves, road curvature, and lighting conditions.

The effect of the number of lanes and lane with on traffic safety is inconsistent in previous literature. Some studies suggest that the number of lanes is negatively associated with the risk of traffic accidents. For example, **?** found that crashes on roadways with more lanes tended to be less severe, which may result from the fact that more lanes give more space and separation between vehicles. They also reported that crashes in higher speed limit segments were more likely to be severe crashes. In contrast, several other studies suggested that an increase in the number of lanes and lane width were positively associated with traffic fatalities (**????**). This reversed relationship could possibly be explained by an increased chance of lane-change-related conflict opportunities (**?**).

It is generally believed that lower speed limits can reduce the chances of traffic crashes, as well as the severity of crashes. For example, **?** used state-level data from 1991 to 2005 to examine the association between truck-specific restrictions and fatality rates. They found that higher speed limits were associated with increased fatality rates, although different speed limits across vehicle types had no significant effect. Another study by **?** also provided evidence that both overall and truck-involved fatalities were positively associated with maximum speed limits.

Road geometric design features, such as road curvature and terrain type, were also reported to be risk factor of traffic safety events. **?** used zero-inflated negative binomial models to examine the effects of road geometry features and crash frequency. Based on 1,787 truck-involved crashes from 1,310 highway segments in four years, they found that AADT, segment length, degree of horizontal curvature, terrain type, land use and width, median type, right side shoulder width, lighting condition, rutting depth, and posted speed limits were significantly associated with the likelihood of truck-involved crash frequency. **?** found that crashes on roadway curved were associated with higher likelihood of major and possible injuries in urban single-vehicle large truck at-fault accidents, but this association is not statistically significant in multi-vehicle accidents.

2.5 Predictive models

2.5.1 Overview

There are two cultures in current statistical or data science field: explanation and prediction (**??**). The pro-explanation culture has long been adopted by most disciplines, such as epidemiology, economics, and psychology. In these disciplines, researchers commonly use generalized linear models, such as logistic regression and Poisson regression, to explain the association between the outcome and predictor variables. In contrast, the pro-prediction culture has recently been adopted in data science disciplines, in which they use black box algorithms such as random forests, decision trees, and neural networks to achieve similarly high prediction accuracy in training and testing sets. Pro-explanation models tend to excel at explaining the association between predictors and the outcome variable and being less likely to overfit the data. However, compared with machine learning and deep learning algorithms, pro-explanation models are less likely to capture potential interaction between predictor variables since they are driven by conceptual frameworks. Therefore, pro-explanation models generally have less prediction accuracy compared with black box algorithms.

Traffic safety field has a pro-explanation culture, although it is shifting towards a pro-prediction by adopting cutting-edge machine learning and deep learning algorithms. The most commonly used statistical models in this field are logistic regression and Poisson regression. Logistic regression is commonly applied to predict crash likelihood (probability) using predictors such as driver features, weather, and traffic (**?**). In contrast, Poisson regression is applied to predict the crash frequency (the number of crashes) within a time period using similarly aggregated traffic and weather characteristics. The following paragraphs will briefly introduce the two models and then compare the two cultures of predictive models in statistical and machine learning perspective.

The parameterization of a binary logistic regression is shown in Model (2.1).

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p_i) \\ \log(\frac{p_i}{1-p_i}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \end{aligned} \tag{2.1}$$

Where Y_i is a binary variable that indicates whether an event occurred or not for the i-th observation. p_i is the mean parameter of a Bernoulli distribution, which is constrained on $[0, 1]$. The logit transformation of p_i then has the range from $-\infty$ to $+\infty$, which equals a linear combination of the predictors $x_1, x_2, \cdots, x_k$ and associated parameters $\beta_0, \beta_1, \cdots, \beta_k$.

The most commonly used outcomes for binary logistic regressions are injury versus non-injury crashes or fatal versus non-fatal crashes (**?**). For example, **?** used a two-level hierarchical Bayesian logistic model to predict the likelihood of high-severity crashes compared to low-severity crashes in New Mexico, accounting for both crash- and driver-level effects. They found that road curve, functional and disabled vehicle damage, single-vehicle crashes, female, older drivers, drug or alcohol involvement were associated with increased odds of severe crashes. **?** used logistic regression with rare events bias correction and Firth method to study significant risk factors for crashes in Greece. They found a negative association between crash likelihood and speed in crash locations. The proportion of trucks on the road was included in their model but not found to be significant. Other traffic safety studies using logistic regressions include but were not limit to **?**, **?**, **?**. There are two excellent systematic reviews on traffic crash likelihood predictions by **?** and **?**.

Other variants of binary logistic regression are binary probit models (**??**), ordered logistic or probit models (**??**), multinomial logit models (**?**). There are only minor differences between a probit model and a logistic model. A logistic model uses the inverse logit of the linear predictors to calculate the probability of an event, as shown in Equation (2.2); a probit model uses the cumulative normal density function of the linear predictors to calculate the probability, as shown in Equation (2.3). The error function $\text{erf}(x)$ is an integral without an

analytical solution: $\text{erf}(x) = \frac{2}{\pi}\int_0^x e^{-t^2}dt$.

$$p = \text{logit}^{-1}(\mathbf{X}'\beta) = \frac{\exp(\mathbf{X}'\beta)}{1+\exp(\mathbf{X}'\beta)} \tag{2.2}$$

$$p = \Phi(\mathbf{X}'\beta) = \frac{1}{2}\Big[1+\text{erf}(\frac{\mathbf{X}'\beta}{\sqrt{2}})\Big] \tag{2.3}$$

Ordered logistic or probit regressions aim to model an ordered multi-category outcome variable. The most common case is study the severity of crashes, such as no-injury crashes, minor-injury crashes, and fatal-injury crashes (**?**). These ordered models account for the ranked nature of different severity levels but make the proportional odds assumption (**?**). When the proportional odds assumption is violated, researchers often switch to multinomial logit or probit models, in which the outcome variable is deemed as nominal.

When researchers have crash data that are aggregated over a long time-period such as one year, it often makes sense to study the number of crashes instead of whether a crash occurred or not since they are often more than one crash. The most commonly used statistical model is therefore Poisson model, as it handles count data that are right-skewed, long tailed, and only have non-negative integer values. The parameterization of a Poisson regression is shown in model(2.4).

$$\begin{aligned} Y_i^\star &\sim \text{Poisson}(\mu_i) \\ \log\mu_i &= \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k, \end{aligned} \tag{2.4}$$

where $Y_i^\star$ is the number of events for the i-th observation, which must be a non-negative integer. μ_i is both the mean and the variance parameter of the Poisson distribution, and it must be a non-negative numeric value. The logarithm of μ_i transforms μ_i into the range of $(-\infty,+\infty)$, which equals a linear combination of the predictors $x_1, x_2, \cdots, x_k$ and associated parameters $\beta_0, \beta_1, \cdots, \beta_k$. Note that the mean parameter equals the variance parameter in a Poisson distribution, which is often violated in real-life data. When the variance of the data

is greater than expected, it is called over-dispersion. Otherwise, it is called under-dispersion. Over-dispersion is much more common than under-dispersion in practice.

Here are some applications of Poisson regression in traffic safety research. **?** used Poisson regressions to explore the association between the rate of crashes driver-level characteristics among 560,695 commercial truck drivers in the United States. They found that past safety performance, out-of-service rate, body mass index (BMI), age, and the number of unique companies were strong predictors of the rate of truck crashes. Other variants of a Poisson model include negative binomial models, quasi-Poisson models, and zero-inflated Poisson or negative binomial models (**??**). Negative binomial or quasi-Poisson models are developed to account for the over-dispersion and under-dispersion in count data, for which a Poisson model fails to account. Zero-inflated Poisson or negative binomial models are developed to account for the rare-event nature of traffic crash data (**????**). There is an excellent review paper on statistical models for crash frequency data by **?**.

Recently, recurrent event models have also been increasing applied to model the change in intensity of SCEs in the traffic safety field. For example, **?** proposed to use a mixed-effects Poisson process to model unintentional lane deviation events, with the baseline intensity and time-varying coefficients modeled by penalized B-splines. They first conducted a simulation study to assess the performance of the proposed model with different curvature of time-varying coefficients and the magnitude of event rate. Simulated 500 data sets with 500 shifts per set suggested satisfactory estimates for the true Gamma fragility parameter ϕ as estimated by an expectation–maximization algorithm, where larger values of ϕ indicated greater heterogeneity between shifts and more intense events. The bias ϕ in the simulation ranged from -0.01 to -0.09, which was around 2% smaller and 0.6% smaller than the true value in low and high event rate settings respectively. They applied the proposed model to 96 commercial truck drivers including 1,880 shifts. The study found that shifts with normal sleep time (7-9 hours) had a lower intensity compared with insufficient ($<$ 7 hours) and abundant ($\geq$ 9 hours) sleep time shifts.

2.5.2 Bayesian models

Traditional frequentist models that view parameters as unknown but fixed values. In the Bayesian perspective, parameters are viewed as random variables that have probability distributions (**?**). Researchers should have subjective prior beliefs (a probability distribution) on these parameters $p(\theta)$ before they collect any data. After observing the data $\mathbf{X}$, the researchers could change their prior beliefs. Therefore, the posterior distribution $p(\theta|\mathbf{X})$ is an unconditional distribution that is a balance between the prior beliefs and the data. This balance is given analytically by the Bayes Theorem (Equation (2.5)).

$$\begin{aligned} p(\theta|\mathbf{X}) &= \frac{p(\theta)p(\mathbf{X}|\theta)}{p(\mathbf{X})} \\ &= \frac{p(\theta)p(\mathbf{X}|\theta)}{\int p(\theta)p(\mathbf{X}|\theta)d\theta} \end{aligned} \tag{2.5}$$

The $p(\mathbf{X}|\theta)$ is the likelihood function, which reflects the data generating process that gives rise to the observed data. The denominator $\int p(\theta)p(\mathbf{X}|\theta)d\theta$ is a normalizing constant that force the posterior distribution to be integrated to one. The prior and likelihood function are straightforward since they both have analytical forms, while the trickiest part of Bayesian inference is the normalizing constant in the denominator (**??**).

The normalizing constant need to make the posterior distribution be integrated to one since the posterior is a probability density distribution. When there are more than two parameters in the model, the normalizing constant often becomes analytically intractable since it involves integration in multiple dimensions. Therefore, modern Bayesian inference often uses numerical methods such as Markov chain Monte Carlo (MCMC) to directly sample from this posterior distribution. However, MCMC often fail or take an inhibitively long time to solve the problem in the case of high-dimensional data or tall data.

There are several strengths of Bayesian models over traditional Frequentist models. First, the probabilistic distribution of parameters, posterior credible intervals, and posterior predictive distributions account for the uncertainty in parameters and the data generating

process, and they also have straightforward and intuitive interpretations. Second, Bayesian models incorporate prior information $p(\theta)$ into the statistical model, which can be useful when there is sufficient prior background information. This prior distribution is particularly useful for estimation in high-dimensional, sparse data, and complex model settings as these regularizing priors can solve convergence issues in traditional maximum likelihood estimation (MLE) (**??**). Lastly, powered by numerical methods and simulation, Bayesian models are applicable in complex data generating process. In practice, researchers only need to specify the priors and likelihood function, and MCMC will sample from the posterior parameter space. Compared to traditional Frequentist approaches such as the MLE that is complicated in estimation for complex models, the difficulty of writing the likelihood function is minimal in Bayesian estimation (**?**).

2.5.3 Hierarchical models

Most studies on traffic safety assume that the sampling unit is a spatial-temporal segment, which is a specific section of a road with relatively high rate of crashes during a period. However, it is not sufficient to only examine the occasions where the crashes are more likely to occur; we must also study the non-crashes and compare them with crashes. On the other hand, these studies that focus on road segments ignore driver-level unobserved effects, which is not ignorable in traffic safety studies. It is reported that the chance of having crashes for truck drivers with crash history in the past year is nearly twice as high as those without a crash history in the past year (**?**). Most motor carrier insurance companies and employers also view historical safety events as an important measure of the driver's performance. Therefore, it is more natural to use driver-focused models to account for unobserved variation and characteristics (**?**).

In Bayesian perspective, a hierarchical model is a statistical model with the probability distribution of one parameter depends on another parameter (**?**). Suppose we have a model with two parameters α, β and data D. The joint prior distribution of the two parameters is

$p(\alpha, \beta)$. According to the Bayes Theorem, the posterior distribution is proportional to the product of the prior distribution and the likelihood function: $P(\alpha, \beta|D) \propto P(\alpha, \beta)P(D|\alpha, \beta)$. In a hierarchical model setting, the product can be factored as a chain of products among parameters, also known as conditional independence, such as $P(\alpha, \beta)P(D|\alpha, \beta) = P(D|\beta)P(\beta|\alpha)P(\alpha)$. In this parameterization, the parameter α is known as the hyperparameter because it gives rise to the parameter β (the parameter of a parameter) (**?**).

Compared with traditional fixed-effects models that either pool all groups of data or estimate separate models individually for each group, a hierarchical model has the advantage of partial pooling across different groups (**?**). This partial pooling shrinks group-level parameter estimates towards the group mean and shares information across groups. Therefore, with reasonable assumptions on the data generating process, estimates from a hierarchical model are generally more robust to extreme observations and reasonably accurate for those groups with sparse data (**??**).

However, hierarchical models are particularly known for its complexity in estimation for both Frequentist MLE and Bayesian estimation. The de facto way of current Bayesian estimation is Markov chain Monte Carlo (MCMC). However, in the hierarchical model setting, it is difficult for MCMC to efficiently sample from the posterior distributions of hyperparameters due to the correlation between different levels of parameters, as well as the large number of parameters created by the hierarchical structure.

2.5.4 Markov chain Monte Carlo (MCMC)

In modern statistics, Bayesian inference almost indispensably relies on Markov chain Monte Carlo (MCMC) to overcome the intractable denominator issue in Bayes Theorem (Equation (2.5)). A *Monte Carlo simulation* is a technique to understand a target distribution by generating a large amount of random values from that distribution (**?**). A *Markov chain* has the property that the probability distribution of the observation i only depends on the previous observation $i - 1$, not on any one prior to observation $i - 1$, as demonstrated in

Equation (2.6).

$$P\left(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}\right) = P\left(X_i = x_i | X_{i-1} = x_{i-1}\right) \tag{2.6}$$

Integrating Markov chains and Monte Carlo simulations, the MCMC method can characterize an unknown unconditional distribution by sampling from the distribution, without knowing its all mathematical properties (**?**). It has been widely applied in fields such as statistics, physics, chemistry, and computer science (**?**). The most notable application of MCMC is probably in Bayesian inference, in which it has been used to draw samples from the posterior distribution and calculate relevant statistics (such as posterior mean and intervals).

The first proposal of using MCMC dates to the paper by **?**, in which they tried to solve an intractable integral with a random walk MCMC. The Metropolis algorithm starts with a randomly defined initial value of the parameter θ. From a pre-defined symmetric proposal probability distribution $p(\theta|\mathbf{x})$, it then draw a proposal parameter value $\theta^{(\text{prop})}$, which only depends on the current parameter value $\theta^{(t)}$. This proposal value will be accepted with the probability of α defined in Equation (2.7). This proposal and acceptance with probability steps will be iterated for a pre-define number of times. When the Metropolis algorithm reaches a steady state, these proposal values are random values drawn from the posterior distribution of parameter θ, which can be used to describe and characterize the posterior distribution.

$$\alpha = \min\left(1, \frac{p(\theta^{(\text{prop})}|\mathbf{x})}{p(\theta^{(t)}|\mathbf{x})}\right) \tag{2.7}$$

After decades of successful empirical trials in physics, **?** proposed a more generalized form of the Metropolis algorithm, in which the proposal distribution can be arbitrary, but the acceptance probability $\alpha^\star$ is modified as shown in Equation (2.8). This Metropolis-Hasting (M-H) algorithm is the most widely-known MCMC algorithm used in different fields. Let $p(\theta|\mathbf{X})$ be the posterior distribution we want to know, then the *Metropolis-Hasting algorithm* is:

1. Let $\theta^{(1)}$ denote an initial value for the continuous state Markov chain,
2. Set $t = 1$,
3. Let q be the proposal density which can depend on the current state $\theta^{(t)}$. Simulate one observation $\theta^{(\text{prop})}$ from $q(\theta^{(\text{prop})}|\theta^{(t)})$,
4. Compute the following probability:

$$\alpha^{\star} = \min\left(1, \frac{p\left(\theta^{(\text{prop})}|\boldsymbol{x}\right)}{p\left(\theta^{(t)}|\boldsymbol{x}\right)} \frac{q\left(\theta^{(t)}|\theta^{(\text{prop})}\right)}{q\left(\theta^{(\text{prop})}|\theta^{(t)}\right)}\right) \tag{2.8}$$

5. Set $\theta^{(t+1)} = \theta^{(\text{prop})}$ with the probability of $\alpha^{\star}$; otherwise set $\theta^{(t+1)} = \theta^{(t)}$. Set $t \leftarrow t + 1$ and return to 3 until the desired number of iterations is reached.

Although the M-H algorithm is simple and powerful for performing MCMC, its performance highly depends on the statistical structure and the proposal distribution. When there are a few parameters in the model and the proposal distribution is not well-designed, the M-H algorithm will have a very low acceptance rate, which makes the M-H algorithm extremely inefficient. In view of this issue, Gibbs sampler was proposed with the idea that the proposed values are always accepted and each parameter is updated one at a time by generating samples from the conditional distributions (**???**). The development of the software *Bayesian inference Using Gibbs Sampler (BUGS)* was critical in popularizing applied Bayesian analyses since it supports a variety of statistical distributions, automatic application of the Gibbs Sampler, and numerous textbooks, tutorials and papers (**??**). Suppose $\theta = [\theta_1, \theta_2, \cdots, \theta_k]$ is a k-dimensional parameter. Let $\mathbf{X}$ denote the data. The *Gibbs sampling* algorithm is then:

1. Begin with an estimate $\theta^{(0)} = [\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_k^{(0)}]$ in the parameter space,
2. Set $t = 1$,
3. Simulate $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_k^{(t-1)}, \mathbf{X})$,
4. Simulate $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_k^{(t-1)}, \mathbf{X})$,
5. $\cdots$,
6. Simulate $\theta_k^{(t)}$ from $p(\theta_k|\theta_1^{(t)}, \theta_3^{(t)}, \cdots, \theta_{k-1}^{(t)}, \mathbf{X})$,

7. Set $t \leftarrow t + 1$ and repeat steps $3 - 6$ for a pre-specified number of iterations and make sure the Gibbs sampler reaches the steady state after sufficient iterations.

The M-H algorithm and Gibbs sampler gain popularity among applied researchers with the development of open source software such as `R` and `BUGS` in the recent 30 years. However, as more and more data are available in applied field, the performance of the two most popular MCMC algorithms has been widely criticized, which drives the development of more efficient MCMC algorithms and software packages (**?**).

2.6 Scalable Bayesian models

An explosive growth of data size and dimensionality in recent years poses a major challenge to Bayesian estimation using MCMC. Traditional MCMC algorithm need to evaluate the entire data at each step of iteration, which could be expensive for computation in the case of tall data (**?**). In applied analysis, researchers often need to set thousands of MCMC iterations to reach stable posterior distribution, which takes hours or days to implement a single model. Besides, when there are high dimensional data where high-probability regions are concentrated on an extremely limited region of sample space, it is extremely difficult for M-H algorithm or Gibbs sampler to generate effective samples from these small posterior regions (**?**). Hierarchical models even complicate this issue by adding random parameters for each subgroup, which further grows the dimensionality of parameter space. Furthermore, when there is high correlation between different parameters that often occur in high-dimensional data settings, neither the M-H algorithm or Gibbs sampler can efficiently generate effective samples from the posterior distribution. All the problems motivate researchers in different fields to develop different scalable algorithms to make Bayesian inference for big data.

2.6.1 Hamiltonian Monte Carlo (HMC)

Originally proposed by **?** with the name of Hybrid Monte Carlo, the Hamiltonian Monte Carlo (HMC) modifies the random-walk behavior in M-H algorithm into a deterministic one by adding auxiliary momentum parameters p_n, so it explores the high-density regions in big data settings more efficiently compared to the traditional M-H algorithm or the Gibbs sampler (**??**). Although HMC was originally proposed in 1987 (**?**), it is only widely adopted by applied researchers in the recent five years, thanks to the development of the No-U-Turn Sampler (NUTS) (**?**) and the statistical programming language `Stan` (**?**).

Let $\boldsymbol{q}$ denote the position vector and $\boldsymbol{p}$ denote the momentum vector in the conservative dynamics physics system. Note that $\boldsymbol{q}$ and $\boldsymbol{q}$ must have the same length. The combination $(\boldsymbol{q}, \boldsymbol{p})$ then defines a position-momentum phase space, which can be calculated using the conditional distribution (**??**), which could be further defined in terms of the *Hamiltonian*.

$$\pi(\boldsymbol{p}, \boldsymbol{q}) = \pi(\boldsymbol{p}|\boldsymbol{q})\pi(\boldsymbol{q}) = e^{-H(\boldsymbol{p},\boldsymbol{q})}$$

After a little bit of transformation, we have:

$$\begin{aligned} H(\boldsymbol{p}, \boldsymbol{q}) &= -\log \pi(\boldsymbol{p}, \boldsymbol{q}) \\ &= -\log \pi(\boldsymbol{p}|\boldsymbol{q}) - \log \pi(\boldsymbol{q}) \\ &= K(\boldsymbol{p}, \boldsymbol{q}) + V(\boldsymbol{q}) \end{aligned} \tag{2.9}$$

In the perspective of physics, the *Hamiltonian* $H(\boldsymbol{p}, \boldsymbol{q})$ is the total energy of the system, which composes of two parts: *kinetic energy* $K(\boldsymbol{p}, \boldsymbol{q})$ and *potential energy* $V(\boldsymbol{q})$. Note that the potential energy $V(\boldsymbol{q}) = -\log \pi(\boldsymbol{q})$ is essentially the negative log of the posterior distribution of the parameter posterior density $\boldsymbol{q}$. In a static system, the Hamiltonian is a constant. The

evolution of this system is governed by the *Hamiltonian equations*:

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \frac{\partial H}{\boldsymbol{p}} = \frac{\partial K}{\boldsymbol{p}} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial H}{\boldsymbol{q}} = -\frac{\partial K}{\boldsymbol{q}} - \frac{\partial V}{\boldsymbol{q}} \end{aligned} \tag{2.10}$$

We can randomly generate high density proposals in the parameters space by taking advantage of the Hamiltonian system. Here is the general idea of the *HMC algorithm* (**?**):

1. Let $\theta^{(0)}$ denote a random initial value from a proposal distribution,
2. Set $t = 1$,
3. Generate a random initial momentum m from a proposal distribution (typically a multivariate normal distribution),
4. Use the leapfrog algorithm to solve the trajectory moving over the high-density posterior parameter space under the Hamiltonian mechanism for a period,
5. Calculate the new momentum m' and new position $\theta^{(\text{prop})}$
6. Compute the following probability:

$$\alpha^H = \min\left(1, \frac{p\left(\theta^{(\text{prop})}|\boldsymbol{x}\right) p(\theta^{(\text{prop})})}{p\left(\theta^{(t)}|\boldsymbol{x}\right) p(\theta^{(t)})} \frac{q(m')}{q(m)}\right) \tag{2.11}$$

7. Set $\theta^{(t+1)} = \theta^{(\text{prop})}$ with the probability of α^H; otherwise set $\theta^{(t+1)} = \theta^{(t)}$. Set $t \leftarrow t + 1$ and return to 3 until the desired number of iterations is reached.

The HMC is essentially an improved form of M-H algorithm by using the Hamiltonian to generate distant and effective proposals instead of naive random-walk and revised form of the acceptance probability (Equation (2.11)).

Two parameters need to be tuned when implementing the HMC: step size ϵ and the optimal trajectory length T. The optimal trajectory length is the product of the number of steps L and step size ϵ (**??**). The step size ϵ decides how similarly the sympletic methods (typically the leapfrog algorithm) imitates the true unnormalized posterior density. If ϵ is

too small, it will take a lot of steps for the leapfrog algorithm to explore the posterior space. If ϵ is too big, the leapfrog algorithm will loop around and return to a place near its original step. The trajectory length $T = \epsilon L$, which need to be tuned in similar style with ϵ: if L is too short, it will be hard to simulate distant proposal and the algorithm is inefficient; if L is too long, the trajectory will loop back and become computationally inefficient. Hand tuning these two parameters was the major obstacle to implement HMC for applied researchers.

A milestone of automatically tuned HMC is the No-U-Turn Sampler (NUTS) proposed by **?**, which solves the difficulty of hand tuning ϵ and T. NUTS calculates the optimal step size ϵ and number of steps L through a tree building algorithm (**?**). The tree depth k is defined as the number of doublings, resulting in 2^k leapfrog steps to build the trajectory. This k is then decided by repeating the doubling iterations until the trajectory 'makes a U-turn' (loops back) or diverges (the Hamiltonian expands to infinity). Therefore, the NUTS can automatically create trajectories that can efficiently explore the high-density parameter space without having to hand tune ϵ and T.

2.6.2 Subsampling MCMC

With rapid development of automated data collection system, more tall and wide data are becoming commonly available to researchers. A tall dataset has many observations or rows, while a wide dataset has many variables or columns. The emergence of big data poses a threat to the existing MCMC methods, as most of these methods require that the full data likelihood be evaluated at each iteration, which will be computationally intensive in the case of tall and wide data. One way to tackle the computational burden of evaluating the full data likelihood is subsampling MCMC, which means evaluating the likelihood based on multiple subsets of data and then combining the results. Subsampling MCMC via simple random sample often does not work as it does not account for the variability of the log likelihood estimator among different subsamples. The most popular technique of performing subsampling MCMC is via introducing auxiliary variables that reduce the variability of log

likelihood estimators (**?**).

The first well-known subsampling MCMC algorithm is the *firefly MCMC* by **?**, which introduces an auxiliary variable for each observation that can be turned on or off to determine if the observation should be included in likelihood evaluation. Starting from this firefly MCMC algorithm, an increasing number of studies have been published on subsampling MCMC algorithms. **?** proposed to use a sequential hypothesis test to generate *accept-reject samples* with high confidence on a fraction of data. Similar studies that use accept-reject samples include **?** and **?**. Another category of widely discussed subsampling MCMC algorithm is Pseudo-Marginal MCMC (PMCMC), which replaces the likelihood or the natural logarithm of likelihood with an unbiased estimate from a subset of data based on control variates at each MCMC iteration (**??**). They proposed two types of bias-correction log-likelihood estimates: a) parameter expanded control variates via Taylor expansion around a reference value in parameter space, and b) data expanded control variate via Taylor expansion around the nearest centroid in data space. Other subsampling MCMC algorithms include Block-Poisson estimator (**?**), delayed acceptance (**?**), noisy MCMC (**?**), and zig-zag process MCMC (**?**).

Apart from the subsampling MCMC algorithms mentioned above, subsampling MCMC using the Hamiltonian mechanism deserves special attention as it more efficiently explores the posterior in high-dimensional parameter space. **?** proposed a stochastic gradient HMC, which introduces a friction term that counteracts the effects of noisy gradient. In contrast, **?** argued that the stochastic gradient HMC proposed by **?** compromised the scalability of the HMC with respect to the complexity of the target distribution. The paper claimed that subsampled data does not have sufficient information to efficiently explore the target distributions, and devastates the scalable performance of HMC. A algorithm called HMC with energy conserving subsampling (HMC-ECS) by **?** extended the PMCMC algorithm proposed by **?** to HMC by introducing a fictitious momentum vector $\vec{p}$, which has the same dimension as the parameter vector θ.

2.7 Conceptual framework

The conceptual model in this study is based on three frameworks:

1. *Truck Driver Fatigue Model* by **?**,
2. 5×*ST-level hierarchy theory* in traffic safety by **?**,
3. *Commercial motor vehicle driver fatigue framework* by **?**.

Summarized from literature review and focus groups, the *Truck Driver Fatigue Model* includes three general categories of factors to driver fatigue, and each category includes several comparatively specific constructs: truck driving environment (regularity of time, quality of rest, and trip control), economic pressure (scheduling demands of commerce, driver internal economic or personal factors, and carrier economic factors), and organizational carrier support (operational practices and general safety measures) (**?**). **?** proposed a 5-level hierarchy theory in studying traffic safety: geographic region, traffic site, traffic crash, driver-vehicle unit, and occupant. The framework proposed in **?** listed four predictor domains including driver, vehicle, carrier, and environment, as well as five outcome variables, crash rate, serious crash rate, fatal crash rate, safety critical event rate, and fatigue.

Figure 2.1 demonstrates the conceptual framework used in this study. A two-level hierarchy structure is proposed in this study, driver level and trip level. Driver level factors include driver features and fatigue; trip level factors include traffic, road geometry, and weather. These factors are assumed to be directly associated with SCEs, which can be modeled by statistical and reliability models. Final, the SCEs are hypothesized to be directly associated with crashes.

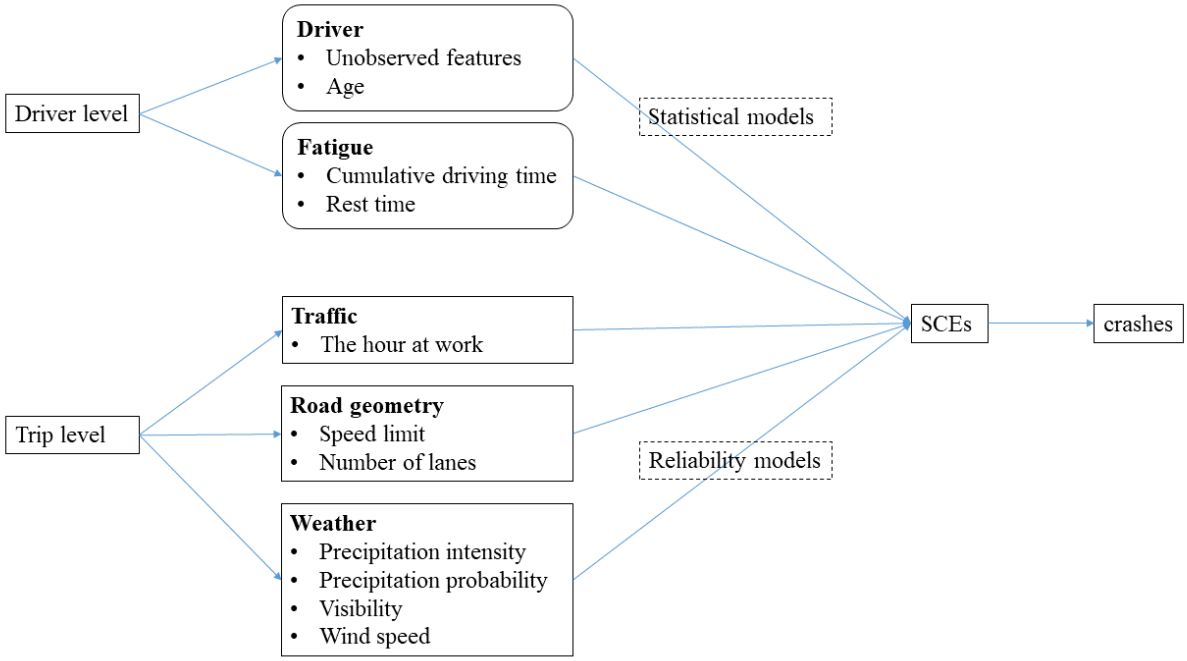

Figure 2.1: *Conceptual model. SCEs represent safety critical events.*

CHAPTER 3

RESEARCH AIMS

The overarching goal of this proposed dissertation is to construct scalable Bayesian hierarchical models for NDS data and understand how cumulative driving time and other factors will impact the performance of truck drivers. However, there are several gaps in traffic safety studies based on the previous literature review. First, an increasing number of studies are using SCEs as the outcome variable, but *the association between crashes and SCEs has not been confirmed among truck drivers.* Second, SCEs are much more common and there can be multiple SCEs within a short period, but *recurrent events models were not widely applied* to understand the risk factors for SCEs. Third, *the high-resolutional and high-dimensional data collected by NDS poses a challenge to Bayesian estimation.* Considering these limitations, the aims of this research will focus on developing innovative and scalable Bayesian hierarchical statistical and reliability models to understand NDS data. Accordingly, the specific aims of this dissertation are as follows:

1. **Aim1: To examine the association between truck crashes and SCEs using a Bayesian Gamma-Poisson regression**. I hypothesize that the rate of crashes is positively associated with the rate of SCEs among truck drivers controlling for the miles driven and other covariates.

2. **Aim2: To construct three scalable Bayesian hierarchical models to identify potential risk factors for SCEs**. I hypothesize that the patterns of SCEs vary

significantly from drivers to drivers and can be predicted using risk factors including cumulative driving time, weather, road geometry, age, speed, speed variation, and others.

- **Sub-aim 1: to contruct a Bayesian hierarchical logistic regression to model the probability of SCEs in 30-minute intervals**. I hypothesize that the probability of SCEs is positively associated with the cumulative driving time and risk factors, and it varies significantly from drivers to drivers.
- **Sub-aim 2: to contruct a Bayesian hierarchical Poisson regression to model the rate of SCEs in 30-minute intervals**. I hypothesize that the rate of SCEs is positively associated with the cumulative driving time and risk factors, and it varies significantly from drivers to drivers.
- **Sub-aim 3: to contruct a Bayesian hierarchical non-homogeneous Poisson process with the power law process intensity function to model the intensity change of SCEs within each shift**. I hypothesize that the intensity of SCEs increases in later stage of shifts, can be predicted by the risk factors, and varies from drivers to drivers.

3. **Aim3: To propose an innovative reliability model that accounts for both within shift cumulative driving time and between-trip rest time**. I hypothesize that the intensity function can be recovered by some proportion or by some amounts during rests between trips, and intensity function varies significantly from drivers to drivers.

CHAPTER 4

METHODS

4.1 Data sources

4.1.1 Real-time ping

A commercial trucking and transportation company in the United States will provide me real-time ping data generated between April 1st, 2015 and March 29th, 2016. During this time, a small device was installed in each of their truck, which will ping irregularly (typically every 2-10 minutes). Each ping will collect real-time data on the vehicle number, date and time, latitude, longitude, driver identification number (ID), and speed at that second. The driver ID is de-identified and no real driver names will be involved. In total, there are 1,494,678,173 pings. A sample of the ping data is demonstrated in Table 4.1.

Table 4.1: *A demonstration of ping data*

trip_id	ping_time	speed	latitude	longitude	driver
100160724	2015-10-23 08:09:26	5	33.94288	-118.1681	canj1
100160724	2015-10-23 08:22:58	4	33.97146	-118.1677	canj1
100160724	2015-10-23 08:23:12	8	33.97178	-118.1677	canj1
100160724	2015-10-23 08:23:30	4	33.97233	-118.1678	canj1
100160724	2015-10-23 08:38:00	40	34.00708	-118.1798	canj1

4.1.2 Truck crashes and SCEs

Real-time time-stamped SCEs and associated GPS locations for all trucks were collected by the truck company and accessible to me as outcome variables. Three types of critical events were recorded: 1) Hard brake, 2) Headway, 3) Rolling stability. Once some kinematic thresholds regarding the driving behavior were met, the sensor will be automatically triggered and the information of these SCEs (latitude, longitude, speed, driver ID) will be recorded. A sample of the SCEs data and crashes are demonstrated in Table 4.2 and Table 4.3.

Table 4.2: *safety critical events*

driver	event_time	event_type
canj1	2015-10-23 14:46:08	HB
canj1	2015-10-26 15:06:03	HB
canj1	2015-10-28 11:58:24	HB
canj1	2015-10-28 17:42:36	HB
canj1	2015-11-02 07:13:56	HB

Table 4.3: *A demonstration of crashes table*

Accident ID	Open date	Open time	Driver	Type	Cause	N_injuries	Fatalities
I1417883	2014-06-10	22:00:00	gres0	L13	99	0	0
I1418899	2014-06-18	10:52:00	gres0	L13	1	0	0
I1430678	2014-10-02	13:38:00	gres0	L13	1	0	0
I1427445	2014-09-04	19:46:00	gres0	L13	1	0	0
I1429286	2014-09-22	05:00:00	gres0	L13	1	0	0
I1432924	2014-10-23	07:00:00	gres0	L25	1	0	0
15384570	2015-11-04	13:01:00	canj1	L70	3	0	0

4.1.3 Driver demographics

A table that includes the birth date of each driver will be provided by the J.B. Hunt Transport Services. The age of the driver can be calculated from this table and merged back to the trips, shifts, and crashes tables via a common unique driver ID. A sample of the driver data is demonstrated in Table 4.4.

4.1.4 Weather data from the `Dark Sky API`

Weather variables, including *precipitation intensity*, *precipitation probability*, *wind speed*, and *visibility*, will be retrieved from the `Dark Sky API`. The `Dark Sky API` allows the users to query historic minute-by-minute weather data anywhere on the globe (**?**). According to the official document, the `Dark Sky API` is supported by a wide range of weather data sources, which are aggregated together to provide the most precise weather data possible for a given location (**?**). Among several different weather data providers I tested, the `Dark Sky API` provides the most accurate and complete weather variables.

Table 4.4: *drivers*

driver	age
canj1	46
farj7	54
gres0	55
hunt	48
kell0	51

To reduce the cost of querying weather data, we will focus on 496 drivers conducting regional work, which generated around the 13 million real-time ping data. These latitude and longitude coordinates will be rounded to two decimal places, which are worth up to 1.1 kilometers. We will also round the time to the nearest hour and ignore those stopping pings. This reduction algorithm will scale the original 13 million real-time ping data down to around five million unique latitude-longitude-date-time combinations. We will use the R package `darksky` to obtain weather variables for these reduced five million unique combinations (**?**). The weather data for these combinations will then be merged back to the original ping data. A sample of the weather data is shown in Table 4.5

Table 4.5: *A demonstration of weather data from the DarkSky API*

ping_time	latitude	longitude	precip_intensity	precip_probability	wind_speed	visibility
2015-10-23 08:09:26	33.94288	-118.1681	0	0	0.21	9.82
2015-10-23 08:22:58	33.97146	-118.1677	0	0	0.22	9.81
2015-10-23 08:23:12	33.97178	-118.1677	0	0	0.22	9.81
2015-10-23 08:23:30	33.97233	-118.1678	0	0	0.22	9.81
2015-10-23 08:38:00	34.00708	-118.1798	0	0	0.24	9.81

4.1.5 Road geometry data from the `OpenStreetMap`

Two road geometry variables for the 496 regional truck drivers will be queried from the `OpenStreetMap` (OSM) project: *speed limits* and *the number of lanes*. The OSM data are collaboratively collected by over two million registered users via manual survey, GPS devices, aerial photography, and other open-access sources (**?**). The OpenStreetMap Foundation supports a website to make the data freely available to the public under the Open Database License.

We will query the speed limits and the number of lanes by specifying a bounding box by defining a center point, as well as the width and height in meters in the `center_bbox()` function available from the `osmar` R package (**?**). We will use real-time longitudes and latitudes as the center point and defined a 100×100 meters box to retrieve the two variables. If the 100×100 meters box is too small to have any road geometry data, we will expand the box to 500×500 and then 1000×1000 to obtain geometry data. If the OSM API returned data from multiple geometry structures, we will take the mean of the returned values as the output. A sample of the road geometry data is shown in Table 4.6.

Table 4.6: *A demonstration of road geometry data from the OpenStreetMap API*

driver	latitude	longitude	speed_limit	num_lanes
farj7	30.32650	-89.86389	65	2
farj7	30.34032	-91.73116	65	2
farj7	30.34174	-91.72572	60	2
farj7	30.35075	-91.69085	60	2
farj7	30.35165	-91.68755	60	2

4.2 Data aggregation

In order to make the MCMC estimation for Bayesian models tractable, I will use the following data reduction algorithms to aggregate real-time ping data to *trips* and *shifts*: a *trip* is a continuous period of driving without stopping for more than 30 minutes; a *shift* is a

long period of driving without stopping for more than 8 hours.

4.2.1 Shifts

The trips data will be further divided into different shifts if the specific driver was not moving for eight hours. A sample of the shifts data is shown in Table 4.7.

```
ping_time latitude longitude precip_intensity
```

1 2015-10-23T08:09:26Z 33.94288 -118.1681 0 2 2015-10-23T08:22:58Z 33.97146 -118.1677 0 3 2015-10-23T08:23:12Z 33.97178 -118.1677 0 4 2015-10-23T08:23:30Z 33.97233 -118.1678 0 5 2015-10-23T08:38:00Z 34.00708 -118.1798 0 precip_probability wind_speed visibility 1 0 0.21 9.82 2 0 0.22 9.81 3 0 0.22 9.81 4 0 0.22 9.81 5 0 0.24 9.81

Table 4.7: *A demonstration of transformed shifts data*

x
latex

4.2.2 Trips

For each of the truck drivers, if the real-time ping data showed that the truck was not moving for more than 30 minutes, the ping data will be separated into two different trips. A sample of the trips data is shown in Table 4.8.

Table 4.8: *A demonstration of transformed trips data*

driver	trip_id	start_time	end_time	trip_time	distance
canj1	100160724	2015-10-23 08:09:26	2015-10-23 08:37:26	28	4.473
canj1	100160725	2015-10-23 09:04:24	2015-10-23 11:21:24	137	46.721
canj1	100160726	2015-10-23 12:00:36	2015-10-23 15:37:36	217	164.576
canj1	100160727	2015-10-23 16:38:10	2015-10-23 18:37:10	119	52.907
canj1	100160728	2015-10-26 07:49:04	2015-10-26 10:52:04	183	104.085

4.2.3 30-minute intervals

As the length of a trip can vary significantly from 5 minutes to more than 8 hours, I will transform the trips data into 30-minute fixed intervals according to the starting and ending time of trips. The 30-minute interval data dissect unnecessarily lengthy trips into small chunks and enable statistical analyses based on these small-interval data. A sample of the 30-minute interval data is shown in Table 4.9.

Table 4.9: *A demonstration of transformed 30-minute intervals*

driver	interval_id	start_time	end_time	interval_time	distance
canj1	197089	2015-10-23 08:09:26	2015-10-23 08:38:00	28	4.538
canj1	197090	2015-10-23 09:04:24	2015-10-23 09:34:24	30	2.645
canj1	197091	2015-10-23 09:34:24	2015-10-23 10:04:24	30	0.984
canj1	197092	2015-10-23 10:04:24	2015-10-23 10:34:24	30	5.928
canj1	197093	2015-10-23 10:34:24	2015-10-23 11:04:24	30	17.348

4.3 Data merging

Figure 4.1 demonstrates the data aggregation and merging workflow. The left part shows the data aggregation from the original ping data to trips, 30-minute intervals, and shifts, which have been demonstrated in Section 4.2. The right part demonstrates the process of merging covariates table (SCEs, drivers, road geometry, and weather) back to the aggregated tables (trips, 30-minute intervals, and shifts tables). The specific details of the merging process.

1. *SCEs*: the SCEs will be merged to the two aggregated tables by drivers and if the time of SCEs fall between the start and end time of the aggregated tables,
2. *Drivers*: the age of drivers are merged to the two aggregated tables using driver ID,
3. *Road geometry*: the road geometry variables will be merged to the original ping data by driver ID, latitude, and longitude. Then they variables will be aggregated by taking the mean for each 30-minute interval and shift,

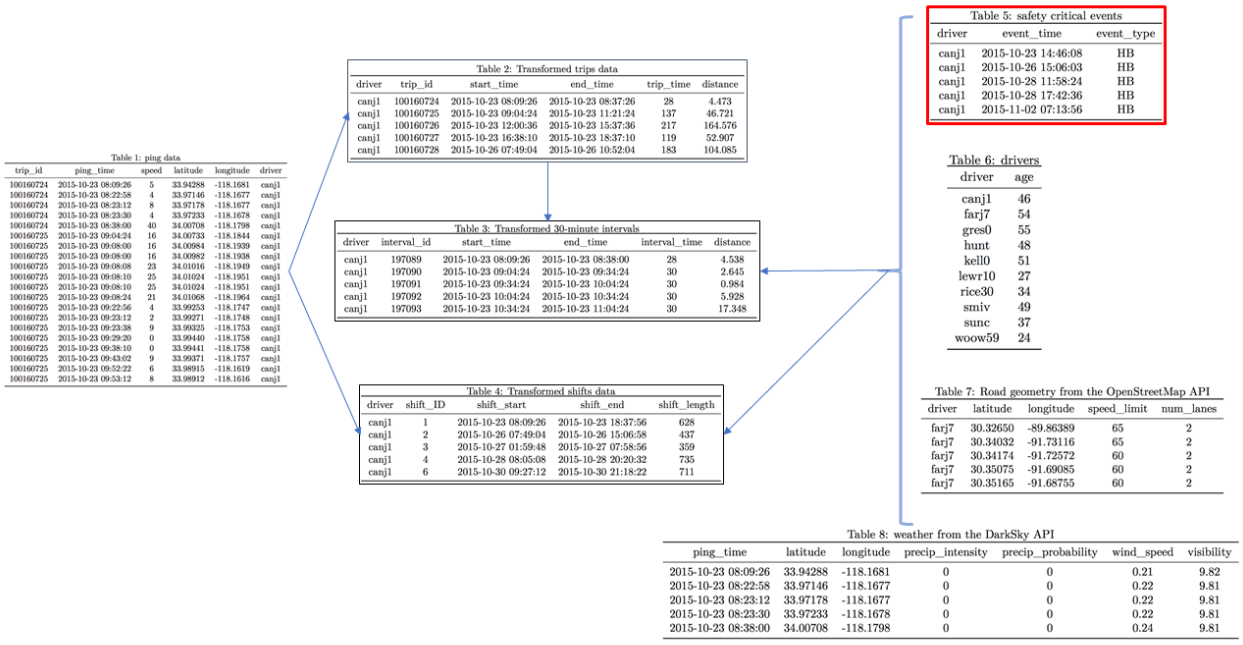

Figure 4.1: *Flow chart of data aggregation and merging*

4. *Weather*: the weather variables will be merged to the original ping data by driver ID, latitude, longitude, date, and time. These weather variables will then be aggregated by taking the mean for each 30-minute interval and shift.

The resulting 30-minute intervals and shifts tables are demonstrated in Table 4.10 and Table 4.11. The predictor variables such as cumulative driving time, driver's age, weather and road geometry variables are truncated and not shown to fit in the page. Table 4.12 demonstrates the SCEs table, with time to events calculated as the time difference in hours between the time of the SCE and the starting time of the corresponding shift.

Table 4.10: *30 minutes intervals data for hierarchical logistic and Poisson regression*

driver	start_time	end_time	interval_time	distance
canj1	2015-10-23T08:09:26Z	2015-10-23T08:38:00Z	28	4.538
canj1	2015-10-23T09:04:24Z	2015-10-23T09:34:24Z	30	2.645
canj1	2015-10-23T09:34:24Z	2015-10-23T10:04:24Z	30	0.984
canj1	2015-10-23T10:04:24Z	2015-10-23T10:34:24Z	30	5.928
canj1	2015-10-23T10:34:24Z	2015-10-23T11:04:24Z	30	17.348

Table 4.11: *A demonstration of shifts data for hierarchical non-homogeneous Poisson process*

driver	shift_ID	n_SCE	SCE_time	SCE_type	start_time	end_time
canj1	1	1	2015-10-23 14:46:08	HB	2015-10-23T08:09:26Z	2015-10-23T18:37:56Z
canj1	2	1	2015-10-26 15:06:03	HB	2015-10-26T07:49:04Z	2015-10-26T15:06:58Z
canj1	3	0	NA	NA	2015-10-27T01:59:48Z	2015-10-27T07:58:56Z
canj1	4	2	2015-10-28 11:58:24;2015-10-28 17:42:36	HB;HB	2015-10-28T08:05:08Z	2015-10-28T20:20:32Z
canj1	6	0	NA	NA	2015-10-30T09:27:12Z	2015-10-30T21:18:22Z

Table 4.12: *A demonstration of SCEs data for hierarchical non-homogeneous Poisson process*

driver	shift_ID	start_time	event_time	shift_length	time2event
canj1	1	2015-10-23 08:09:26	2015-10-23 14:46:08	10.467	6.600
canj1	2	2015-10-26 07:49:04	2015-10-26 15:06:03	7.283	7.267
canj1	4	2015-10-28 08:05:08	2015-10-28 11:58:24	12.250	3.883
canj1	4	2015-10-28 08:05:08	2015-10-28 17:42:36	12.250	9.617
canj1	7	2015-11-02 06:26:48	2015-11-02 07:13:56	13.667	0.783

4.4 Analytical Plan for Aim 1

The first aim seeks to determine the association between the rate of crashes and the rate of SCEs at the level of drivers.

Data: this aim will use the original ping data table that has 1,494,678,173 pings in total, as demontrated in Table 4.1. Since drivers with less than 100 real-time pings will be recognized as potential outliers and only include drivers with at more than 100 pings. The cleaned ping data will be aggregated to trips according to the procedure defined in Section 4.2.

Outcome and predictor variables: the outcome variable will be the number of crashes for each driver. The primary independent variable will be the number of SCEs per 10,000 miles. These SCEs will be further decomposed into the number of hard brakes, headways, and rolling stability per 10,000 miles in similar analysis. The covariates will be the total miles driven, the percent of night driving, and the age of the drivers.

Statistical models: since the outcome variable is a count variable, a Poisson model or a negative binomial model is a natural choice for this type of outcome variable (**?**). However,

these two models are less likely to fully account for the variance across drivers. Therefore, I propose to use a Gamma-Poisson model to examine the association between crashes and SCEs. Here is how the proposed Gamma-Poisson model will be implemented. Let us assume that:

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$

$$X|\lambda \sim \text{Poisson}(\lambda)$$

Then we have:

$$X \sim \text{Gamma-Poisson}(\alpha, \beta)$$

The Gamma-Poisson distribution is a α-parameter distribution, with the α as a measure of overdispersion. The Gamma-Poisson distribution has the probability mass function of:

$$f(x) = \frac{\Gamma(x+\beta)\alpha^x}{\Gamma(\beta)(1+\alpha)^{\beta+x}x!}, \quad x = 0, 1, 2, \ldots$$

The mean and variance of a Gamma-Poisson distribution are:

$$\begin{aligned} E(X) &= \alpha\beta \\ V(X) &= \alpha\beta + \alpha^2\beta \\ &= \alpha\beta(1+\alpha) \end{aligned}$$

The log-linear Gamma-Poisson model will be specified as:

$$\log\beta = \mathbf{X}\gamma - \log m,$$

where $\mathbf{X}$ is the predictor variables matrix, including the percent of night driving and the age of the drivers, γ is the associated $2 * 1$ parameter vector, m is the total miles driven as an offset term in the Poisson distribution, and α is a fixed overdispersion parameter that does not depend on any covariates.

All data reduction, cleaning, and statistical analysis will be done on the RStudio Server

on the Ohio Supercomputer Center (OSC). The OSC provides high performance computing resources and expertise to academic researchers (**?**). The Bayesian log-linear Gamma-Poisson model will be conducted using self-defined `Stan` functions, which can be accessed via the `rstan` package in statistical computing environment R 3.5.1 (**??**).

Potential problems and alternative plans: the sheer size of the original ping data may be a problem in this aim. The ping data has 1,494,678,173 rows and 9 columns, which takes up more than 140 gigabytes (GB) when stored as a single comma-separated values (csv) file. Although I will use the OSC server that has Random-Access Memory (RAM) of more than 500 GB, it may still be hard and slow to read and process this giant file. In that case, I will separate the single giant csv file into several small csv files according to driver ID, then aggregate the pings to trips for each small csv file. After the ping data are aggregated to trips, it is unlikely that the log-linear Gamma-Poisson model fail. In that unlikely event, I can turn to negative binomial models or use traditional MLE estimates instead of Bayesian estimation.

4.5 Analytical Plan for Aim 2

The purpose of aim 2 is to develop three scalable hierarchical Bayesian statistical and reliability models for the SCEs of truck drivers and identify potential risk factors. Bayesian hierarchical logistic regression, Poisson regression, and NHPP with the PLP intensity function will be used. All three hierarchical models will account for both driver-level and trip-level variables. Hamiltonian Monte Carlo with energy conserving subsampling (HMC-ECS) proposed by **?** will be used to scale the Bayesian hierarchical models for the 30-minute interval table (more than one million rows) with 496 random intercepts and slopes. The details of the statistical models will be explained in the following sections.

4.5.1 Bayesian hierarchical logistic regression

Here the probability of a critical event occurred will be modeled using a Bayesian hierarchical logistic regression, as shown in Equation (4.1). I will categorize the number of safety events during a 30-minute interval into a binary variable $Y_{i,d}$ of either 0 or 1, where 0 indicates that no critical event occurred while 1 indicates that at least 1 critical event during that interval. The analysis will be based on the merged 30-minute interval data, as shown in Table 4.10.

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p_i) \\ \log \frac{p_i}{1-p_i} &= \beta_{0,d(i)} + \beta_{1,d(i)} \cdot \text{CT}_i + \sum_{j=1}^{J} x_{ij}\beta_j \\ \beta_{0,d} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2), \quad d = 1, 2, \cdots, D \\ \beta_{1,d} &\sim \text{i.i.d. } N(\mu_1, \sigma_1^2), \quad d = 1, 2, \cdots, D \end{aligned} \tag{4.1}$$

Where μ_0 and σ_0 are hyper-parameters for random intercepts $\beta_{0,d}$, μ_1 and σ_1 are hyper-parameters for random slopes $\beta_{1,d}$, and $\beta_2, \beta_3, \cdots, \beta_J$ are fixed parameters for covariates x_{ij}. Since we do not have much prior knowledge on the parameters, I will assign weakly informative priors (**?**) for these parameters shown in Equation (4.2).

$$\begin{aligned} \mu_0 &\sim N(0, 5^2) \\ \mu_1 &\sim N(0, 5^2) \\ \sigma_0 &\sim \text{Gamma}(1, 1) \\ \sigma_1 &\sim \text{Gamma}(1, 1) \\ \beta_2, \beta_3, \cdots, \beta_J &\sim N(0, 10^2) \end{aligned} \tag{4.2}$$

Since μ_0 and μ_1 can be any real number, I will assign two weakly informative normal distributions with mean of 0 and standard deviation of 5 as the priors for these two hyperparameters. The priors for the hyperpriors need to be relatively more restrictive than priors for

fixed-effects parameters $\beta_2, \beta_3, \cdots, \beta_J$ (**?**). In comparison, σ_0 and σ_1 must be strictly positive, so I will Gamma$(1, 1)$ with wide distribution on positive real numbers as their priors.

4.5.2 Bayesian hierarchical Poisson regression

The logistic regression considers whether SCEs occurred in a 30-minute fixed interval or not, but it ignores the intensity of SCEs. Therefore I further propose a Bayesian hierarchical Poisson regression to model the association between cumulative driving time and the rate of SCEs, as shown in Equation (4.3). Each driver will have a different a random intercept and a random slope on cumulative driving time. The analysis will also be based on the merged 30-minute interval data, as shown in Table 4.10.

$$\begin{aligned} N_i &\sim \text{Poisson}(T_i \cdot \lambda_i) \\ \log \lambda_i &= \beta_{0,d(i)} + \beta_{1,d(i)} \cdot \text{CT}_i + \sum_{j=1}^{J} x_{ij}\beta_j \\ \beta_{0,d} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2), \quad d = 1, 2, \cdots, D \\ \beta_{1,d} &\sim \text{i.i.d. } N(\mu_1, \sigma_1^2), \quad d = 1, 2, \cdots, D \end{aligned} \tag{4.3}$$

Where N is the number of SCEs for driver $d(i)$ in time interval i, and it has a Poisson distribution with the mean and variance parameter λ. T_i is the length of the time interval, which serves as an offset term. Most of T_is in the 30-minute interval table are 30 minutes. Other variables are identical as those described in Equation (4.1); the only thing that is changed here is the outcome distribution and offset term T_i. Therefore, I will assign identical prior distributions for these parameter as specified in Equation (4.2).

4.5.3 Non-homogeneous Poisson process (NHPP)

Despite Poisson regression consider the frequency and rate of SCEs in a given interval, it assumes that the intensity of SCEs is a constant. This constant intensity assumption may not be true in real-life transportation practice. Based on the merged shifts data set shown in

Table 4.7, I present a non-homogeneous Poisson process (NHPP) with a power law process (PLP) intensity function. This model will answer whether SCEs occurred more frequently at early stages of shifts, towards the end of shifts, or does not show significant patterns.

Figure 4.2 shows a sample of SCEs distributions in different shifts. Each arrow represents a shift while each red cross shows a SCE. These recurrent events data fit into the analysis framework of point process and reliability models. A point process is a stochastic model that describes the occurrence of events in a given period (**?**). The mean function of a point process is $\Lambda(t) = E(N(t))$, where $\Lambda(t)$ is the expected number of failures through time t. Two notations that are important in reliability models are *Rate of Occurence of Failures (ROCOF)* and *Intensity function*.

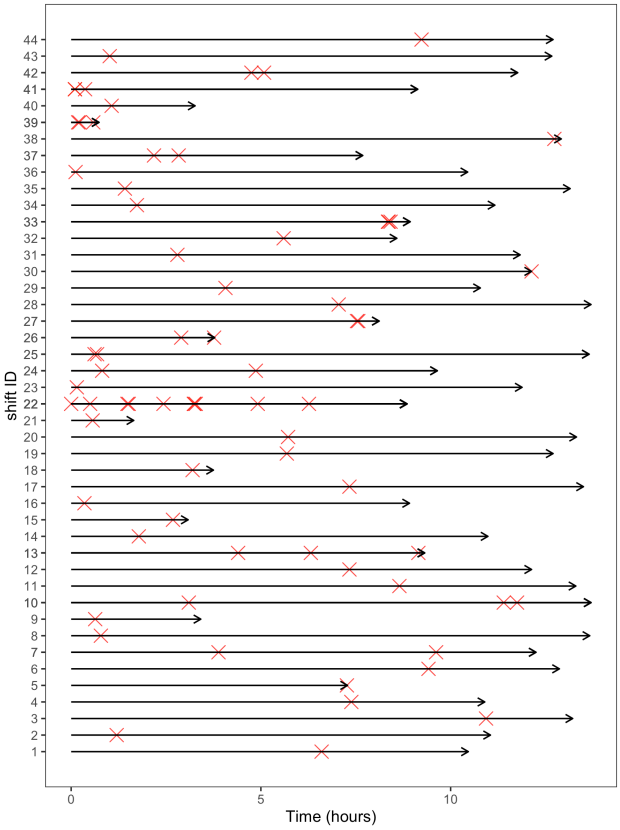

Figure 4.2: *An arrow plot of time to SCEs in each shift*

1. *ROCOF*: When the mean function $\Lambda(t)$ is differentiable, the ROCOF is $\mu(t) = \frac{d}{dt}\Lambda(t)$. The ROCOF can be interpreted as the instantaneous rate of change in the expected number of failures,
2. *Intensity function*: The intensity function of a point process is $\lambda(t) = \lim_{\Delta t \to 0} \frac{P(N(t,t+\Delta t] \geq 1)}{\Delta t}$.

When there is no simultaneous events, ROCOF is the same as the intensity function. A commonly used reliability model is the *Nonhomogeneous Poisson Process (NHPP)*, which is

a Poisson process whose intensity function is non-constant. The Power law process (PLP) is a special case of a NHPP when the intensity function is:

$$\lambda(t) = \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1}, \tag{4.4}$$

where $\beta > 0$ and $\theta > 0$, also known the Weibull intensity function. The mean function $\Lambda(t)$ is the integral of the intensity function: $\Lambda(t) = \int_0^t \lambda(t)dt = \int_0^t \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1} = \left(\frac{t}{\theta}\right)^{\beta}$.

There are two forms of truncation in a NHPP: 1) *Failure truncation* when testing stops after a predetermined number of failures, 2) *Time truncation* when testing stops at a predetermined time t. Since the drivers typically decide to stop working based on a certain amount of working time, not based on the number of SCEs they already have, time truncation is the case in this study. In a time truncated case, the likelihood function for $f(n, t_1, t_2, \cdots, t_n)$ is shown in Equation (4.5) (the prove can be found at page 54 in **?**).

$$\begin{aligned} f(n, t_1, t_2, \cdots, t_n) &= f(n) f(t_1, t_2, \cdots, t_n | n) \\ &= \frac{e^{-\int_0^\tau \lambda(u)du}[\int_0^\tau \lambda(u)du]^n}{n!} n! \frac{\prod_{i=1}^n \lambda(t_i)}{[\Lambda(\tau)]^n} \\ &= \Big(\prod_{i=1}^n \lambda(t_i)\Big) e^{-\int_0^\tau \lambda(u)du} \\ &= \Big(\prod_{i=1}^n \frac{\beta}{\theta}(\frac{t_i}{\theta})^{\beta-1}\Big) e^{-(\tau/\theta)^\beta}, \\ n &= 0, 1, 2, \cdots, \quad 0 < t_1 < t_2 < \cdots < t_n \end{aligned} \tag{4.5}$$

After the likelihood function of a NHPP is given, the NHPP with PLP can be specified. Let $T_{d,s,i}$ denotes the time to the d-th driver's s-th shift's i-th critical event. The total number critical events of d-th driver's s-th shift is $n_{d,s}$. The ranges of these notations are:

- $i = 1, 2, \cdots, n_{d,S_d}$,
- $s = 1, 2, \cdots, S_d$,
- $d = 1, 2, \cdots, D$.

I assume that the times of critical events within the d-th driver's s-th shift were generated from a PLP, with a fixed shape parameter β and varying scale parameters $\theta_{d,s}$ across drivers d and shifts s. In a PLP, the intensity function of the NHPP is $\lambda(t) = \frac{\beta}{\theta}(\frac{t}{\theta})^{\beta-1}$. The model is described in Equation (4.6).

$$\begin{aligned} T_{d,s,1}, T_{d,s,2}, \cdots, T_{d,s,n_{d,s}} &\sim \text{PLP}(\beta, \theta_{d,s}) \\ \beta &\sim \text{Gamma}(1,1) \\ \log\theta_{d,s} &= \gamma_{0d} + \gamma_1 x_{d,s,1} + \gamma_2 x_{d,s,2} + \cdots + \gamma_k x_{d,s,k} \\ \gamma_{01}, \gamma_{02}, \cdots, \gamma_{0D} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2) \\ \gamma_1, \gamma_2, \cdots, \gamma_k &\sim \text{i.i.d. } N(0, 10^2) \\ \mu_0 &\sim N(0, 5^2) \\ \sigma_0 &\sim \text{Gamma}(1,1) \end{aligned} \tag{4.6}$$

The shape parameter β shows the reliability changes of drivers. When $\beta > 1$, the intensity function $\lambda(t)$ is increasing, the reliability of drivers is decreasing, and SCEs are becoming more frequent; when $\beta < 1$, the intensity function $\lambda(t)$ is decreasing, the reliability of drivers is increasing, and SCEs are becoming less frequent; when $\beta = 1$, the NHPP is simplified as a homogeneous Poisson process with the intensity of $1/\theta$. The $\theta_{d,s}$ is a scale parameter that does not reflect reliability changes.

All data reduction, cleaning, statistical analysis, and visualization will be done on the RStudio Server and Jupyter Server on the Ohio Supercomputer Center (OSC). The OSC provides high performance computing resources and expertise to academic researchers (**?**). The scalable Bayesian statistical and reliability models will be conducted using the HMC-ECS algorithm (self-defined functions in Python 3.6.0) or HMC (the `rstan` package in statistical computing environment R 3.5.1) (**???**).

Potential problems and alternative plans: the sheer size of the 30-minute interval table and merged shifts table may be a problem in this aim. The 30-minute interval table has

one million rows and 10 variables, and the merged shift table has more than 200,000 rows and 10 variables. In the meanwhile, each of the three models will have 496 random intercepts and slopes, which is extremely difficult to estimate in the Bayesian setting. Although I propose to use the HMC-ECS to estimate the random effect, there are still chances that the model does not work. In that case, I will sample 50 to 200 typical drivers, then conduct the analysis based on this smaller sample data. In the unlikely event that the models still fails based on this smaller data, I can restrict the hierarchical models to only have random intercepts or use traditional MLE instead of Bayesian estimation.

4.6 Analytical Plan for Aim 3

Aim 3 seeks to innovate the NHPP using a PLP intensity function proposed in Aim 2. I propose to account for the rest time within a shift by adding one more parameter κ, *the percent of reliability recovery during a break within a shift*. This new reliability model (*jump-point PLP, JPLP*) will be between a NHPP where the intensity function is not influenced by between-trip rests ("as bad as old") and a renewal process where the intensity function is fully recovered by between-trip rests ("as good as new"). The intensity function of the proposed jump-point PLP will be recovered for a certain percent κ every time the driver took a short break (less than eight hours) between trips.

Figure 4.4 demonstrates the complete intensity function of a PLP (the solid curve), simulated SCEs (red crosses on the x-axis), and between-trip rests (the green intervals) of a NHPP used in Aim 2. This figure demonstrates its limitation: when the driver takes a break, the intensity function does not change at all. However, when the driver takes a break, the fatigue level of the driver theoretically should be decreasing and the intensity of SCEs should therefore decrease.

Therefore, the jump-point PLP I propose will account for the effect of taking a rest. Figure 4.3 shows the proposed complete intensity function (black solid curves), continuous PLP intensity function of a NHPP (the grey solid curve), simulated SCEs (red crosses on the

x-axis), and rest times between trips (green intervals). Everytime the driver takes a rest, the complete intensity function will be moved down by a certain percent κ, here the $\kappa = 0.8$.

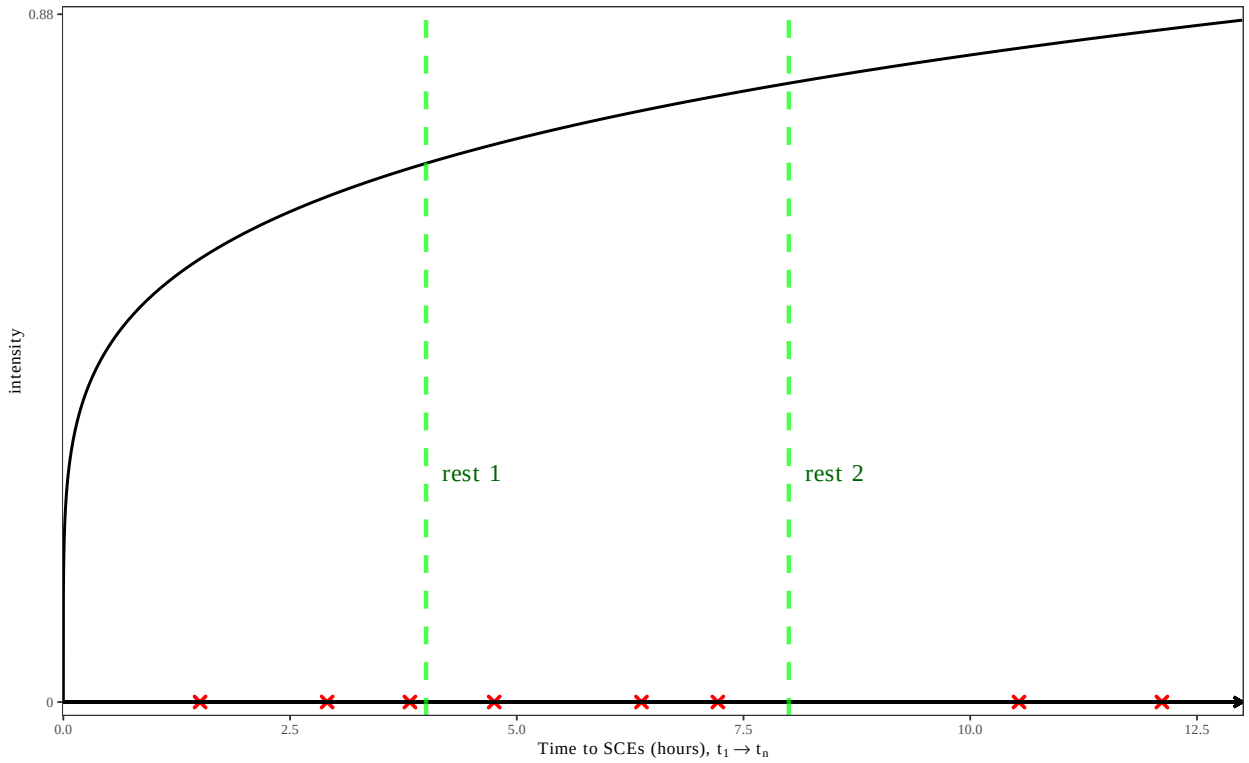

Figure 4.3: *Intensity function, time to SCEs, and rest time within a shift generated from a NHPP with a PLP intensity function,* $\beta = 1.2$, $\theta = 2$

The data and notations $T_{d,s,i}, d, s, i$ will be identical as the PLP specified in Aim 2. Here I assume that the times of critical events within the d-th driver's s-th shift were generated from a JPLP, with a fixed shape parameter β, varying scale parameters $\theta_{d,s}$ across drivers d

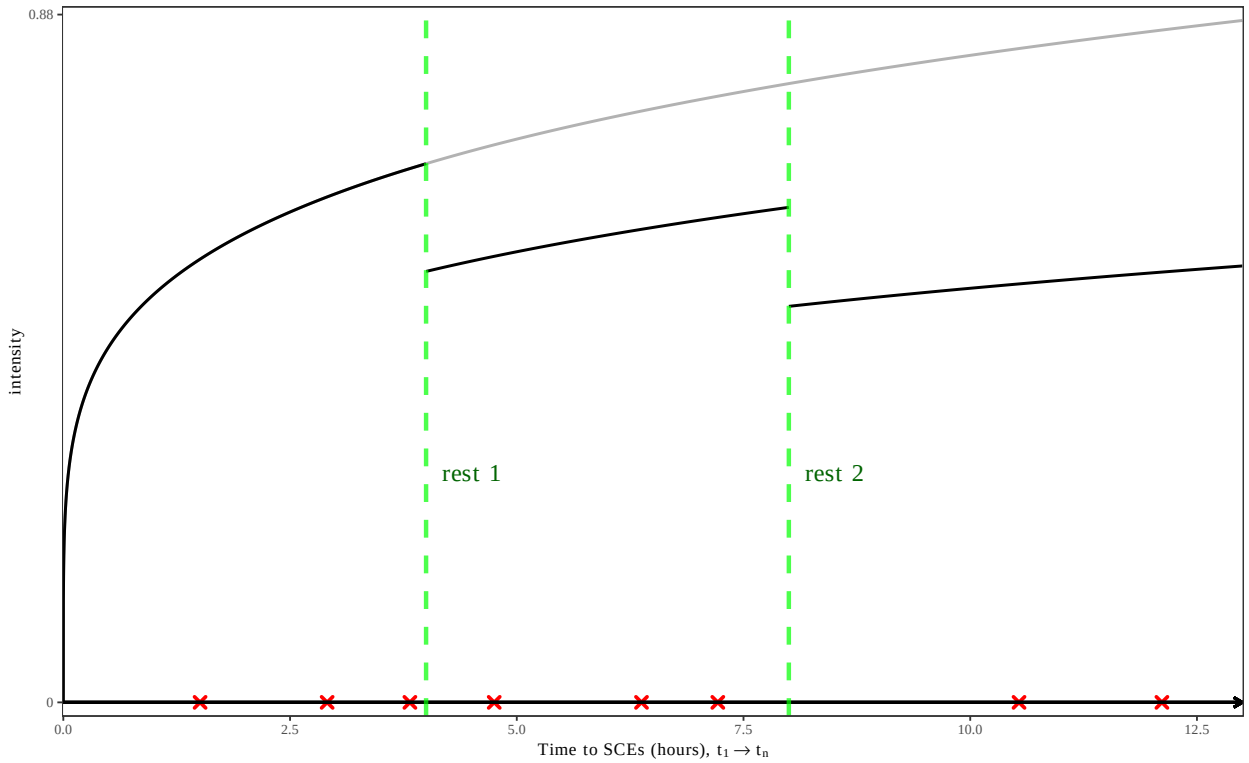

Figure 4.4: *Intensity function, time to SCEs, and rest time within a shift with a jump-point PLP intensity function,* $\beta = 1.2$, $\theta = 2$, $\kappa = 0.8$

and shifts s, and a parameter κ, as shown in Equation (4.7).

$$
\begin{aligned}
T_{d,s,1}, T_{d,s,2}, \cdots, T_{d,s,n_{d,s}} &\sim \text{JPLP}(\beta, \theta_{d,s}, \kappa) \\
\beta &\sim \text{Gamma}(1,1) \\
\log \theta_{d,s} &= \gamma_{0d} + \gamma_1 x_{d,s,1} + \gamma_2 x_{d,s,2} + \cdots + \gamma_k x_{d,s,k} \\
\kappa &\sim \text{Uniform}(0,1) \\
\gamma_{01}, \gamma_{02}, \cdots, \gamma_{0D} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2) \\
\gamma_1, \gamma_2, \cdots, \gamma_k &\sim \text{i.i.d. } N(0, 10^2) \\
\mu_0 &\sim N(0, 5^2) \\
\sigma_0 &\sim \text{Gamma}(1,1)
\end{aligned}
\tag{4.7}
$$

The shape parameter β shows the reliability changes of drivers. When $\beta > 1$, the intensity function $\lambda(t)$ is increasing, the reliability of drivers is decreasing, and SCEs are becoming

more frequent; when $\beta < 1$, the intensity function $\lambda(t)$ is decreasing, the reliability of drivers is increasing, and SCEs are becoming less frequent; when $\beta = 1$, the NHPP is simplified as a homogeneous Poisson process with the intensity of $1/\theta$. $\theta_{d,s}$ is a scale parameter. κ is a parameter that reflects the percent of intensity function recovery once the driver takes a break.

Potential problems and alternative plans: in the unlikely event that the JPLP fails to be models, I will use the *modulated PLP* proposed by **?**. The modulated PLP has well-defined data generating process, intensity function, and joint-likelihood functions. If the JPLP does not work, I will revise the modulated PLP into a hierarchical modulated PLP, on which this Aim 3 will be based. The hierarchical JPLP and hierarchical modulated PLP will be estimated using `Stan` programs by adding self-defined likelihood function, which can be accessed via the `rstan` package in statistical computing environment R 3.5.1 on the OSC (**???**).

BIBLIOGRAPHY

VITA AUCTORIS

Miao Cai was born and raised in Xinzhou district, Wuhan, Hubei Province, China.